Predicción de desorden

Lucía Chemes y Juliana Glavina

Recursos a utilizar:

- DisProt https://www.disprot.org
- IUPred2A https://iupred2a.elte.hu/plot
- MobiDB http://mobidb.bio.unipd.it/

Métodos de predicción de desorden

Objetivos:

- Familiarizarse con distintos métodos de predicción de desorden
- Interpretación de los resultados de los distintos métodos

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: (1) predicción de desorden a partir de composición de secuencia, (2) a partir *machine learning* sobre estructuras determinadas por cristalografía de rayos X y (3) los 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.

Ejercicio 1.

1. 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.
2. ProViz es una herramienta que permite visualizar alineamientos y estructura de dominios de una proteína. Ingresa a la web de proviz, y busca la proteína Calcineurina A (Q08209): http://proviz.ucd.ie/proviz.php?uniprot_acc=Q08209&alignment=QFO

   Observa las regiones indicadas como desordenadas y estructuradas por IUPred. ¿Existe diferencia en la composición de secuencia? ¿Se observan diferencias en el grado de conservación? ¿Cuales parecen estar mejor alineadas? ¿A qué pueden deberse estas diferencias?

Ejercicio 2. Base de datos MobiDB

Objetivo:

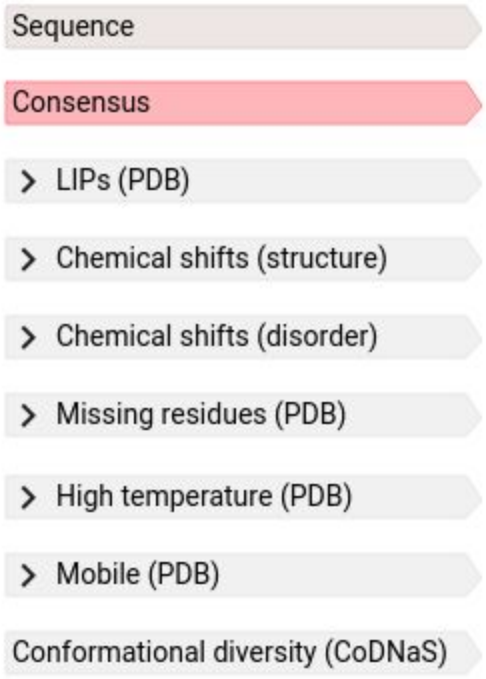
- Familiarizarse con la base de datos MOBIdb

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).

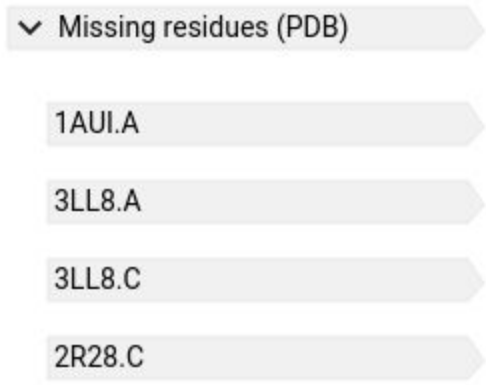
2. Ve a la pestaña *Predictions*. ¿Cuáles regiones son predichas como desordenadas por la mayoría de los métodos? ¿Qué métodos predicen más desorden y cuáles menos? ¿Hay mucha variación?

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. Exploraremos 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 3. 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
MMQDLRLILIIVGAIAIIALLVHGFWTSRKERSSMFRDRPLKRMKSKRDDDSYDEDVEDDEGVGEVRVH
RVNHAPANAQEHEAARPSPQHQYQPPYASAQPRQPVQQPPEAQVPPQHAPHPAQPVQQPAYQPQPEQPL
QQPVSPQVAPAPQPVHSAPQPAQQAFQPAEPVAAPQPEPVAEPAPVMDKPKRKEAVIIMNVAAHHGSEL
NGELLLNSIQQAGFIFGDMNIYHRHLSPDGSGPALFSLANMVKPGTFDPEMKDFTTPGVTIFMQVPSYG
DELQNFKLMLQSAQHIADEVGGVVLDDQRRMMTPQKLREYQDIIREVKDANA
```

2. Utilizando IUPred2A, ¿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”.

   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.

EJERCICIOS ADICIONALES

Base de datos 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.

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.

Cada una de las entradas en la base de datos posee un identificador único.

Objetivos:

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

Ejercicio Adicional 1.

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).

A. 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.

B. ¿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?

C. ¿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?

D. ¿Se observa algún dominio globular conservado?

    a. Expande "Domains". ¿A qué corresponden los segmentos coloreados? ¿Qué tipo de evidencia poseen dichos fragmentos?

E. ¿La evidencia experimental recolectada coincide con las predicciones realizadas en el ejercicio 1?

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?