Capítulo 1

Biología estructural

Alberto Pascual et al.

1.1. Proteínas desordenadas

Mª Inmaculada Yruela Guerrero

1.1.1. Introducción

En general se tiene la idea, acuñada durante el s. XX, de que existe una estrecha relación entre la estructura y la función de una proteína (*secuencia de aminoácidos* → *estructura 3D* → *función*), y que una condición indispensable para que una proteína desarrolle una función es que tenga en su estado nativo una estructura 3D. Sin embargo, en los últimos 20 años los estudios han revelado que hay un número elevado de proteínas que no adoptan estructuras tridimensionales definidas y realizan importantes funciones biológicas en su estado nativo. La posible utilidad de tales regiones fue sugerida por primera vez hace 70 años por Linus Pauling, quién especuló sobre su flexibilidad en la producción de anticuerpos. A nivel estructural, los primeros indicios de regiones desestructuradas en proteínas surgieron cuando tan sólo se habían determinado 20 estructuras de proteínas por rayos-X, en las que aparecían regiones no discernibles debido a su pobre densidad electrónica, y que sin embargo tenían una relevante función.

Esta clase de proteínas cuestionaba en parte el paradigma central de la biología molecular formulado por Francis Crick (1958) que postula que a cada secuencia de aminoácidos le corresponde una estructura tridimensional. La pérdida de densidad electrónica en las estructuras de rayos-X puede surgir por un fallo al resolver el problema de la fase, por defectos del cristal o por sucesos de eliminación proteolítica accidental durante el proceso de purificación de la proteína. Sin embargo, una explicación común para la falta de densidad electrónica es que el átomo, residuo, cadena lateral o segmento no observado no disperse los X-rayos de forma coherente, debido a la variación en su posición de una proteína a la siguiente o próxima, es decir, los átomos no observados son desordenados.

En 1978, el mismo año en que el desorden funcional fue definido por cristalografía de rayos-X, la técnica de resonancia paramagnética electrónica (NMR) reveló que la cola de histona H5, altamente

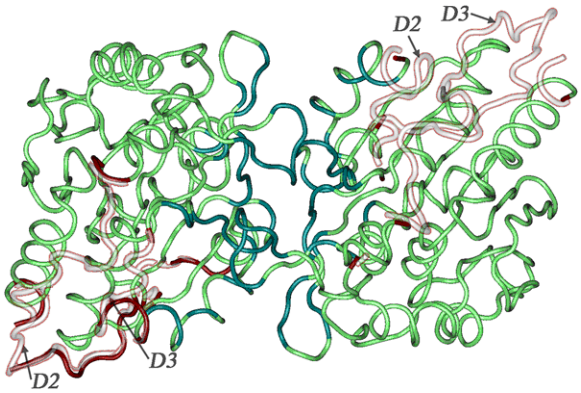

Figura 1.1: Estructura de la chaperona hsp31 (pdb 1PV2). Las regiones estructuradas se muestran en verde, y las regiones desestructuradas en blanco (regiones no definidas en pdb 1PV2). El posible orden que se muestra de las regiones en blanco está simulado en base a la estructura del pdb 1ONS (misma proteína en estado monomérico sin desorden) [**?**].

cargada era desordenada y podía ser clasificada como una proteína desordenada o desestructurada (Aviles et al. 1978). Actualmente la literatura contiene numerosos datos de regiones desordenadas que son esenciales para la función de las proteínas.

Las proteínas que carecen de estructuras definidas se conocen actualmente por el nombre de *proteínas intrínsicamente desordenadas* (PID) y están presentes en todos los organismos vivos. Las PIDs pueden contener regiones desordenadas, y ser así parcialmente desordenadas, o carecer de un plegamiento estructurado en su conjunto, siendo por tanto completamente desordenadas de forma aislada. Desde un punto de vista termodinámico el desorden en una proteína se define como un estado estructural random coil. El desorden puede encontrarse en bucles flexibles o giros, dominios, unión entre dominios o en proteínas completas. Las PIDs no pueden ser descritas por una sola conformación pues adoptan múltiples estructuras. Deben representarse como un conjunto de éstas, algunas compactas, otras extendidas, de estabilidad similar y que se intercambian a una gran velocidad, más de un millón de veces por segundo. La presencia de regiones desordenadas confiere flexibilidad, lo que es una ventaja para el reconocimiento de múltiples moléculas (ARN, ADN, otras proteínas, pequeños ligandos o moléculas). Las PIDs suelen desempeñar funciones que dependen de la unión a otras moléculas, pueden unirse a diversas dianas moleculares e incluso pueden adoptar estructuras diferentes en los distintos complejos finales. Pueden permitir transiciones entre diferentes estados conformacionales o estructurales. Los procesos de unión de las proteínas desordenadas están caracterizados por una baja afinidad, es decir, las uniones son generalmente débiles, pero en cambio son altamente específicas.

1.1.2. Desorden en proteínas

Las PIDs se pueden caracterizar mediante diversos métodos experimentales, cada uno con sus propios puntos fuertes y débiles. Sin embargo, esta caracterización resulta en la mayoría de los casos parcial. A continuación se describen las principales técnicas experimentales usadas.

La cristalografía y difracción de rayos-X, como se mencionó anteriormente, no puede detectar regiones desestructuradas en proteínas por su falta de densidad electrónica. La mayor incertidumbre

