

# Bioinformática Estructural

---

## Resumen

La biología estructural permite el estudio de las estructuras de macromoléculas, el origen de esta estructura y su relación con la función biológica. La función específica está íntimamente ligada a la conformación tridimensional; la configuración estructural de las biomoléculas depende a su vez de su composición básica. Por ello, se busca obtener y comprender diferentes resultados de modelado de estructuras proteicas para resolver problemas biológicos. También es importante comprender las bases teóricas, tanto conceptuales como algorítmicas, de la predicción y análisis de estructura de macromoléculas, e interpretar los resultados de estos programas.

Apuntes adaptados a partir de los proporcionados por Modesto Redrejo en <https://strbio.github.io/>

Sandra Mingo Ramírez

UAM - 2024/25

5 de febrero de 2025 17:00

Universidad Autónoma de Madrid  
Bioinformática y Biología Computacional

[Código en Github](#)

# Índice general

|            |   |           |
|------------|---|-----------|
| <b>I</b>   | <b>Introducción y modelado de proteínas</b>   | <b>2</b>  |
| <b>I</b>   | <b>Aplicaciones y métodos de bioinformática estructural en biología y biomedicina</b> | <b>3</b>  |
| I.1        | Metas en la bioinformática estructural . . . . .                                      | 3         |
| I.2        | Introducción a las estructuras proteicas . . . . .                                    | 4         |
| I.2.1      | Gráfico de Ramachandran . . . . .   | 9         |
| I.2.2      | Pliegues (folds), dominios y motivos de proteínas . . . . .                           | 11        |
| <b>II</b>  | <b>Bases de datos de proteínas</b>  | <b>12</b> |
| II.1       | Comparación de estructura y alineamiento . . . . .                                    | 12        |
| II.2       | Principales bases de datos de proteínas . . . . .                                     | 13        |
| II.2.1     | Bases de datos estructurales . . . . .  | 15        |
| II.2.2     | Bases de datos de secuencias . . . . .  | 16        |
| II.3       | Estrategias actuales y futuras en las bases de datos de proteínas . . . . .           | 18        |
| <b>III</b> | <b>Estructuras de proteínas</b>   | <b>20</b> |
| III.1      | Obtener y trabajar con estructuras de proteínas . . . . .                             | 20        |
| III.2      | Determinación experimental de las estructuras de proteínas . . . . .                  | 20        |
| III.2.1    | Cristalografía de rayos X o difracción de rayos X de un solo cristal . . . . .        | 21        |
| III.2.2    | Resonancia magnética nuclear . . . . .  | 22        |
| III.2.3    | Criomicroscopía electrónica . . . . .   | 23        |
| III.3      | Garantía de calidad estructural . . . . .   | 24        |
| III.3.1    | Parámetros globales en estructuras basadas en experimentos . . . . .                  | 25        |
| III.3.2    | Parámetros estereoquímicos . . . . .  | 26        |
| III.4      | Visualización de estructuras de proteínas . . . . .                                   | 26        |
| III.4.1    | Formatos de ficheros de estructuras proteicas . . . . .                               | 26        |
| III.4.2    | Ocupancia y factor B . . . . .  | 27        |
| III.4.3    | Aplicaciones de visualización de macromoléculas biológicas . . . . .                  | 28        |

## **Parte I**

# **Introducción y modelado de proteínas**

# Capítulo I

## Aplicaciones y métodos de bioinformática estructural en biología y biomedicina

### I.1. Metas en la bioinformática estructural

La bioinformática estructural (SB por sus siglas en inglés) es una disciplina amplia que abarca recursos de datos, algoritmos y herramientas para investigar, analizar, predecir e interpretar estructuras biomacromoleculares. En este curso, nos centraremos específicamente en la bioinformática estructural de proteínas, incluyendo la visualización y el análisis de la estructura de biomacromoléculas, así como la predicción de estructuras y complejos de proteínas. La premisa de la SB es que la información estructural de alta resolución sobre los sistemas biológicos permite un razonamiento preciso sobre sus funciones y los efectos de las modificaciones y perturbaciones.

Los objetivos de SB requieren al menos cuatro líneas de investigación diferentes:

- **Visualización:** Tratar con una o muchas estructuras complejas e integrar varias fuentes de información como secuencias, datos estructurales, campos electrostáticos, localizaciones de sitios funcionales y áreas de variabilidad.
- **Clasificación:** Agrupación jerárquica de estructuras similares para identificar orígenes comunes y vías de diversificación. Al igual que en otros campos de la biología, la clasificación es tediosa pero necesaria para comprender el espacio estructural.
- **La predicción:** de estructuras sigue siendo un área de gran interés y un campo de investigación en sí mismo. Como veremos a continuación, el número de secuencias diferentes es mucho mayor que la disponibilidad de estructuras, lo que hace de la predicción una herramienta esencial y útil.
- **Simulación:** Las estructuras obtenidas experimentalmente son ante todo modelos estructurales estáticos. Sin embargo, las propiedades de estas moléculas son a menudo el resultado de sus movimientos dinámicos. La definición de las funciones energéticas que rigen el plegamiento de las proteínas y su posterior

dinámica estable pueden analizarse mediante simulaciones de dinámica molecular, aunque las capacidades de cálculo pueden ser limitantes para alcanzar escalas de tiempo biológicamente relevantes.

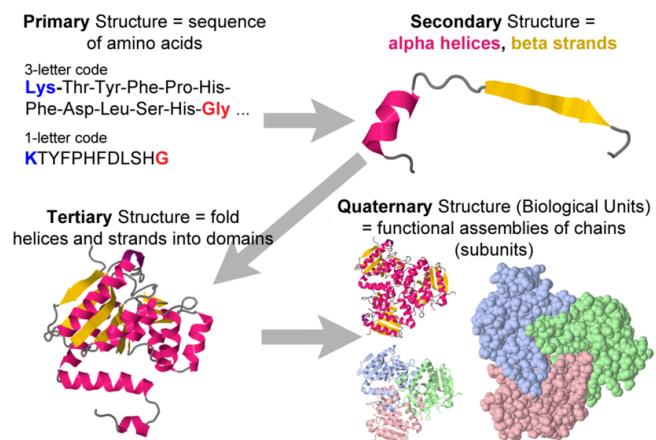
Impulsado por enormes cantidades de datos e importantes avances técnicos, este campo ha experimentado una transformación sustancial en los últimos veinte años. La mejora de las capacidades experimentales para analizar la estructura de las proteínas y otras moléculas y estructuras biológicas y el avance de la predicción de estructuras asistida por Inteligencia Artificial (IA) han aumentado sustancialmente la capacidad de los investigadores de las ciencias de la vida para abordar diversas cuestiones relativas a la diversidad, evolución y función de las proteínas. Esta transformación se ha potenciado en los últimos 5 años, y sus implicaciones para la biología, la biotecnología y la biomedicina siguen siendo en gran medida impredecibles.

## I.2. Introducción a las estructuras proteicas

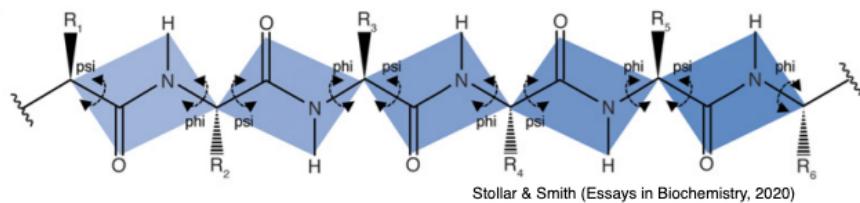
Las proteínas son componentes esenciales de la vida, que intervienen en diversas funciones vitales como elementos estructurales, elementos de andamiaje o enzimas activas que catalizan reacciones metabólicas. Las proteínas están compuestas por polímeros de aminoácidos, y la secuencia de aminoácidos de una proteína concreta se denomina **estructura primaria** de la proteína. Las cadenas de aminoácidos pueden plegarse espontáneamente en estructuras tridimensionales, estabilizadas principalmente por enlaces de hidrógeno entre aminoácidos. La secuencia de aminoácidos determina las diferentes capas de la estructura tridimensional. En la naturaleza existen L-aminoácidos, pero no D-aminoácidos. Cada uno de los 20 aminoácidos naturales tiene propiedades fisicoquímicas específicas que influyen en su conformación preferida. Por lo tanto, el nivel inicial de plegamiento se conoce como **estructura secundaria**, que forma patrones comunes como se verá más adelante. Estos segmentos de patrones de estructura secundaria son capaces de plegarse en formas tridimensionales debido a las interacciones entre las cadenas laterales de los aminoácidos, lo que se conoce como **estructura terciaria** de la proteína. Además, dos o más cadenas peptídicas individuales pueden agregarse para formar proteínas multisubunidad, lo que se conoce como **estructura cuaternaria**.