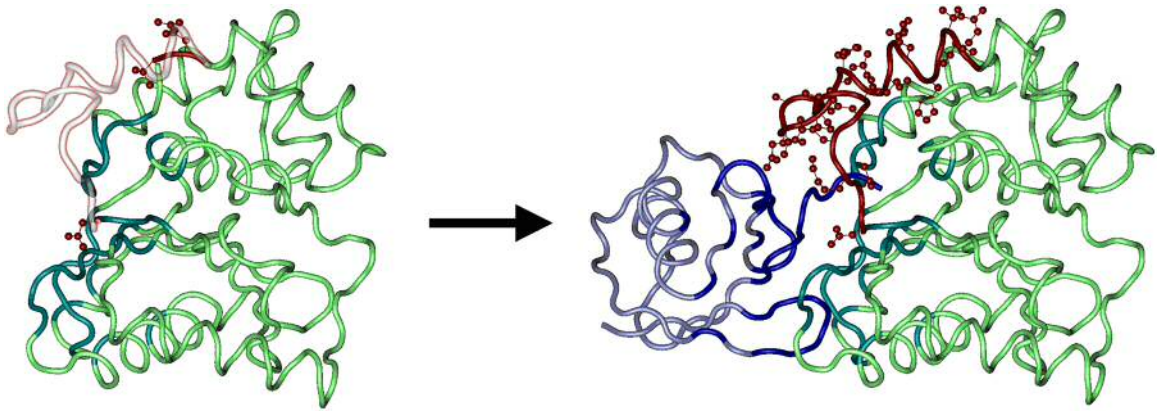

Figura 1.2: Transición desorden-orden en la formación del complejo ubiquitina C-terminal ubiquitina hidrolasa. En la figura de la izquierda se muestra el monómero ubiquitina hidrolasa (1UCH) con la región desordenada marcada en blanco y los aminoácidos que flanquean esta región en rojo. En la figura de la derecha se muestra el complejo formado entre la ubiquitina hidrolasa y la ubiquitina con los mismos aminoácidos marcados en rojo [**?**].

de esta técnica es que, sin experimentos adicionales, no se puede definir con total exactitud si una región con falta de densidad electrónica es un dominio PID o es el resultado de dificultades técnicas. Una metodología alternativa a la cristalografía de rayos-X convencional para estudiar proteínas PID es la dispersión de rayos X de ángulo pequeño (SAXS).

La resonancia magnética nuclear (RMN) puede resolver estructuras 3D de proteínas en solución. El hecho de que no sea necesario cristalizar la proteína para resolver su estructura 3D hace que esta técnica proporcione una estimación menos sesgada del desorden en comparación con la determinación por cristalografía y difracción de rayos-X. Bajo circunstancias favorables esta técnica proporciona información sobre la movilidad de cada residuo. Sin embargo, comparado con los resultados obtenidos con proteínas ordenadas, los datos relativos a proteínas PID son relativamente escasos. Esto indica que el estudio de proteínas PID con esta técnica también tiene dificultades. Las proteínas PID suelen formar agregados a las concentraciones necesarias para experimentos de RMN, además de presentar alta heterogeneidad con interconversiones estructurales en el orden de milisegundos que ocasionan un elevado ensanchamiento de los picos espectrales. Estas dificultades hacen que con la técnica de RMN los datos sobre desorden no sean tan abundantes.

La espectroscopía de dicroismo circular (CD) también puede proporcionar información estructural de las proteínas en solución. Los espectros de CD UV-lejano proporcionan estimaciones de estructura secundaria y pueden distinguir entre estructuras globulares ordenadas y bucles o giros flexibles en estado de glóbulo fundido (carente de estructura compacta globular). Por otro lado, el CD UV-cercano muestra picos estrechos para grupos aromáticos cuando se ordena la proteína, pero estos picos desaparecen en el estado de glóbulo fundido debido a la movilidad promedio de los átomos. La combinación del uso del CD UV-cercano y CD UV-lejano puede distinguir si se ordena una proteína o si se encuentra desordenada. Sin embargo, este método es sólo semicuantitativo y no proporciona información sobre aminoácidos específicos. Por tanto una limitación de esta técnica es que no proporciona una clara información para las proteínas que contienen tanto regiones ordenadas como desordenadas.

La digestión con proteasas proporciona indicios de la flexibilidad de una proteína estructurada, la

flexibilidad no es una mera exposición superficial, es el factor determinante para la búsqueda de sitios de corte de digestión. Estudios han demostrado que las regiones PID tienen una hipersensibilidad a las proteasas. Así, la digestión con proteasas es especialmente útil cuando se utiliza en combinación con otros métodos. Por ejemplo, la digestión con proteasas puede usarse con difracción de rayos-X, para resolver si una región con pérdida de densidad electrónica es debido a un estado de desorden. También se usa en combinación con el dicroísmo circular o la espectrometría de masas.

La espectroscopía de resonancia paramagnética electrónica (EPR) combinada con el marcaje sitio-dirigido de etiquetas de spin (SDSL) es una de las técnicas hoy en día más adecuadas para estudiar la estructura y dinámica de PIDs. La espectroscopía SDLS-EPR ha alcanzado actualmente un nivel que hace que su aplicación en este campo sea cada vez más extendido (Drescher 2012).

1.1.3. Predición de desorden en proteínas

Las técnicas experimentales que se han descrito anteriormente constituyen una herramienta muy valiosa para el estudio de PIDs, aunque presentan ciertas limitaciones. Una de ella es la identificación de este tipo de proteínas en estudios post-genómicos. En el caso del proteoma humano, hasta la fecha se han identificado unas 600 proteínas total o parcialmente desestructuradas y se ha descrito su función por medio de técnicas experimentales. Pero este número sólo constituye una pequeña parte del total de proteínas estimadas. Ante este panorama un enfoque bioinformático resulta indispensable para avanzar en la identificación y caracterización de PIDs.

Los primeros métodos bioinformáticos se basaron en los primeros estudios teóricos de proteínas individuales. Estos estudios sugerían que después de ser sintetizada una cadena de aminoácidos para producir una proteína, la cadena se pliega de una manera que depende de su composición. En concreto, los aminoácidos voluminosos e hidrofóbicos (los que repelen las moléculas de agua, que de forma natural rodean a las proteínas) se sitúan en el interior de la molécula proteíca. Por el contrario los aminoácidos que se colocan en la superficie plegada de la misma suelen ser pequeños e hidrofílicos (interaccionan con las moléculas de agua circundantes).

Así, en los inicios, el planteamiento subyacente a los métodos computacionales de predicción de desorden era comparar las secuencias de aminoácidos de las proteínas que se conocían como PIDs con las que presentaban formas plegadas rígidas. Los primeros cálculos realizados por Dunker en 1997 descubrieron que las PIDs presentaban mayor número de aminoácidos hidrofílicos que las proteínas rígidas o compactas. Por tanto la relación entre aminoácidos hidrofilicos e hidrofóbicos podría predecir el grado de desestructuración o desorden de una proteína concreta.

Los métodos computacionales desarrollados hasta la fecha se basan en distintos tipos de cálculos y aproximaciones. Según esto los podemos clasificar en 3 tipos: a) métodos basados en cálculos Ab-initio, donde las predicciones se apoyan solamente sobre la información de la composición de la secuencia proteíca y utilizan técnicas como redes neuronales o clasificadores bayesianos entre otros; b) métodos basados en moldes en los que se examinan estructuras (o no estructuras) de secuencias similares; c) métodos basados en meta-predicciones que combinan las predicciones de varios métodos computacionales.

A continuación nos detendremos en analizar los métodos basados en cálculos Ab-initio. Estos métodos utilizan medidas matemáticas de la composición proteíca. Las regiones desordenadas tienen baja complejidad por lo que los primeros métodos computacionales que se desarrollaron se basaron en medidas matemáticas que distinguen entre regiones de secuencia de proteínas globulares y

no globulares. Las estructuras globulares se caracterizan por ser compactas y estar determinadas por secuencias de aminoácidos de alta complejidad. Estas difieren ligeramente de las estructuras no globulares que contienen secuencias al azar aleatorio. Los algoritmos SEG (Wootton, 1994), CAST (Promponas et al. 2000) o GBA (Li and Kahveci, 2006) son métodos en general eficaces para discriminar entre regiones globulares y no globulares de forma automática. Los análisis estadísticos realizados en secuencias de proteínas muestran que aproximadamente una cuarta parte de los aminoácidos forman parte de regiones de baja complejidad y que más de la mitad de las proteínas tienen al menos una de estas regiones.

Otras características tenidas en cuenta en los cálculos Ab-initio son la hidrofobicidad y la carga neta de la proteína. El esqueleto proteíco y las cadenas laterales de las proteínas se mueven constantemente debido al movimiento térmico y a la energía cinética de los átomos. Este movimiento es dependiente del carácter hidrofóbico del segmento proteíco. Los métodos basados en esta propiedad emplean cálculos de la distribución de factores-B en las estructuras cristalinas. Los factores-B reflejan la fluctuación de los átomos sobre sus posiciones promedio y proporcionan información importante sobre la dinámica de la proteína. Estos cálculos usan regresiones vectoriales tipo SVR (Support Vector Regression). Los enfoques computacionales para predecir el movimiento térmico son útiles para el análisis de las propiedades dinámicas de proteínas con estructuras desconocidas y por tanto para PIDs. Las regiones desordenadas son raramente hidrofóbicas y presentan una alta probabilidad de fluctuación.

La carga neta influye en la capacidad de los polipéptidos de una proteína en formar contactos estabilizantes. En las proteínas globulares existen un gran número de interacciones inter-residuos, que aportan la energía estabilizadora para superar la pérdida de entropía durante el plegamiento. Por el contrario, en PIDs las secuencias no tienen la capacidad de formar suficientes interacciones inter-residuos. Las regiones desordenadas muestran una alta carga neta.

Las redes neuronales artificiales son también usadas para desarrollar métodos de predicción de desorden en proteínas. Estos métodos predicen, a partir de una secuencia de proteína, la probabilidad de encontrar segmentos desordenados.

1.1.4. Métodos bioinformáticos de predicción de desorden

En la última década más de una veintena de métodos bioinformáticos se han desarrollado para predecir regiones desordenadas o desestructuradas a partir de la secuencia de una proteína (ej. DisEMBL, DISOPRED2, DRIPPRED, DISpro, FoldIndex, GlobProt2, IUPred, PONDR, RONN, SPRITZ, entre otros). La clave del éxito de estos programas bionformáticos reside en que la secuencia de aminoácidos no sólo determina la estructura tridimensional de una proteína, sino también la ausencia de la misma. A continuación se detallan (en un orden arbitrario) algunos de ellos y se muestran, en algunos casos, los resultados que proporcionan en su formato de salida. Es interesante mencionar que estos métodos computacionales se pueden utilizar directamente a través de sus propios servidores web o bien se pueden descargar en servidores locales (según permisos de licencia). Las predicciones obtenidas por cada uno de estos métodos son difíciles de comparar entre sí debido a las diferencias existentes entre los parámetros y/o variables que manejan (ej. factores-B, regiones sin coordenadas en archivos PBD, etc). Los experimentos CASP[1] son una valiosa herramienta que permite obtener una validación de estos métodos.

[1]CASP. *http://predictioncenter.org*

DisEMBL

DisEMBL[2] es un método basado en redes neuronales artificiales que contempla los siguientes tres criterios para definir las regiones PID. Las regiones PID pueden ser: i) bucles flexibles (*loops/coils*) tal como los define Kabsch y Sander (1983). Los aminoácidos presentes en los bucles flexibles no son frecuentes en regiones estructuradas tales como hélices α (H), hélices 3_{10} (G) o láminas β (E). Sin embargo, es importante señalar que los bucles flexibles no necesariamente son regiones desordenadas o desestructuradas, aunque el desorden sólo se encuentra en estas regiones flexibles. ii) bucles "calientes" con un alto grado de movilidad determinada a partir de factores de temperatura C_α (factores-B). Este tipo de bucles son un subgrupo de los bucles flexibles pero con una alta dinámica añadida. iii) coordenadas en estructuras de rayos-X, definidas como entradas REMARK465 en archivos PDB (Protein Data Bank), sin mapa de densidad electrónica. Las regiones ausentes en archivos PDB se consideran por tanto desordenadas o desestructuradas. DisEMBL además proporciona un interfaz de tubería para predicciones a gran escala, esenciales, por ejemplo, a escala de genómica estructural (Linding et al. 2003).