Es importante señalar que el enlace peptídico en sí no permite la rotación, ya que posee características parciales de doble enlace. Por lo tanto, la rotación está restringida a los enlaces entre el  $C\alpha$  y el grupo  $C = O$  (el ángulo phi ( $\phi$ )) y el  $C\alpha$  y el grupo  $NH$  (el ángulo psi ( $\psi$ )). Así pues, la cadena principal del polipéptido consiste en una secuencia repetida de dos enlaces giratorios seguidos de un enlace no giratorio (péptido). Sin embargo, no todos los  $360^\circ$  de los ángulos  $\phi$  y  $\psi$  son factibles debido a posibles **choques estéricos** entre cadenas laterales vecinas. Para determinados ángulos y combinaciones de aminoácidos, las restricciones espaciales impiden que los átomos ocupen la misma ubicación física, lo que explica en parte las distintas propensiones de ciertos aminoácidos a adoptar diferentes tipos de estructuras secundarias.

Además, las cadenas laterales de los aminoácidos poseen sus propios ángulos de torsión, conocidos como  $\chi_1$ ,  $\chi_2$ ,  $\chi_3$ , etc (figura I.3). Estos ángulos de torsión influyen significativamente en las estructuras secundarias y, sobre todo, terciarias de



**Figura I.1:** Los distintos niveles de la estructura proteica.

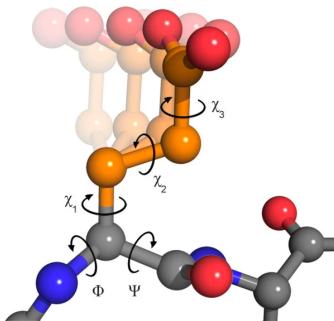


**Figura I.2:** Esquema de una cadena polipeptídica genérica. Las cadenas laterales de los residuos se denotan como R. Los rectángulos coloreados indican conjuntos de seis átomos que son coplanares debido al carácter de doble enlace del enlace peptídico. Las flechas indican los enlaces que son libres de rotar con el ángulo de rotación sobre el N-C<sub>α</sub> conocido como phi ( $\phi$ ) y sobre el C<sub>α</sub>-C conocido como psi ( $\psi$ ). Obsérvese que sólo se etiquetan los enlaces del esqueleto peptídico y que, en la mayoría de los casos, el enlace del grupo R es libre de rotar.

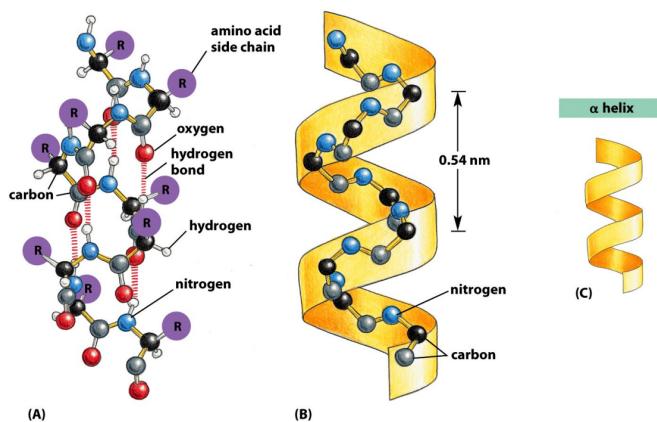
las proteínas. Las distintas combinaciones de torsiones de las cadenas laterales definidas por los ángulos  $\chi$  se denominan **rotámeros**.

Dentro de estas limitaciones, las dos conformaciones locales primarias que evitan el impedimento estérico y maximizan el enlace de hidrógeno entre la columna vertebral y el backbone son las estructuras secundarias  $\alpha$ -hélice y  $\beta$ -hoja. Linus Pauling propuso inicialmente la hélice  $\alpha$  como zurda en 1951, pero la estructura cristalina de la mioglobina en 1958 reveló que la forma diestra es más común. En las hélices diestras típicas, el grupo NH de la espina dorsal se une mediante enlaces de hidrógeno al grupo C=O de la espina dorsal del aminoácido situado cuatro residuos antes en la secuencia de la proteína. Esta forma de espiral regular tiene los grupos R apuntando hacia fuera, lejos de la espina dorsal peptídica, y requiere unos 3,6 residuos para completar una vuelta completa de la hélice (figura I.4).

Las diferentes secuencias de aminoácidos tienen distintas tendencias a formar estructuras  $\alpha$ -helicoidales. La metionina, la alanina, la leucina, el glutamato y la lisina tienen propensiones especialmente altas a formar hélices, mientras que la prolina y la glicina tienen propensiones pobres a formar hélices. La prolina a menudo rompe o retuerce una hélice porque carece de un hidrógeno amida para formar enlaces de hidrógeno y su voluminosa cadena lateral interfiere con el esqueleto del giro precedente. La glicina, con sólo un hidrógeno como grupo R, es demasiado flexible y costosa desde



**Figura I.3:** Ángulos diedros en el glutamato: Los ángulos diedros son los principales grados de libertad de la columna vertebral (ángulos  $\phi$  y  $\psi$ ) y la cadena lateral (ángulos  $\chi$ ) de un aminoácido. El número de ángulos  $\chi$  varía entre cero y cuatro para los 20 aminoácidos estándar. La figura muestra una representación esférica del glutamato, que tiene tres  $\chi$  ángulos.



**Figura I.4:** Hélice alfa.

el punto de vista entrópico para mantener la estructura  $\alpha$ -helicoidal, lo que la convierte en una rompedora de hélices  $\alpha$ .

Las láminas  $\beta$  (figura I.6) están formadas por dos o más cadenas polipeptídicas extendidas denominadas hebras  $\beta$  que discurren una junto a otra en disposición paralela o antiparalela. En una lámina  $\beta$ , los residuos se disponen en zigzag y los enlaces peptídicos adyacentes apuntan en direcciones opuestas. El grupo NH y el grupo C=O de cada aminoácido forman enlaces de hidrógeno con el grupo C=O y el grupo NH, respectivamente, de las cadenas adyacentes. Las cadenas pueden ir en direcciones opuestas (lámina  $\beta$  antiparalela) o en la misma dirección (lámina  $\beta$  paralela). Las cadenas laterales de cada residuo se alternan en direcciones opuestas, dando a las láminas  $\beta$  caras hidrofílicas e hidrofóbicas, formando a menudo un patrón de alternancia de residuos hidrofílicos e hidrofóbicos en la estructura primaria.

Los residuos aromáticos grandes (tirosina, fenilalanina, triptófano) y los aminoácidos  $\beta$ -ramificados (treonina, valina, isoleucina) suelen encontrarse en las hebras  $\beta$ . Como en el caso de las hélices  $\alpha$ , las hebras  $\beta$  suelen estar terminadas por glicinas, que son especialmente comunes en los giros  $\beta$  (el conector más común entre hebras), como aminoácidos con ángulos  $\phi$  positivos.

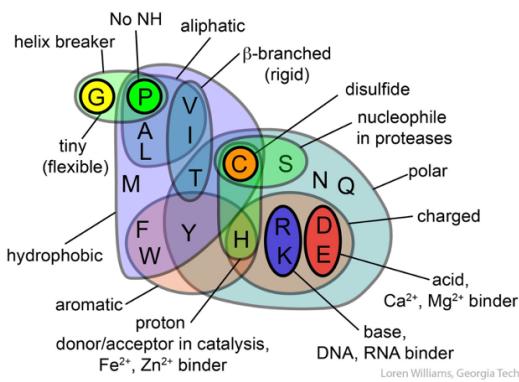


Figura 1.5: Aminoácidos clasificados según su tipo.

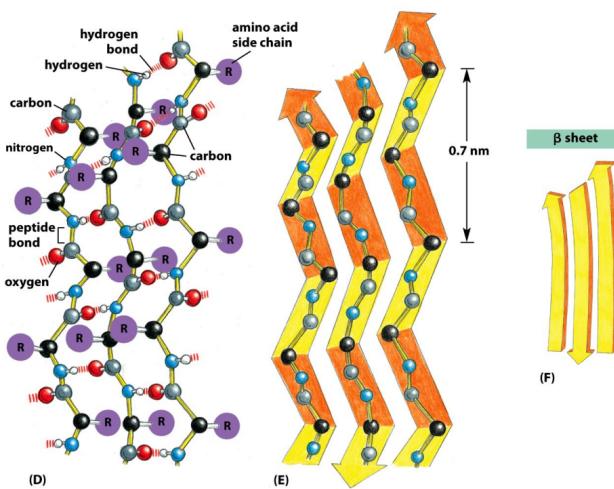


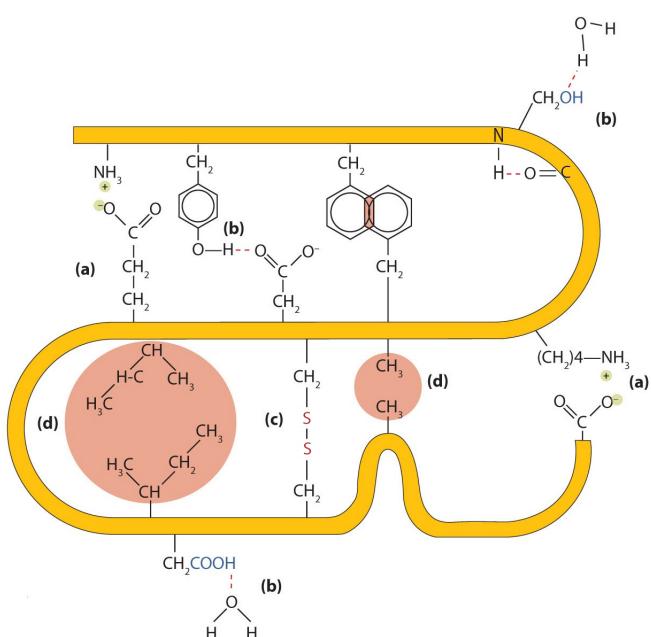
Figura 1.6: Descripción detallada de una lámina beta formada por tres hebras beta.