- *Ejemplo:*

  La interfaz web es fácil de usar. En primer lugar hay que introducir la proteína problema, bien por su código (SWISS-PROTSWALL, ej. P61313) o entrada (ej. RL15_HUMAN), o bien por la secuencia de aminoácidos, ver **??**.

```
>sp|P61313|RL15_HUMAN 60S ribosomal protein L15 OS=Homo sapiens
  GN=RPL15 PE=1 SV=2
  MGAYKYIQELWRKKQSDVMRFLLRVRCWQYRQLSALHRAPRPTRPDKARRLGYKAKQGYV
  IYRIRVRRGGRKRPVPKGATYGKPVHHGVNQLKFARSLQSVAEERAGRHCGALRVLNSYW
  VGEDSTYKFFEVILIDPFHKAIRRNPDTQWITKPVHKHREMRGLTSAGRKSRGLGKGHKF
  HHTIGGSRRAAWRRRNTLQLHRYR
```

Figura 1.3: Sequencia de la proteína ribosomal L15 humana.

El resultado de salida muestra: i) una gráfica que representa la probabilidad de desorden a lo largo de la secuencia para cada uno de los criterios que definen las regiones PID (bucles flexibles, bucles "calientes" o entradas REMARK465 sin coordenadas en archivos PDB), **??**. Las líneas horizontales corresponden al nivel de expectativa aleatoria para cada predicción. ii) tres salidas de texto con las secuencias en las que están marcados los aminoácidos que cumplen cada uno de los criterios (**??**).

Código 1.1: Ejemplo de salida de DisEMBL

```
Disordered by Loops/coils definition
>none_LOOPS 34-57, 67-89, 116-129, 135-188
mgaykyiqel wrkkqsdvmr fllrvrcwqy rqlSALHRAP RPTRPDKARR
LGYKAKQgyv iyrirvRRGG RKRPVPKGAT YGKPVHHGVn qlkfarslqs
vaeeragrhc galrvLNSYW VGEDSTYKFf evilIDPFHK AIRRNPDTQW
ITKPVHKHRE MRGLTSAGRK SRGLGKGHKF HHTIGGSRra awrrrntlql
hryr
```

[2]DisEMBL. *http://dis.embl.de*

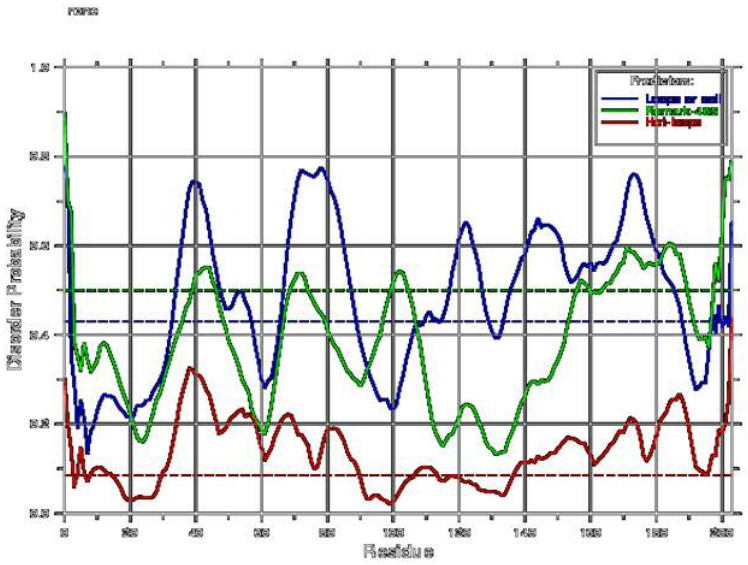

Figura 1.4: Gráfica de salida de DisEMBL

Disordered by Hot-loops definition
>none_HOTLOOPS 31-90, 138-204
mgaykyiqel wrkkqsdvmr fllrvrcwqy **RQLSALHRAP RPTRPDKARR**
LGYKAKQGYV IYRIRVRRGG RKRPVPKGAT YGKPVHHGVN qlkfarslqs
vaeeragrhc galrvlnsyw vgedstykff evilidp**FHK AIRRNPDTQW**
ITKPVHKHRE MRGLTSAGRK SRGLGKGHKF HHTIGGSRRA AWRRRNTLQL
HRYR

Disordered by Remark-465 definition
>none_REM465 165-190
mgaykyiqel wrkkqsdvmr fllrvrcwqy rqlsalhrap rptrpdkarr
lgykakqgyv iyrirvrrgg rkrpvpkgat ygkpvhhgvn qlkfarslqs
vaeeragrhc galrvlnsyw vgedstykff evilidpfhk airrnpdtqw
itkpvhkhre mrgl**TSAGRK SRGLGKGHKF HHTIGGSRRA** awrrrntlql
hryr

DISOPRED2

DISOPRED2[3] identifica regiones PID en base a la información obtenida de 750 secuencias no redundantes con estructuras de rayos-X de alta resolución. El desorden se identifica con aquellos aminoácidos que tienen coordenadas sin mapa de densidad electrónica. Para cada proteína problema se genera un perfil de secuencia usando el motor de búsqueda PSI-BLAST contra una base de datos de secuencia filtrada. El vector de entrada para cada aminoácido se construye a partir de los perfiles de una ventana simétrica de 15 posiciones (Ward et al. 2004). DISOPRED2 se diferencia de otros métodos en que está entrenado directamente sobre las secuencias proteícas no sobre la composición de aminoácidos.

- *Ejemplo:*

[3]DISOPRED2. *http://bioinf.cs.ucl.ac.uk/disopred*

La interfaz web es fácil de usar. En primer lugar hay que introducir la secuencia de aminoácidos de la proteína problema (**??**).

Los resultados de salida presentan diferentes formatos: i) una gráfica que representa la probabilidad de desorden a lo largo de la secuencia, **??**. ii) un archivo de texto con la secuencia donde se marcan con un asterisco los residuos desordenados (**??**). iii) un archivo de texto con todos los aminoácidos que contiene la secuencia y la probabilidad de desorden asociada (**??**).