Las estructuras secundarias, terciarias y cuaternarias de las proteínas se mantienen gracias a las interacciones entre aminoácidos (figura 1.7). Estas interacciones suelen clasificarse en cuatro tipos, que pueden ser tanto intra- como intermoleculares:

- Enlace iónico:** Los enlaces iónicos surgen de las atracciones electrostáticas entre cadenas laterales de aminoácidos cargadas positiva y negativamente. Por ejemplo, la atracción entre un ion carboxilato del ácido aspártico y un ion amonio de la lisina ayuda a estabilizar una región plegada específica de una proteína.
- Enlace de hidrógeno:** Los enlaces de hidrógeno se forman entre un átomo de oxígeno o nitrógeno altamente electronegativo y un átomo de hidrógeno unido a otro átomo de oxígeno o nitrógeno, como los de las cadenas laterales de aminoácidos polares. Los enlaces de hidrógeno son cruciales para las interacciones intra e intermoleculares en las proteínas, como en las hélices alfa.
- Enlaces disulfuro.** Cuando dos aminoácidos cisteína se acercan durante el plegamiento de la proteína en condiciones redox adecuadas, la oxidación puede unir sus átomos de azufre, formando un enlace disulfuro. A diferencia de los enlaces iónicos o de hidrógeno, se trata de enlaces covalentes, por lo que son un ejemplo clásico de reacción espontánea, que se produce como modificación

posttraduccional. Aunque son sensibles a los agentes reductores, estabilizan en gran medida la estructura terciaria y son vitales para la estructura cuaternaria de muchas proteínas, como los anticuerpos.

**4. Interacciones hidrofóbicas.** Las fuerzas de dispersión surgen cuando un átomo normalmente no polar se convierte momentáneamente en polar debido a una distribución desigual de electrones, dando lugar a un dipolo instantáneo que induce un desplazamiento de electrones en un átomo no polar vecino. Las fuerzas de dispersión son débiles, pero pueden ser importantes cuando otros tipos de interacciones no existen o son mínimas. El término interacción hidrofóbica suele utilizarse erróneamente como sinónimo de fuerzas de dispersión. Las interacciones hidrofóbicas surgen porque las moléculas de agua establecen enlaces de hidrógeno con otras moléculas de agua (o grupos de proteínas capaces de establecer enlaces de hidrógeno). Como los grupos no polares no pueden formar enlaces de hidrógeno, la proteína se pliega de tal forma que estos grupos quedan enterrados en la parte interior de la estructura proteica, minimizando su contacto con el agua.

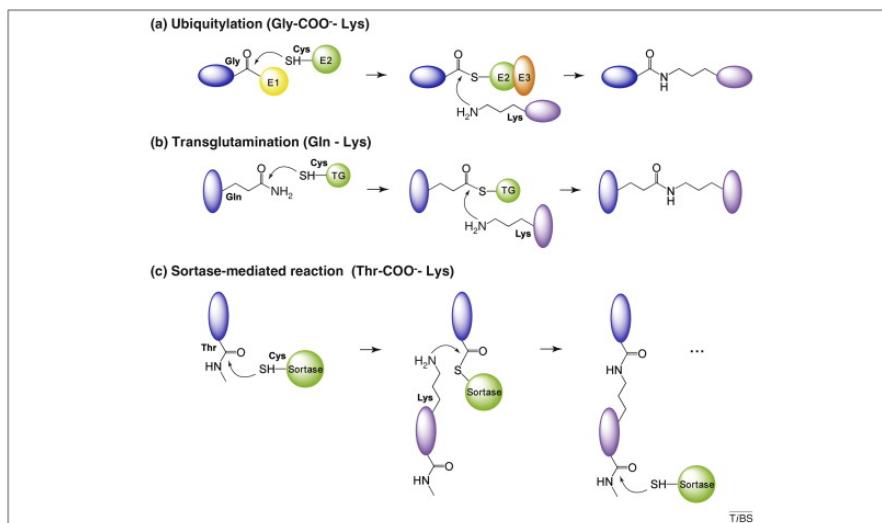


**Figura I.7:** Cuatro interacciones estabilizan la estructura terciaria de una proteína: (a) enlace iónico, (b) enlace de hidrógeno, (c) enlaces disulfuro y (d) fuerzas de dispersión.

Otras interacciones intramoleculares menos frecuentes podrían ser relevantes en algunas proteínas, como los llamados enlaces isopéptidos, formados entre dos grupos proteicos, al menos uno de los cuales no es un grupo  $\alpha$ -amino o  $\alpha$ -carboxi. Algunos ejemplos son la ubiquitilación, la sumoilación, la transglutaminación, el anclaje de proteínas a la superficie celular mediado por sortasas y la formación de pilus. Todos estos procesos comparten varias características (figura I.8):

- Todos implican la reacción de un grupo  $\epsilon$ -amino de la lisina de una proteína con el grupo  $\alpha$ -carboxi principal de otra proteína, excepto en el caso de la transglutaminación, en la que la lisina se dirige a un grupo carboxiamida de la cadena lateral de la glutamina.

- Todos los procesos están mediados por enzimas e implican un intermediario tioéster transitorio formado por la cisteína del sitio activo. Este intermediario se resuelve mediante un ataque nucleofílico por el grupo  $\varepsilon$ -amino de la lisina, que completa la formación del enlace.



**Figura I.8:** Formación de enlaces isopeptídicos intermoleculares mediada por enzimas. Se muestran ejemplos de tres procesos biológicos diferentes: ubiquitilación, transglutaminación y ensamblaje de pilus mediado por sortasa en bacterias Gram positivas. Las proteínas unidas por enlaces isopeptídicos están coloreadas en azul y morado y las enzimas formadoras de enlaces isopeptídicos en verde.

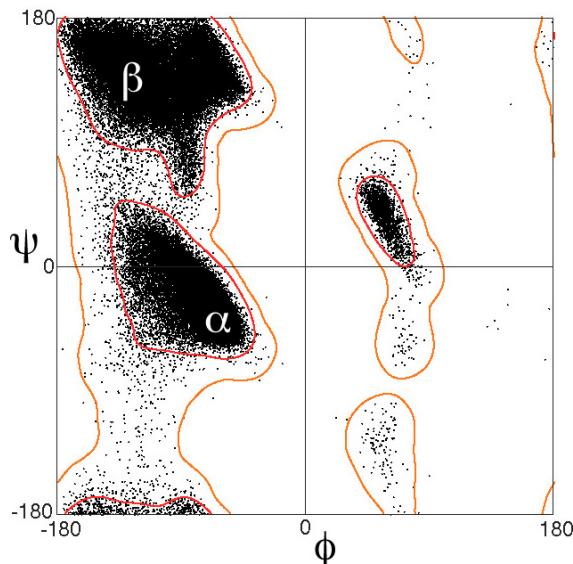
A diferencia de estos procesos dependientes de enzimas, los enlaces isopeptídicos entrecruzados (intrachain isopeptide bonds) se forman autocatalíticamente en la pilina principal Spy0128 de *S. pyogenes* y en otras proteínas de la superficie de células Gram+, así como en la cápside del fago HK97. En este caso, la reacción de formación del enlace es una reacción inducida por la proximidad que se produce cuando los aminoácidos participantes se sitúan juntos en un entorno hidrofóbico, ya sea a través del plegamiento de la proteína concurrente con la formación del enlace peptídico en el ribosoma o por la reorganización de la cápside (en HK97).

En cuanto a la ingeniería proteica, es posible crear aminoácidos no naturales reactivos. Esto se ha utilizado para aumentar la termoestabilidad de proteínas como anticuerpos, crear recombinantes y unir covalentemente proteínas a superficies o nanopartículas.

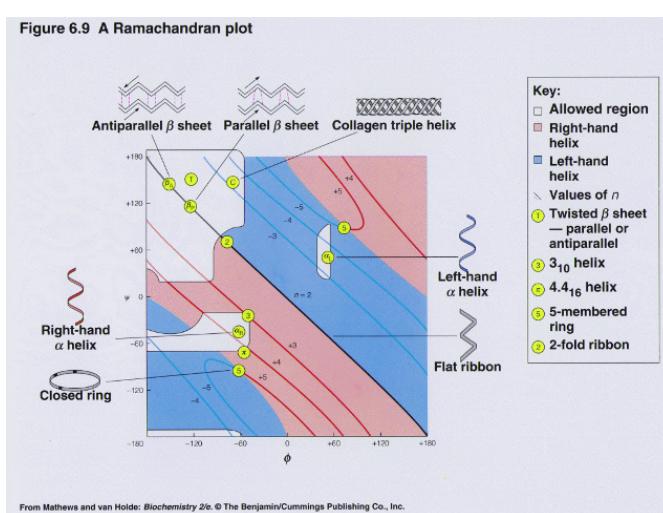
### I.2.1. Gráfico de Ramachandran

Muchas combinaciones de ángulos  $\phi$  y  $\psi$  están prohibidas debido al principio de exclusión estérica, que dicta que dos átomos no pueden ocupar el mismo espacio simultáneamente. Este concepto fue demostrado inicialmente por Gopalasamudram Ramachandran, que desarrolló un gráfico para visualizar los valores de ángulo permitidos, conocido como gráfico de Ramachandran. Este gráfico puede mostrar los ángulos de un aminoácido específico, de todos los aminoácidos de una proteína o incluso de muchas proteínas. El análisis de los ángulos  $\phi$  y  $\psi$  en proteínas conocidas

revela que aproximadamente tres cuartas partes de todas las combinaciones posibles de  $\phi$ ,  $\psi$  no están permitidas (figura I.9) y se corresponden con motivos comunes de estructura secundaria (figura I.10).



**Figura I.9:** Diagrama general de Ramachandran. La densidad de puntos refleja la probabilidad de cada combinación de ángulos, definiendo las regiones central (línea roja) y de tolerancia (naranja).



**Figura I.10:** Definición de alternativas de estructura secundaria por su combinación de ángulos  $\phi$ ,  $\psi$ .

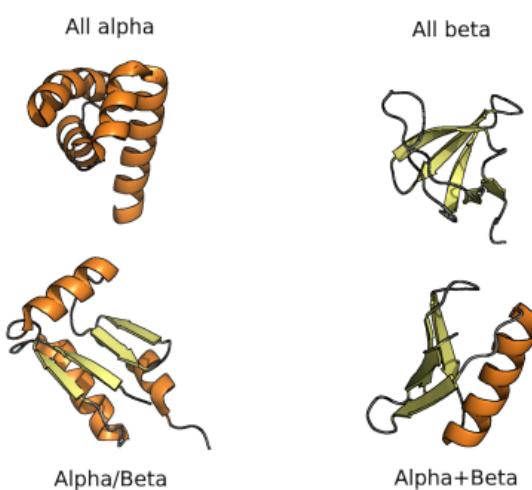
Los residuos funcionalmente relevantes son más propensos que otros a tener ángulos de torsión que se sitúan en las regiones permitidas pero desfavorecidas de un diagrama de Ramachandran. La geometría específica de estos residuos relevantes desde el punto de vista funcional, aunque desfavorable desde el punto de vista energético, puede ser importante para la función de la proteína, ya sea catalítica o de otro tipo. Tales conformaciones deben ser estabilizadas por la proteína mediante enlaces H, empaquetamiento estérico u otros medios, y rara vez se dan en residuos muy expuestos a disolventes.

Suele haber espacios designados para las hélices  $\alpha$  y las láminas  $\beta$ , pero también puede haber outliers que muestren aminoácidos concretos.

### I.2.2. Pliegues (folds), dominios y motivos de proteínas

La estructura terciaria tridimensional global de una proteína se conoce comúnmente como su **pliegue**, definiendo así la forma y orientación global ignorando los loops. Dentro del pliegue proteico global, podemos reconocer distintos dominios y motivos. Los **dominios** son secciones compactas de la proteína que representan regiones estructural y (normalmente) funcionalmente independientes. Eso significa que un dominio mantiene sus características principales, aunque se separe de la proteína global. Por otro lado, los **motivos** son pequeñas subestructuras que no son necesariamente independientes y que constan sólo de unos pocos tramos de estructura secundaria. De hecho, los motivos también pueden denominarse superestructuras secundarias y son frecuentes en la secuencia. En resumen, un dominio corresponde a un fold, y una cadena peptídica puede tener uno o varios dominios.

La diversidad de pliegues, dominios y motivos proteicos, así como su combinación, puede utilizarse para clasificar jerárquicamente las estructuras proteicas, como en muchos otros campos de la biología. La primera clasificación se propuso en los años 70 y consistía en cuatro grupos de pliegues, como se muestra en la siguiente figura. Todas las proteínas  $\alpha$  se basan casi por completo en una estructura  $\alpha$ -hélice, y todas las estructuras  $\beta$  se basan en  $\beta$ -láminas. La estructura  $\alpha/\beta$  se basa en una mezcla de  $\alpha$ -hélices y  $\beta$ -láminas, a menudo organizadas como  $\beta$ -hebras paralelas conectadas por  $\alpha$ -hélices. Por otro lado, las estructuras  $\alpha+\beta$  consisten en motivos discretos de  $\alpha$ -hélice y  $\beta$ -lámina que no están entrelazados (como ocurre en las proteínas  $\alpha/\beta$ ). Por último, las proteínas pequeñas abarcan polipéptidos con estructuras secundarias nulas o escasas.



**Figura I.11:** Las cuatro clases de proteínas estructurales de la clasificación de Chothia y Levitt.

# Capítulo II

## Bases de datos de proteínas

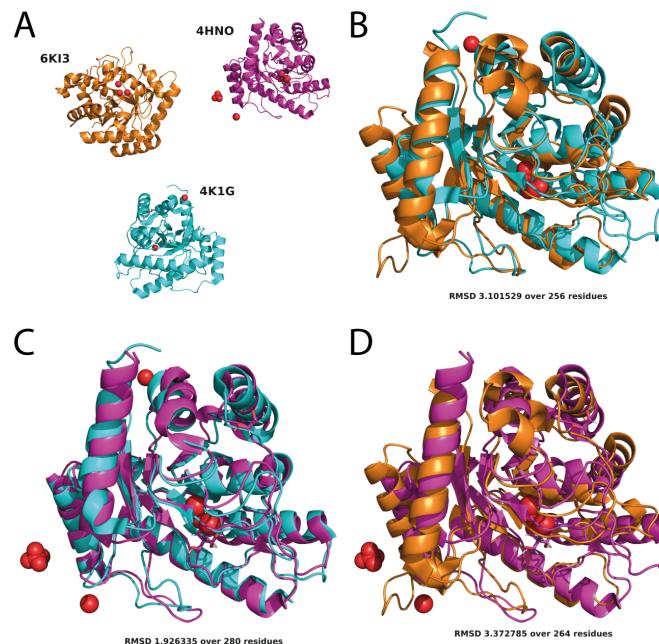
### II.1. Comparación de estructura y alineamiento

Para comprender la diversidad y función de las proteínas, es importante comparar sus secuencias y estructuras. Esto ayuda a encontrar patrones comunes y a comprender su diversidad e historia evolutiva. Midiendo y analizando estas similitudes, los científicos pueden clasificar las proteínas y determinar sus relaciones en términos de función y evolución. Este proceso también es crucial en el modelado de proteínas, ya que ayuda a identificar, evaluar y elegir modelos intermedios.

Es esencial aclarar la distinción entre alineamiento y superposición, ya que estos términos se confunden con frecuencia en la literatura. Un **alineamiento estructural** pretende identificar similitudes y diferencias entre dos estructuras, mientras que la **superposición de estructuras** muestra las estructuras basándose en criterios específicos, normalmente derivados de un alineamiento estructural previo. Por consiguiente, la superposición trata de minimizar la distancia entre estructuras identificando una transformación que consiga la menor desviación cuadrática media (RMSD) o las máximas equivalencias dentro de un límite RMSD.

La RMSD puede calcularse para cualquier par de moléculas. En el contexto de las proteínas, solemos referirnos a la RMSD de los alfa-carbones. Una alineación superior facilitará una mejor superposición. Por lo tanto, aunque la alineación y la superposición son procesos distintos, la RMSD puede servir como indicador de ambos; cuanto menor sea la RMSD, mejor será la alineación/superposición. Es importante señalar que la RMSD es una medida de distancia real, no una puntuación. Eso implica que sólo podemos obtener la RMSD para los residuos alineados, no para toda la secuencia de cualquiera de las dos proteínas. Por lo tanto, una RMSD de 1  $\text{\AA}$  puede indicar una distancia cercana pero, si implica a muy pocos aminoácidos, no sugiere necesariamente una buena similitud. Tanto el valor RMSD como el número de residuos alineados deben tenerse en cuenta para un análisis preciso.