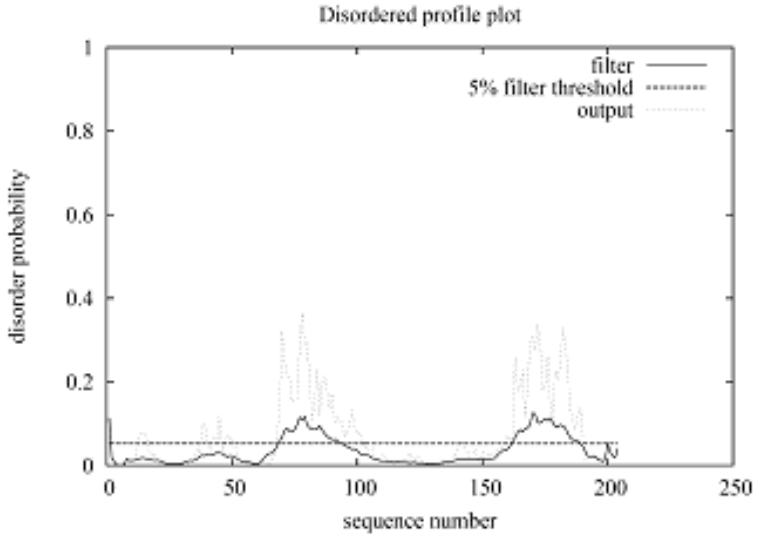

Figura 1.5: Gráfica de salida de DISOPRED2

Código 1.2: Ejemplo de salida de DISOPRED2

```
DISOPRED predictions for a false positive rate threshold of: 2%

conf: 8000000000000000000000000000000000000000000000000000000000000000
pred: *.............................................................
  AA: MGAYKYIQELWRKKQSDVMRFLLRVRCWQYRQLSALHRAPRPTRPDKARRLGYKAKQGYV
              10        20        30        40        50        60
conf: 0000001456787788888888888777666554433120000000000000000000000000
pred: ..........*..**********.....................................
  AA: IYRIRVRRGGRKRPVPKGATYGKPVHHGVNQLKFARSLQSVAEERAGRHCGALRVLNSYW
              70        80        90       100       110       120
conf: 0000000000000000000000000000000000001145677788888888888888
pred: ...................................................**********
  AA: VGEDSTYKFFEVILIDPFHKAIRRNPDTQWITKPVHKHREMRGLTSAGRKSRGLGKGHKF
             130       140       150       160       170       180
conf: 888766665300000000040002
pred: ***..................
  AA: HHTIGGSRRAAWRRRNTLQLHRYR
             190       200

Asterisks (*) represent disorder predictions and dots (.)
prediction of order. The confidence estimates give a rough
indication of the probability that each residue is disordered.
```

Código 1.3: Ejemplo de salida de DISOPRED2

```
        -----DISOPRED version 2-----
Disordered residues are marked with asterisks (*)
   Ordered residues are marked with dots (.)
Predictions at a false positive rate threshold of: 2%

    1 M *    0.114   0.114
    2 G .    0.018   0.018
    3 A .    0.009   0.009
    4 Y .    0.002   0.002
...
   72 K *    0.214   0.089
   73 R .    0.207   0.081
   74 P .    0.152   0.080
   75 V *    0.156   0.087
   76 P *    0.160   0.097
   77 K *    0.283   0.115
   78 G *    0.365   0.109
   79 A *    0.288   0.117
   80 T *    0.294   0.097
   81 Y *    0.201   0.090
   82 G *    0.109   0.087
   83 K *    0.151   0.086
   84 P *    0.230   0.086
   85 V *    0.118   0.095
   86 H .    0.197   0.085
   87 H .    0.210   0.076
...
```

FoldIndex

FoldIndex[4] predice si una secuencia de proteína es intrínsecamente desordenada en base al algoritmo propuesto por Uversky y col. (2000) que considera el promedio de la hidrofobicidad de cada residuo y la carga neta de la secuencia. FoldIndex© tiene una tasa de error comparable a la de los más sofisticados métodos de predicción. Usa ventanas correderas que permiten la identificación de grandes regiones dentro de una proteína con plegamientos diferenciables a los de la proteína completa (Prilusky 2005).

- *Ejemplo:*

  La interfaz web es fácil de usar. En primer lugar hay que introducir la secuencia de aminoácidos de la proteína problema (**??**).

  En la **??** y el **??** se puede ver el tipo de resultados que proporciona este programa.

Código 1.4: Ejemplo resumido de salida de FoldIndex

```
Number Disordered Regions:  2
Longest Disordered Region:  97
Number Disordered Residues: 162
Predicted disorder segment: [  1]-[ 97] length: 97 score: -0.32 ± 0.11
Predicted disorder segment: [140]-[204] length: 65 score: -0.27 ± 0.14

Las regiones PID están marcadas en rojo
```

[4]FoldIndex. *http://bip.weizmann.ac.il/fldbin/findex*

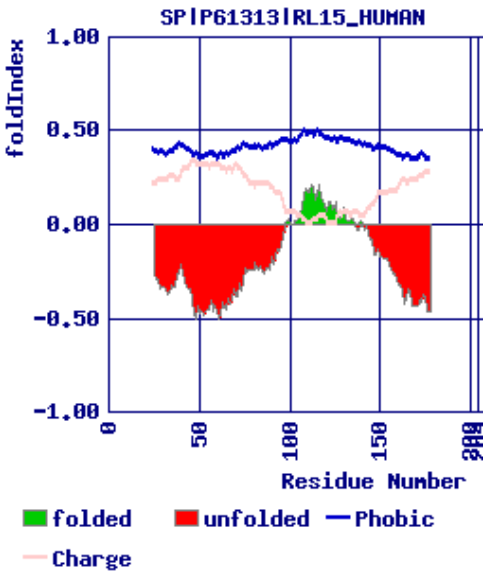

Figura 1.6: Gráfica de salida de FoldIndex

```
  1 MGAYKYIQEL WRKKQSDVMR FLLRVRCWQY RQLSALHRAP RPTRPDKARR
 51 LGYKAKQGYV IYRIRVRRGG RKRPVPKGAT YGKPVHHGVN QLKFARSLQS
101 VAEERAGRHC GALRVLNSYW VGEDSTYKFF EVILIDPFHK AIRRNPDTQW
151 ITKPVHKHRE MRGLTSAGRK SRGLGKGHKF HHTIGGSRRA AWRRRNTLQL
201 HRYR
```

IUPred

IUPred[5] se basa en la estimación de la capacidad de los polipéptidos en formar contactos estabilizantes. Las PIDs no tienen la capacidad de formar suficientes interacciones inter-residuos. Este método usa una expression cuadrática para la composición de aminoácidos que tiene en cuenta que la contribución de un aminoácido (tanto ordenado como desordenado) depende no sólo de sus propiedades químicas sino también de su entorno en la secuencia, incluyendo sus potenciales de interacción. Las energías de secuencias estimadas con regiones PID claramente se desplazan hacia energías menos favorables en comparación con proteína ordenadas (Dosztanyi et al, 2005).