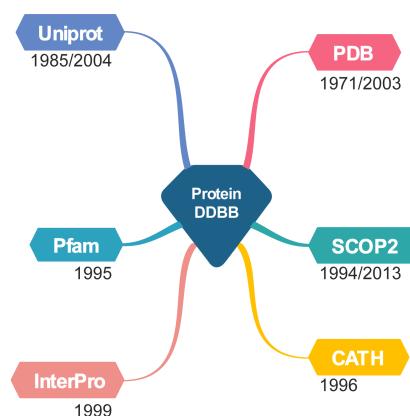
Global Distance Test se utiliza en CASP al ser menos sensible a outliers y permite comparar estructuras de secuencias idénticas. Se normaliza el número de residuos que caigan bajo un límite.



## II.2. Principales bases de datos de proteínas

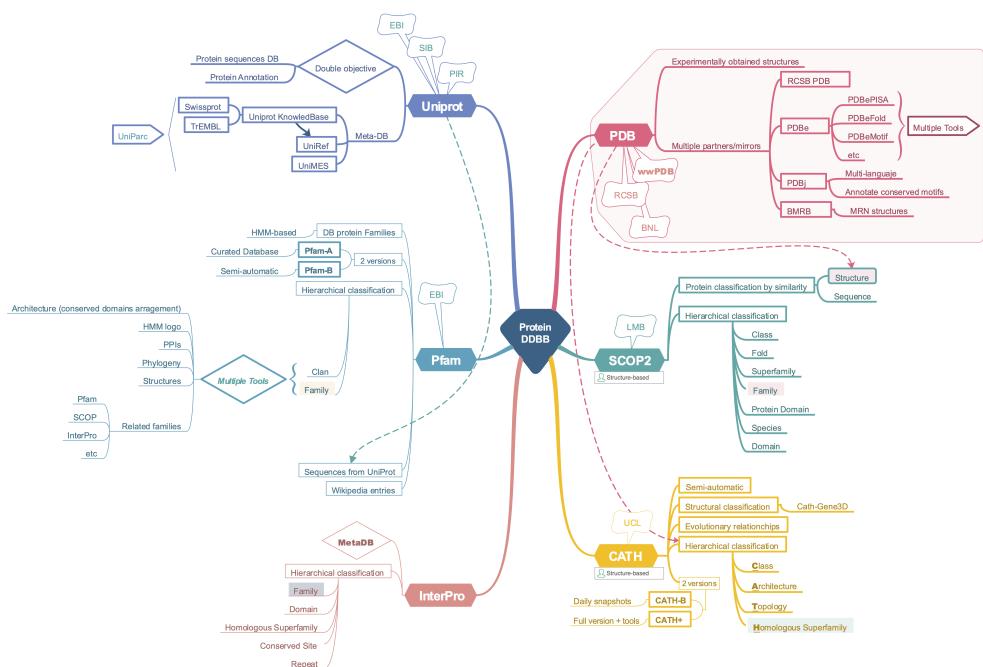
La clasificación de secuencias proteicas nos ayuda a comprender la diversidad de las distintas proteínas mediante el examen de sus secuencias, lo que se conoce como el **espacio de secuencias proteicas** (el concepto matemático de espacio). Por otro lado, la clasificación de las estructuras proteicas consiste en agrupar las proteínas en función de sus relaciones estructurales. Algunas clasificaciones tienen en cuenta la vecindad estructural (continuo estructural), mientras que otras utilizan el concepto de evolución de las proteínas como principal factor de diversificación, lo que da lugar a un **espacio de estructuras proteicas** discreto en lugar de continuo.

Esta sección no pretende ofrecer una revisión exhaustiva de todas las bases de datos de proteínas. En consecuencia, no cubriremos en detalle la base de datos de proteínas del NCBI, que se utiliza ampliamente para diversos fines y probablemente se mencione en otros cursos. La Base de Datos de Proteínas del NCBI sirve principalmente como repositorio principal de secuencias, con un énfasis mínimo en el análisis de la diversidad y clasificación de proteínas. Esta sección destacará las principales diferencias y aplicaciones de Pfam, Uniprot, Prosite, PDB, SCOP y CATH.



En bioinformática, las bases de datos suelen clasificarse en primarias o secundarias. Las **bases de datos primarias** contienen datos obtenidos experimentalmente, como secuencias de nucleótidos, secuencias de proteínas o estructuras macromoleculares. Es importante señalar que, una vez asignado el número de acceso a una base de datos, los datos de las bases de datos primarias permanecen inalterados y forman parte del registro científico. En cambio, las **bases de datos secundarias** incluyen datos derivados del análisis de datos primarios. Estas bases de datos suelen utilizar información procedente de numerosas fuentes, incluidas otras bases de datos y la literatura científica. Suelen ser muy complejas e implican una compleja combinación de algoritmos informáticos y/o análisis e interpretación manuales para generar nuevos conocimientos a partir del registro público de la ciencia.

Aunque la distinción entre bases de datos primarias y secundarias se ha vuelto menos clara en los últimos tiempos debido a la integración de datos procedentes de diversas fuentes, aún pueden distinguirse algunas diferencias. Las principales bases de datos primarias para secuencias de proteínas son NCBI Protein y RCSB-PDB para estructuras proteicas. UniProt también alberga una base de datos primaria de secuencias denominada TrEMBL y, desde 2002, incorpora la base de datos PIR-PSD, que reúne los recursos de Protein Information Resource, EMBL y SIB en una única metabase de datos (véase PIR-PSD). Por otra parte, RCSB-PDB es la principal base de datos estructural primaria, mientras que SCOP2 y CATH son bases de datos secundarias notables.



Todas las bases de datos que describimos aquí permiten el acceso mediante programación y/o API, normalmente con paquetes BioPython y R, lo que aumenta significativamente las posibilidades de programación y análisis de datos por lotes.

## II.2.1. Bases de datos estructurales

### II.2.1.1. RCSB-PDB

La base de datos Protein Data Bank es la principal base de datos estructural primaria de macromoléculas. Contiene principalmente estructuras de proteínas, pero también abarca ácidos nucleicos y complejos nucleoproteicos. PDB cumplió 50 años en 2021 y se puede ver un resumen detallado de su historia en el sitio RCSB-PDB.

Brevemente, el PDB se creó en 1971 en el Laboratorio Nacional de Brookhaven con sólo 7 estructuras. Posteriormente, el **Research Collaboratory for Structural Bioinformatics (RCSB)**, formado por Rutgers, UCSD/SDSC y CARB/NIST, se hizo responsable de la gestión del PDB en 1998 en respuesta a una RFP y un largo proceso de revisión. En 2003, se creó el Worldwide Protein Data Bank (wwPDB) para mantener un único archivo PDB de datos estructurales macromoleculares a disposición libre y pública de la comunidad mundial. Está formado por organizaciones que actúan como centros de depósito, procesamiento y distribución de datos PDB.

Las estructuras del PDB se obtienen en gran medida mediante cristalográfia de rayos X, pero acepta derivaciones de datos de EM y RMN desde 1989 y 1991, respectivamente. De hecho, el BMRB (Biological Magnetic Resonance Bank) se ha asociado con el PDB desde 2006 y el EMBD (Electron Microscopy Data Bank) desde 2021. Además, a partir de septiembre de 2022, el PDB también contiene modelos computados de la base de datos AlphaFold (de la que hablaremos más adelante en este curso) y RoseTTAFold-ModelArchive. **Así pues, la base de datos PDB es el eje principal que centraliza las estructuras biológicas en la actualidad.**

La base de datos PDB tiene cuatro réplicas y sitios web (RCSB, Europa, BMRB y Japón) con información que se solapa principalmente, aunque tienen cierta especialización. El sitio PDB del RCSB tiene también una sección educativa (PDB-101) con información y recursos muy útiles para la enseñanza y el aprendizaje de la biología estructural y el trabajo con estructuras PDB.

Las entradas del PDB contienen toda la información sobre la estructura, desde la secuencia de la proteína y su origen hasta los detalles del experimento, así como la evaluación de la estructura y la visualización. Se puede descargar toda esta información y las coordenadas de la estructura en diversos formatos de archivo.

### II.2.1.2. SCOP

La base de datos Structural Classification of Proteins (SCOP, <http://scop.mrc-lmb.cam.ac.uk>) es una **clasificación de dominios proteicos** organizada según sus relaciones evolutivas y estructurales en categorías jerárquicas. La unidad principal es la **familia**, que agrupa proteínas relacionadas con pruebas claras de su origen evolutivo, mientras que la **superfamilia** reúne dominios proteicos relacionados de forma más distante. Además, las superfamilias se agrupan en **pliegues** distintos en función de las características estructurales globales que comparten la mayoría de sus miembros. Se proporcionan definiciones de dominio para los dos niveles principales de la clasificación SCOP, familia y superfamilia, y los límites de dominio para cada uno de ellos pueden coincidir o diferir.

Para cada grupo, se selecciona un representante basándose en su secuencia (UniProtKB) y estructura (PDB) y se utiliza para la clasificación SCOP. Así, los límites de dominio SCOP se asignan tanto a la entrada PDB como a la UniProtKB.

### II.2.1.3. CATH

CATH ([www.cathdb.info](http://www.cathdb.info)) es un recurso gratuito y de acceso público que identifica dominios proteicos dentro de proteínas del Banco de Datos de Proteínas y los clasifica en grupos relacionados evolutivamente según la información sobre secuencia, estructura y función. Parte de la base de que las proteínas relacionadas que se pliegan de forma similar suelen exhibir funciones similares (esto sólo podría demostrarse si encontramos intermediarios). CATH utiliza un esquema de clasificación jerárquica en el que las unidades comparadas y clasificadas son dominios estructurales. Los dominios, definidos aquí como dominios estructurales globulares capaces de plegarse de forma semiindependiente, se extraen de estructuras de proteínas determinadas experimentalmente y disponibles en la base de datos PDB. Los dominios se clasifican en los siguientes niveles jerárquicos que componen el nombre CATH: Clase (C), Arquitectura (A), Topología (T) y Superfamilias homólogas (H).

CATH utiliza una combinación de varios algoritmos basados en estructuras (SSAP, CATHEDRAL) y en secuencias (alineaciones de secuencias basadas en Needleman-Wunsch, Jackhmmer, Profile Comparer y HHsearch) para evaluar la similitud de los dominios entre sí e identificar proteínas homólogas.

CATH tiene un recurso hermano, Gene3D, que añade secuencias adicionales de dominios de proteínas sin estructura conocida, lo que eleva el número total actual de dominios en CATH-Gene3D a 95 millones.

La base de datos CATH se actualiza con bastante regularidad mediante instantáneas diarias (CATH-B), pero cada 12 meses se publica una versión completa con más herramientas, denominada CATH+. CATH-plus contiene familias funcionales (CATH-FunFams), clusters estructurales y otras herramientas.

## II.2.2. Bases de datos de secuencias

### II.2.2.1. Uniprot

Las bases de datos Uniprot están gestionadas por el consorcio UniProt, creado en 2002 por EMBL-EBI, SIB y PIR. En la actualidad, UniProt puede considerarse una metadatabase, ya que sus entradas contienen información procedente de diversas fuentes. Se creó con dos objetivos principales: establecer una base de datos de secuencias de proteínas completa y no redundante y enriquecer esa base de datos con anotaciones detalladas. Estas anotaciones incluyen familias de proteínas y genes, datos de función y estructura-función, interacciones con otras proteínas o cofactores, localización, patrones de expresión, variantes, etc. Así, pretende cumplir los objetivos tanto de las bases de datos primarias como de las secundarias.

El eje central de las bases de datos UniProt es la Uniprot Knowledgebase. Se trata de una colección de información funcional sobre proteínas, con anotaciones

precisas, coherentes y ricas. UniProtKB consta de dos bases de datos internas: una sección contiene registros anotados manualmente con información extraída de la bibliografía, sugerencias de la comunidad y análisis computacionales revisados por los conservadores. La otra sección incluye registros analizados computacionalmente. Estas secciones se denominan «UniProtKB/Swiss-Prot» (revisada, anotada manualmente) y «UniProtKB/TrEMBL» (no revisada, anotada automáticamente), respectivamente. En los últimos años, UniProtKB ha incorporado datos estructurales de la base de datos AlphaFold, además de referencias cruzadas a información estructural.

UniProt contiene secuencias con distintos niveles de detalle de anotación en dos bases de datos complementarias: Uniparc y Uniref. En resumen, UniParc (UniProt Archive) es una base de datos exhaustiva y no redundante que incluye la mayoría de las secuencias de proteínas disponibles públicamente en todo el mundo. UniParc evita la redundancia almacenando cada secuencia única una sola vez y asignándole un identificador único estable (UPI), que permite identificar la misma proteína a partir de diferentes bases de datos fuente. Un UPI nunca se elimina, cambia o reasigna. Por otro lado, UniRef (UniProt Reference Clusters) proporciona conjuntos agrupados de secuencias de UniProtKB (y registros seleccionados de UniParc) para garantizar una cobertura completa del espacio de secuencias a varias resoluciones, ocultando al mismo tiempo las secuencias redundantes (pero no sus descripciones). La base de datos UniRef100 combina secuencias idénticas en una única entrada UniRef, mostrando la secuencia de una proteína representativa, los números de acceso de todas las entradas fusionadas y enlaces a las bases de datos correspondientes. UniRef90 se construye agrupando secuencias UniRef100 utilizando el algoritmo MMseqs2, de modo que cada clúster consiste en secuencias con al menos un 90 % de identidad de secuencia y un 80 % de solapamiento con la secuencia más larga (la secuencia semilla ) del clúster. Del mismo modo, UniRef50 se construye agrupando secuencias semilla UniRef90 que tienen al menos un 50 % de identidad de secuencia y un 80 % de solapamiento con la secuencia más larga del clúster. UniParc y UniRef sólo contienen secuencias de proteínas; el resto de la información sobre las proteínas debe recuperarse de las bases de datos de origen utilizando referencias cruzadas de bases de datos.

### II.2.2.2. InterPro

InterPro pretende ser una base de datos funcional secundaria, clasificando las proteínas en familias, dominios y sitios importantes. Para clasificar las proteínas de este modo, InterPro utiliza modelos predictivos, conocidos como firmas, proporcionados por varias bases de datos diferentes (hasta 13) que conforman el consorcio InterPro. InterPro combina esas diferentes firmas que representan familias, dominios o sitios equivalentes, y proporciona información adicional como descripciones, referencias bibliográficas y términos de la Ontología Genética (GO), para producir un recurso completo para la clasificación de proteínas.

La base de datos InterPro se actualiza cada 2 meses y es muy útil para la anotación de ORFans o proteínas divergentes. En los últimos años, ha integrado más recursos, incluyendo Pfam, así como datos estructurales y predicciones, dando lugar a un recurso muy práctico para múltiples propósitos en la ciencia de las proteínas.

InterPro se creó como una BBDD de secuencias, pero actualmente se encuentra en un punto intermedio. Ahora se podría decir que es más bien una «metabase de datos» que contiene información sobre secuencias y estructuras.

### II.2.2.3. Pfam

Pfam es una base de datos de proteínas cuyo objetivo es clasificar secuencias por sus relaciones evolutivas. Se fundó en 1995 y ha sido muy útil para la anotación funcional de datos genómicos. El sitio web de Pfam (<http://pfam.xfam.org/>) se cerró a finales de 2022. Sin embargo, la base de datos Pfam no se interrumpió, sino que se integró en el sitio InterPro. Pfam utiliza perfiles HMM para clasificar las proteínas en familias, que se agrupan en clanes.

La versión actual (37.1) contiene 23.794 entradas y 751 clanes. Pfam se diseñó como una base de datos que debe actualizarse con frecuencia en la era genómica de avance rápido. Para ello, utiliza dos tipos de alineación. Cada familia Pfam tiene un alineamiento semilla que contiene un conjunto representativo de secuencias para la entrada. A partir del alineamiento semilla se construye automáticamente un modelo de Markov oculto (HMM) de perfil y se busca en una base de datos de secuencias denominada pfamseq utilizando el software HMMER3 (<http://hmmer.org/>). Todas las regiones de secuencias que satisfacen un umbral curado específico de la familia, también conocido como umbral de reunión, se alinean con el HMM de perfil para crear el alineamiento completo.

Además de las entradas Pfam basadas en HMM (Pfam-A), los perfiles Pfam se utilizan para proporcionar un conjunto de alineaciones de secuencias múltiples no anotadas, generadas computacionalmente, denominadas Pfam-B. Sin embargo, en las últimas versiones de Pfam, los alineamientos Pfam-B sólo se publican actualmente en el sitio FTP de Pfam.

Pfam también se ha utilizado en la creación de otros recursos como Rfam (familias de ARN) y Dfam (elementos transponibles de ADN).

## II.3. Estrategias actuales y futuras en las bases de datos de proteínas

Existe una tendencia significativa hacia el cruce y la integración de datos diversos dentro de las **metadatabases**. Un caso ejemplar es el Human Protein Atlas, que proporciona información sobre proteínas clasificadas por tipo celular o tejido, junto con detalles sobre variantes de splicing, mutantes, etc. Además, es importante reconocer las nuevas bases de datos estructurales, como la base de datos AlphaFold de Deepmind y el Atlas Metagenómico ESM, que albergan millones de estructuras de proteínas predichas mediante métodos de aprendizaje profundo. También existen bases de datos especializadas, como BFDV, que contienen estructuras de proteínas víricas obtenidas a través de AlphaFold (pero que no están en la base de datos de AlphaFold) y en las que se pueden realizar búsquedas mediante Foldseek, un método diseñado para identificar similitudes estructurales.

Dado el reciente impulso en la capacidad de obtener con precisión modelos de proteínas, algunos autores sugirieron (o desearon) que las futuras bases de datos contuvieran no solo variantes de secuencias de proteínas y complejos proteicos, sino también conformaciones diversas para cada estructura, lo que ayudaría a conocer mejor su función y papel biológico.