IUPred, a diferencia de otros métodos, ofrece 3 tipos diferentes de predicción: i) segmentos desordenados largos (>30 residuos consecutivos); ii) segmentos desordenados cortos (<30 residuos consecutivos); iii) dominios estructurados. IUPred también incorpora la herramienta ANCHOR que predice regiones de interacción con otras proteínas o moléculas en el segmento de secuencia desordenada. Estas regiones funcionan como una via de transición desorden-orden durante la interacción con una proteína globular.

- *Ejemplo:*

  La interfaz web es fácil de usar. En primer lugar hay que introducir la secuencia de aminoácidos de la proteína problema (**??**).

[5]IUPred. *http://iupred.enzim.hu*

Los resultados de salida presentan diferentes formatos: i) una gráfica que representa la probabilidad de desorden a lo largo de la secuencia (**??**). ii) un archivo de texto con todos los aminoácidos que contiene la secuencia y la probabilidad de desorden asociada (**??**). Estos formatos son similares a los que produce DISOPRED2.

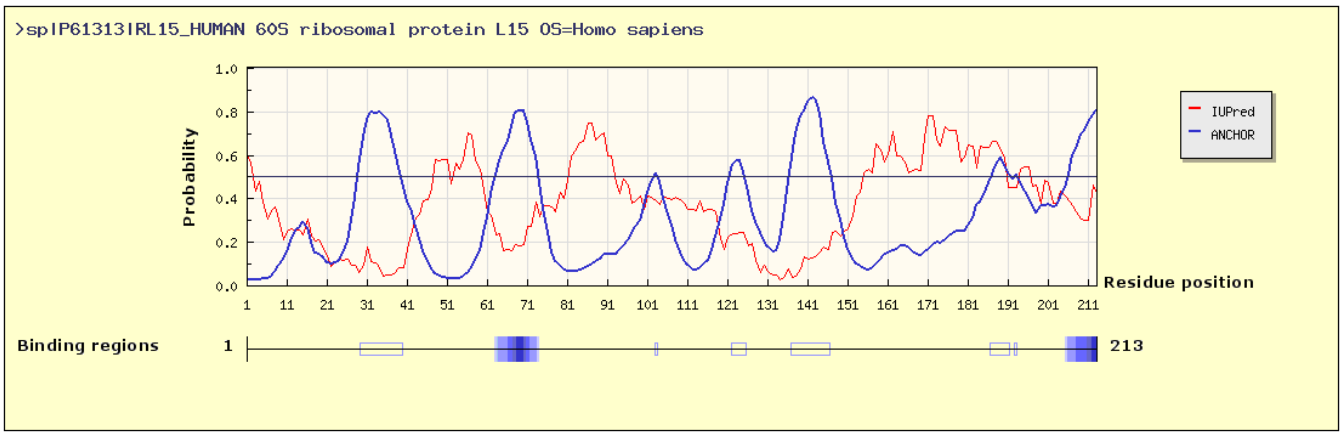

Figura 1.7: Gráfica de salida de IUPred.

Código 1.5: Ejemplo de salida de IUPred

```
>sp|P61313|RL15_HUMAN 60S ribosomal protein L15 OS=Homo sapiens
Predicted Disordered Binding Regions
  From      To    Length
1   63      73        11
2  206      213        8
```

PONDR

PONDR VL-XT [6] funciona a partir de datos de secuencia primaria solamente y está basado en cálculos de redes neuronales utilizando ventanas generalmente de 21 aminoácidos. Propiedades como la composición de aminoácidos y la hidropaticidad se calculan sobre estas ventanas, y los valores son entradas del programa. La red neuronal, que está ensayada para conjuntos específicos de secuencias ordenadas o desordenadas, devuelve un valor para el aminoácido central de la ventana. Las predicciones son refinadas sobre un ancho de ventana de 9 aminoácidos (Romero et al. 2001).

1.1.5. Composición, distribución y función de las proteínas desordenadas

Las PIDs, en general, se caracterizan por una baja complejidad en su secuencia de aminoácidos. Tienen un bajo contenido en aminoácidos de tipo hidrofóbico (Val, Leu, Ile, Met, Phe, Trp, Tyr), que suelen formar parte del esqueleto de proteínas globulares compactas, y tienen una alta proporción de aminoácidos cargados y polares (Gln, Ser, Pro, Glu, Lys, y ocasionalmente Gly y Ala). Tales regiones son muy frecuentes en proteínas reguladoras transcripcionales (factores de transcripción)

[6] IUPred. *http://www.pondr.com*

que hoy se reconocen como proteínas PID. Los análisis de datos de secuencia en genomas completos indican que las PIDs son altamente prevalentes y que su proporción aumenta con la complejidad de los organismos. La composición de nucleótidos de los genes que codifican para las PIDs tienen un elevado contenido GC (guanina y citosina). En bacterias valores altos de GC resultan en un aumento del contenido de Gly, Ala, Arg y Pro, mientras que contenidos bajos de GC derivan en un enriquecimiento de Phe, Tyr, Met, Ile, Asn y Lys. El primer grupo de aminoácidos se encuentra sobrerrepresentado en regiones PID, por lo que es de esperar que valores altos de GC resulten en un aumento significativo de desorden. Este tipo de correlación también se ha observado en organismos eucariotas superiores.