# Capítulo III

## Estructuras de proteínas

### III.1. Obtener y trabajar con estructuras de proteínas

El pintor surrealista belga René Magritte creó una colección de cuadros surrealistas titulada *La trahison des images* (1928-1929). El más famoso de estos cuadros muestra una pipa humeante con la siguiente leyenda debajo: «Ceci n'est pas une pipe» (Esto no es una pipa). Efectivamente. En realidad es un cuadro de una pipa. Esto también es extrapolable a la bioinformática: Una imagen de una proteína, o un archivo informático con las coordenadas de una estructura proteica, no constituye la proteína real. Más bien representa **una** posible conformación de esa proteína.

Incluso las estructuras determinadas experimentalmente tienen dos limitaciones importantes que deben tenerse en cuenta: (1) representan una estructura fija (excepto las basadas en RMN), mientras que las proteínas *in vivo* son flexibles y dinámicas, y (2) están sujetas a errores experimentales y a menudo contienen regiones de baja confianza. Además, incluso las estructuras macromoleculares determinadas experimentalmente son, hasta cierto punto, modelos con proporciones variables entre los datos experimentales y las predicciones computacionales utilizadas para hacer coincidir los datos experimentales (como difracción de rayos X, mapas de densidad crio-EM, RMN, SAXS, FRET...) con estructuras o modelos conocidos previamente. Es importante señalar que, aunque las estructuras de proteínas pueden ser muy valiosas, debemos ser conscientes de sus limitaciones y aplicaciones.

### III.2. Determinación experimental de las estructuras de proteínas

El análisis estructural de las proteínas es crucial para comprender en detalle los mecanismos moleculares que subyacen a sus funciones. Una representación tridimensional facilita la orientación de diversos dominios, motivos o residuos de interés, lo que resulta esencial para comprender variantes poblacionales o patógenas, el diseño de fármacos y la ingeniería de proteínas. Además, las estructuras proteínicas pueden ayudar a predecir la función y las relaciones evolutivas, ya que la conservación

estructural es mayor que la conservación secuencial; el espacio de la estructura proteínica es menor que el espacio secuencial. Sin embargo, la obtención de datos estructurales precisos y detallados puede ser un reto técnico y requerir mucho tiempo. Como ya se ha dicho, el modelado de estructuras proteicas suele ser un valioso complemento o una alternativa. Las estructuras obtenidas experimentalmente suelen obtenerse mediante cristalografía de rayos X, resonancia magnética nuclear (RMN) o criomicroscopía electrónica (crioEM).

### III.2.1. Cristalografía de rayos X o difracción de rayos X de un solo cristal

La cristalografía de rayos X, también conocida como difracción de rayos X de monocrystal, es una técnica utilizada para determinar la estructura atómica de las moléculas dentro de formas cristalinas. Este proceso consiste en crear un cristal de la molécula de interés, que se coloca en un goniómetro y se expone a un haz concentrado de rayos X (Figura III.1). El patrón de difracción resultante producido por los rayos X que atraviesan el cristal permite determinar las posiciones atómicas, los enlaces químicos, el desorden cristalográfico y otros detalles estructurales. La interpretación de la relación entre el patrón de difracción y la densidad de electrones requiere complejos cálculos matemáticos, en particular mediante transformadas de Fourier, para generar un *modelo* tridimensional de la estructura.

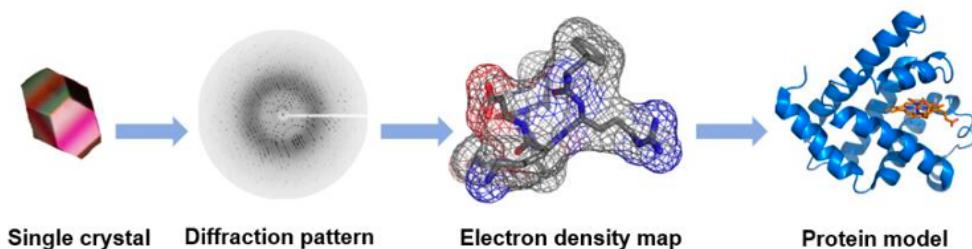


Figura III.1: Flujo de trabajo esquemático de la cristalografía de rayos X.

Cuando recogemos datos de difracción de rayos X de un cristal, medimos las intensidades de las ondas difractadas dispersadas en todas las direcciones. Estas medidas nos dan las amplitudes, pero no la información de fase necesaria para reconstruir una imagen (mapa de densidad) de la molécula, lo que se conoce como el «problema de fase». Este problema se agrava cuando faltan datos o éstos son deficientes. En la cristalografía de proteínas, las fases se obtienen a menudo utilizando las coordenadas atómicas de una proteína similar (reemplazamiento molecular, MR) o identificando las posiciones de los átomos pesados. Los átomos pesados dispersan los rayos X con más intensidad que los ligeros, lo que nos ayuda a determinar sus posiciones dentro del cristal. Comparando los patrones de difracción del cristal original y de uno con átomos pesados añadidos, podemos deducir información de fase mediante la sustitución isomorfa. Los átomos pesados actúan como puntos de referencia para recuperar la información de fase perdida, crucial para reconstruir la estructura tridimensional de la molécula. El reemplazo molecular encuentra modelos que se ajustan a las intensidades experimentales a partir de estructuras conocidas, para lo que suele ser necesario cubrir al menos el 50 % de la estructura total con una

d.s.r.m. C $\alpha$  baja. Alrededor del 70 % o más de las estructuras PDB se han resuelto mediante este método (MR), y el número aumenta a medida que se dispone de más estructuras homólogas. Los avances en la predicción de estructuras proteicas de novo han dado lugar a protocolos como MR-Rosetta, QUARK, AWSEM-Suite, I-TASSER-MR y AlphaFold-guided MR, que generan estructuras señuelo de tipo nativo útiles para resolver el problema de fase.

La difracción de rayos X es una potente técnica que permite obtener estructuras de alta resolución a nivel atómico de proteínas tanto solubles como de membrana, ya sean apoenzimas o holoenzimas unidas a un sustrato, cofactor o fármaco. Sin embargo, la muestra de proteína debe ser cristalizable (es decir, homogénea), lo que requiere una cantidad sustancial de proteína muy pura. Otra limitación de las estructuras de rayos X es que sólo proporcionan una (o muy pocas) formas estáticas de la proteína, y la localización de los átomos de hidrógeno no puede determinarse mediante métodos de difracción convencionales. Debido a su único electrón, los átomos de hidrógeno son difíciles de detectar con precisión con rayos X, que se dispersan en la densidad de electrones. Aunque los átomos de hidrógeno pueden predecirse, esta limitación sigue complicando algunos análisis químicos. Algunas proteínas conservan toda su funcionalidad, lo que permite realizar experimentos de cristalización con ciertas enzimas, pero también hay numerosos ejemplos en los que la cristalización puede conducir a una representación sesgada de la proteína y dar lugar a artefactos estructurales.

### III.2.2. Resonancia magnética nuclear

Todos los núcleos atómicos son partículas cargadas que giran rápidamente y producen frecuencias de resonancia únicas para cada átomo. Cuando se aplica un campo magnético, puede detectarse una señal electromagnética con una frecuencia característica del campo magnético en el núcleo. Este principio constituye la base de la resonancia magnética nuclear (RMN, Figura III.2).

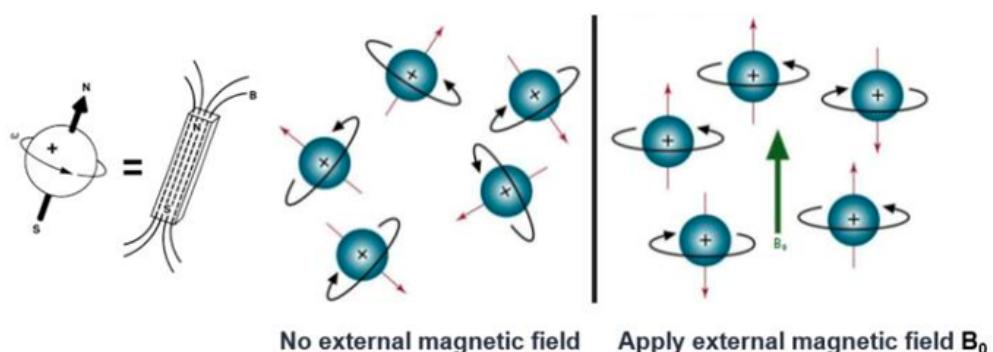
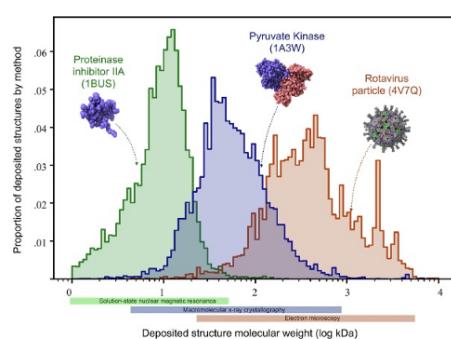


Figura III.2: Bases de la resonancia magnética nuclear.