Las predicciones realizadas con diferentes métodos computacionales estiman que el 30-60 % de las proteínas en organismos eucariotas contienen segmentos desordenados de longitud superior a 30 aminoácidos. En Archaea y bacterias la prevalencia es menor (2-18 %). Estas predicciones se han realizado utilizando los proteomas completos disponibles (Dunker et al. 2000). Los porcentajes que se han calculado para los proteomas de plantas no difieren de los determinados en otros eucariotas. Sin embargo, cuando en plantas se examinan por separado los proteomas cloroplásticos, mitocondrial y nuclear se encuentran diferencias. Los proteomas cloroplástico (2-11 %) y mitocondrial (2-19 %) tienen mucho menos desorden que el nuclear, con valores similares a los presentes en Archaea y bacterias, de acuerdo con su origen filogenético. Por otra parte, es interesante señalar que cuando se examina el patrón de transferencia génica entre el cloroplasto y el núcleo durante la evolución, nos encontramos que los genes de origen cloroplástico, que son ahora codificados por genes nucleares, han adquirido desorden. Por tanto, la dinámica evolutiva del núcleo en plantas añade segmentos de desorden, a excepción de las proteínas que son codificadas en ambos genomas, debido posiblemente a restricciones funcionales (Yruela and Contreras-Moreira, 2012). Las PIDs también se encuentran en organismos más sencillos como los virus. Los fagos, virus especializados en infectar bacterias, se adhieren a la membrana de una célula huésped mediante proteínas que se mantienen unidas al cuerpo del fago por medio de regiones conectoras flexibles.

Las PIDs desempeñan funciones importantes y básicas en la célula, la mayoría asociadas con procesos de regulación, que incluyen la transcripción, la translación, la transducción de señal, la fosforilación, la regulación del ensamblaje de multicomplejos (ej. ribosoma) donde se requieren interacciones altamente específicas y de baja afinidad, entre otros. Por tanto, las PIDs están mayoritariamente asociadas a funciones reguladoras y de señalización, que son importantes en la comunicación celular y la respuesta celular a diversos estímulos. Estas funciones adquieren un mayor protagonismo en organismos eucariotas donde la complejidad de los sistemas celulares es superior. Esta observación puede explicar la correlación positiva entre desorden y complejidad.

La diversidad funcional de las PIDs complementa a la de las proteínas estructuradas. Las regiones PID participan en interacciones proteína-proteína, en el ensamblaje de complejos multi-proteícos y en múltiples actividades de las proteínas. Las proteínas con actividad chaperona tienen alta proporción de regiones PID. Lo mismo ocurre con las regiones de unión específica a ADN, como son los elementos cis, o los zinc-finger. Intuitivamente se puede pensar que una mayor flexibilidad y capacidad de interacción entre moléculas confiere a los organismos de una ventaja evolutiva. A mayor complejidad mayor capacidad de establecer interacciones. Esta característica permite la adaptación a diferentes condiciones del entorno, haciendo que la red de interacciones sea menos sensible a cambios ambientales y continúe su normal funcionamiento, facilita también la unión a diversas dianas y a su vez el control sobre la afinidad de esa unión, ajustando de esta forma el tiempo de transmisión de la señal según las necesidades. Las características dinámicas de las proteínas desordenadas aceleran el proceso de unión, lo que puede resultar crucial en las condiciones típicas de baja concentración de

proteínas involucradas en procesos de regulación. Así, es importante mencionar la hipótesis según la cual las proteínas desordenadas ayudan a la eficiente propagación de las señales celulares ya que la superficie de interacción (con otra molécula) por longitud de cadena peptídica es muy superior a la encontrada en las proteínas de estructura definida, debido de nuevo a su flexibilidad que da lugar a una cadena más expandida. Es importante matizar que la complejidad de los organismos es un fenómeno en el que intervienen múltiples parámetros, no sólo el desorden, sino que también el tamaño de los genomas, la capacidad de regulación por splicing alternativo, número de interacciones potenciales, especificidad de tejido, etc. son a considerar (Shad et al. 2011).

Las regiones PID también podemos encontrarlas conectando dominios o módulos estructurados, en este caso se trata de fragmentos significativamente largos (¿30 aminoácidos) carentes de estructura. En algunos casos estas regiones participan activamente en la función de la proteína entera por estar directamente involucradas en la zona de interacción. En otros casos estas regiones, por un lado, incrementan la movilidad de los dominios estructurados, y por otro, establecen una orientación predeterminada entre ellos, modulando de forma pasiva las posibilidades de interacción de los mismos. Se ha propuesto que los segmentos desordenados de unión inter-dominio incrementan la velocidad con la que se producen grandes cambios conformacionales entre los módulos, facilitando de nuevo la transmisión de señales dentro del entorno

Para terminar esta sección comentaremos algunos resultados recientes obtenidos de estudios bioinformáticos a nivel de genómica comparada. Estos estudios permiten clasificar las regiones desordenadas en tres tipos: a) regiones donde el desorden está conservado entre organismos pero la secuencia de aminoácidos rápidamente cambia por procesos evolutivos (desorden flexible); b) regiones de desorden conservado con alta conservación de la secuencia (desorden conservado); c) desorden no conservado. El primer tipo estaría asociado principalmente a rutas de señalización celular y multifuncionalidad y el segundo a procesos de unión a ARN y a proteínas de tipo chaperona. Por el momento se desconoce la relevancia del tercer tipo (Bellay, 2011).

1.1.6. Enfermedades asociadas a proteínas desordenadas

Numerosas PIDs están relacionadas con diversas enfermedades, algunas neurogenerativas (alzhéimer y parkinson), cáncer, cardiovasculares, diabetes, encefalopatías espongiformes transmisibles. Mediante estudios bioinformáticos se ha encontrado que el 79 % de las proteínas asociadas con el cáncer contienen regiones desordenadas de más de 30 aminoácidos (Iakoucheva et al. 2002). Por el contrario, sólo el 13 % de proteínas de un conjunto con estructuras ordenadas bien definidas contenían tales regiones de desorden predichas. La presencia de desorden en varias proteínas relacionadas con el cáncer se ha observado experimentalmente. Algunos ejemplos son: la proteína p53 (que participa en la red de señalización que regula la expresión de genes), AFP (alfa-Fetoproteína que participa en la regulación de la división célula), BRCA1 (proteína de la susceptibilidad al cáncer de mama), miembros de la familia Bcl-2 (implicadas en la muerte celular programada). El ejemplo más estudiado es la proteína supresora de tumores p53 que realiza su función reguladora interaccionando con otras múltiples proteínas. Aproximadamente el 70 % de esas interacciones está mediado por sus regiones desordenadas, bien a través de mutaciones o por cualquier otro factor. Si p53 pierde su función, la célula típicamente se convierte en cancerosa.

Los análisis realizados en otras patologías han revelado resultados parecidos a los obtenidos en cáncer. Estudios bioinformáticos en los proteomas de los virus del papiloma humano indican que

las proteínas de los virus considerados de alto riesgo para el desarrollo de carcinomas contienen más regiones desordenadas que las proteínas homólogas en virus no malignos.

En el caso de enfermedades cardiovasculares se ha calculado que un 61 % de las proteínas relacionadas son PIDs. Este porcentaje es próximo al calculado para proteínas de señalización (66 %) y es significativamente más alto que el promedio calculado de PIDs en organismos eucariotas (30-60 %). Este alto porcentaje de PIDs sugiere que podrían ser esenciales para la función de las proteínas que intervienen en los procesos relacionados con esta enfermedad, además de para su control y regulación. Los datos disponibles sobre PIDs en enfermedades cardiovasculares indican que hay una buena correlación entre las observaciones experimentales realizadas y los resultados de predicción.

Una de las características de la diabetes de tipo II (diabetes mellitas) es la formación de depósitos amiloides en los islotes de Langerhans del páncreas como respuesta a la disminución progresiva de la acción de la insulina. Esto produce un aumento inicial de la cantidad de hormona, aunque posteriormente decae provocando el aumento de los niveles de glucosa en sangre. La proteína amilina es el componente principal de estos agregados amiloides. Se trata de una pequeña proteína totalmente desordenada que sufre un cambio conformacional y adquiere cierto grado de estructura para posteriormente dar lugar a la formación de los depósitos.

Las encefalopatías espongiformes transmisibles se producen por la acumulación de agregados de una proteína llamada Prion. Un ejemplo de este tipo de patología es la encefalopatía espongiforme bovina o enfermedad de las vacas locas. La proteína Prion consta de dos dominios, uno estructurado y otro desordenado. El primero sufre un profundo cambio conformacional que es parcialmente responsable de los fenómenos de agregación. En los procesos de interacción y unión con las proteínas Prion el dominio desestructurado juega un papel importante.

El Alzheimer está asociado a la acumulación de depósitos proteicos con diferentes características morfológicas conocidos como depósitos amiloides, placas seniles y ovillos neurofribrilares. Las placas seniles se generan por la agregación de las proteínas amiloide β y τ. La proteína amiloide β antes de la agregación carece de estructura y sufre un proceso de compactación previo a la asociación. En el estado agregado se convierte en neurotóxica. Algo similar le ocurre a la proteína τ. Antes de la agregación en ovillos la proteína τ es mayoritariamente desordenada y se agrega tras sufrir un proceso de plegamiento parcial. Otras enfermedades neurodegenerativas, como las sinucleinopatías se caracterizan también por la formación de agregados fibrilares, concretamente de la proteína alfa-sinucleína. Se ha comprobado que en condiciones fisiológicas la proteína alfa-sinucleína está desordenada casi en su totalidad. Cuando varían las condiciones de pH y temperatura, la alfa-sinucleína es capaz de variar su conformación adoptando diversos grados de estructuración y de agregación. Los agregados pueden presentar morfologías muy variables; esferas, fibras y cúmulos amorfos.

Las interconexiones entre el desorden intrínseco, la señalización celular y enfermedades humanas sugieren que las enfermedades conformacionales pueden deberse no sólo al plegamiento anómalo de proteínas, sino también a errores en la identificación y la señalización, así como a fenómenos de plegamiento no natural o no nativo. La mayoría de proteínas que conocemos tienen una determinada conformación y a menudo llevan a cabo una única función. Las PID, sin embargo, son multifuncionales. Al no tener prácticamente estructura son muy flexibles lo que les permite interactuar con otras muchas proteínas del organismo, de aquí que ocupen posiciones clave dentro de las células. Entender su dinamismo, saber con qué proteínas se relacionan y saber de qué manera lo hacen es clave para poder avanzar en el diseño de fármacos específicos. En los últimos años se está realizando un importante esfuerzo para descubrir pequeñas moléculas que puedan reconocer proteínas PID y modificar su función. En este sentido las proteínas PID presentan dos problemas críticos para el

desarrollo de drogas eficaces. Debido a la plasticidad en su capacidad de unión, una molécula podría unirse a la proteína diana y a varias otras cuya función no se desea alterar. También es posible que la proteína diana reconozca diversas zonas de unión en la misma proteína que modifiquen la función con resultados diferentes. Por tanto se necesitan realizar estudios en profundidad para resolver estas cuestiones.

El descubrimiento de moléculas capaces de inhibir con cierta especificidad la interacción entre las proteínas c-Myc y Max es un hecho esperanzador para avanzar en el diseño de fármacos. Ambas proteínas son factores de transcripción. En particular, c-Myc activa la expresión de hasta el 15 % de los genes humanos. La función de c-Myc afecta por tanto importantes procesos biológicos como la proliferación celular, la diferenciación y la muerte celular programada. La disfunción de c-Myc es responsable de numerosos tipos de cáncer humano. Max es capaz de homodimerizar y de interaccionar con otros factores de transcripción como c-Myc. El complejo c-Myc/Max tiene afinidad por determinadas zonas del ADN y esta unión favorece el proceso de transcripción. c-Myc y Max son proteínas desordenadas en su estado libre que adquieren estructura al unirse entre sí. Conseguir inhibir la formación del complejo y alterar su función es un objetivo importante para futuras terapias contra el cáncer.