Es importante señalar que el movimiento del núcleo no es aislado; interactúa tanto intra como intermolecularmente con los átomos circundantes. Por consiguiente, la espectroscopía de resonancia magnética nuclear puede proporcionar información estructural sobre moléculas específicas. Por ejemplo, en las proteínas, las estructuras secundarias como las  $\alpha$ -hélices, las  $\beta$ -hojas y los giros indican diversas disposiciones

de los átomos de la cadena principal en el espacio tridimensional. Las distancias entre los núcleos atómicos en estas estructuras secundarias, sus interacciones y las propiedades dinámicas de los segmentos polipeptídicos revelan directamente la estructura tridimensional de las proteínas. Estas características nucleares contribuyen al comportamiento espectroscópico de la muestra, dando lugar a señales de RMN distintivas. La interpretación computacional de estas señales facilita la determinación de la estructura tridimensional de la proteína.

La principal ventaja del método de RMN es que permite medir directamente la estructura tridimensional de las macromoléculas en su estado natural en solución. La RMN proporciona información sobre la dinámica y las interacciones intermoleculares de estas moléculas. La resolución de la estructura tridimensional puede alcanzar el rango subnanométrico. Sin embargo, el espectro de RMN de biomoléculas grandes es complejo y difícil de interpretar, lo que limita su aplicación al análisis de biomoléculas grandes, normalmente por debajo de 20-30 kDa (Figura III.3). Además, esta técnica requiere cantidades relativamente grandes de muestras puras (varios miligramos) para lograr una relación señal-ruido razonable.

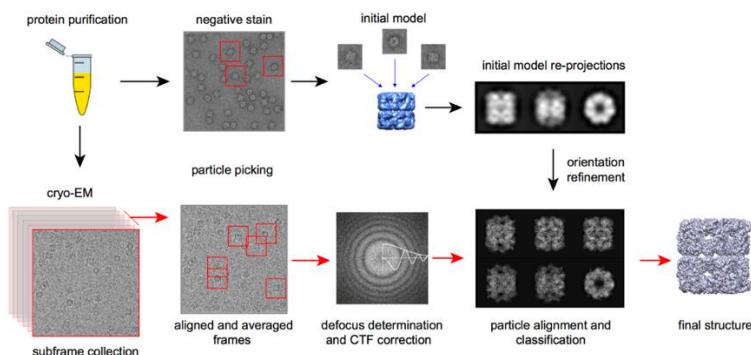


**Figura III.3:** Cobertura del peso molecular mediante la técnica estructural.

### III.2.3. Criomicroscopía electrónica

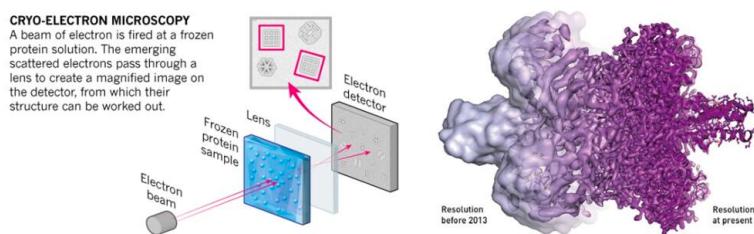
El principio fundamental de la crioEM es la dispersión de electrones, similar a otros métodos de microscopía electrónica. Las muestras se preparan mediante crioconservación antes del análisis. A continuación, se utiliza una fuente de electrones como fuente de luz para medir la muestra. Después de que el haz de electrones atravesie la muestra, un sistema de lentes convierte la señal dispersa en una imagen ampliada que se graba en el detector. Un paso posterior crucial es el procesamiento de la señal, que transforma miles de imágenes de las partículas en diversas orientaciones en una estructura tridimensional de la muestra.

Tradicionalmente, el uso de métodos de microscopía electrónica para la biología estructural se limitaba a grandes complejos macromoleculares, como las cápsides víricas (Figura III.3). Recientemente, también se ha aplicado a partículas más pequeñas. El número de estructuras de proteínas determinadas mediante criomicroscopía electrónica ha aumentado considerablemente en los últimos 5-10 años (consultable en PDB). Este aumento se debe a varias mejoras técnicas de la técnica (Figura III.5), como la preparación y conservación de muestras, el análisis y el procesamiento, que permiten obtener imágenes a nivel atómico. Estos avances fueron reconocidos con la concesión



**Figura III.4:** El proceso de la técnica de análisis de partículas individuales Cryo-EM.

del Premio Nobel de Química 2017 a Jacques Dubochet, Joachim Frank y Richard Henderson .



**Figura III.5:** La revolución de la criomicroscopía electrónica.

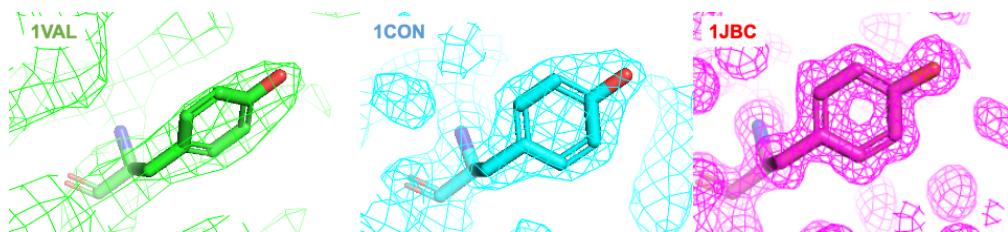
La crioEM se utiliza habitualmente hoy en día, sobre todo para grandes complejos moleculares o partículas víricas. Permite generar estructuras rápidamente, requiere una cantidad mínima de proteínas y puede producir datos fiables incluso con impurezas presentes. Sin embargo, los microscopios de nueva generación sólo suelen ser asequibles para las grandes instituciones, y las partículas pequeñas suelen tener un alto nivel de ruido. Además, procesar un gran número de imágenes puede suponer un reto cuando se pretende obtener estructuras de alta calidad.

### III.3. Garantía de calidad estructural

Toda estructura, independientemente de su origen o método de determinación, es susceptible de error. Las estructuras determinadas experimentalmente son, en realidad, modelos que se han construido para alinearse con los datos experimentales. La calidad de los datos iniciales y la precisión de los procedimientos experimentales influyen significativamente en la fiabilidad de los resultados estructurales. Al igual que ocurre en otras disciplinas científicas, los experimentos independientes pueden dar lugar a modelos relacionados de la misma molécula, aunque suele haber variaciones; no obstante, ambos modelos pueden seguir considerándose representaciones exactas.

### III.3.1. Parámetros globales en estructuras basadas en experimentos

Existen distintos parámetros que nos ayudan a comprender la calidad y fiabilidad de una estructura. En primer lugar, la **resolución** es un buen indicador del nivel de detalle de la estructura, ya que puede afectar en gran medida a la modelización de los datos experimentales.



**Figura III.6:** Efecto de la resolución en la calidad de la densidad electrónica. El residuo Tyr100 de la concanavalina A tal y como se encuentra en las estructuras PDB indicadas a 3 Å, 2 Å y 1,2 Å.

Otro parámetro importante es el **factor R**, que es la diferencia entre los factores de estructura calculados a partir del modelo y los obtenidos a partir de los datos experimentales. Es decir, el factor R es la desviación entre el patrón de difracción calculado del modelo y el patrón de difracción experimental original. Normalmente, las buenas estructuras con una resolución de 1-3 Å, tienen un factor R de 0,2 (es decir, 20 % de desviación). Sin embargo, debe tenerse en cuenta que este factor suele reducirse tras el refinamiento iterativo, lo que resta importancia a su uso como indicador de fiabilidad. Un factor más fiable es el **factor  $R_{free}$** . Éste es menos susceptible de manipulación durante el refinamiento, ya que se basa sólo en una pequeña parte de los datos experimentales (5-10 %) que no se utiliza durante la fase de refinamiento.

Una forma más intuitiva, aunque sólo cualitativa, de entender la precisión de las coordenadas de un átomo determinado es el factor B. El valor de temperatura o factor B se correlaciona con los errores de posición, aunque su definición matemática es más compleja. Los valores normales de un factor B se sitúan entre 14 y 30, mientras que los valores superiores a 30 suelen indicar que el átomo se encuentra en una región flexible o desordenada, y los átomos con un factor B superior a 40 suelen descartarse por ser demasiado poco fiables.

La desviación cuadrática media (RMSD) es un estimador tradicional de la calidad de las estructuras resueltas por RMN. Las regiones con valores altos de RMSD (por encima de 6) son las que están menos definidas por los datos. Sin embargo, hay que tener en cuenta que este parámetro también puede ser engañoso, ya que depende en gran medida del procedimiento utilizado para generar y seleccionar los datos que se envían al PDB. Un experimentalista podría reducir la RMSD seleccionando las «mejores» pocas estructuras para su depósito a partir de un borrador mucho mayor. Además, la RMSD tiene muchas otras aplicaciones, como la comparación de diferentes estructuras o modelos de la misma secuencia o de secuencias relacionadas.

En los últimos años, con el aumento de la cantidad y la calidad de las estructuras EM, también se han propuesto nuevos parámetros. Uno de ellos, el **factor Q**, se

introdujo recientemente para la [validación de estructuras 3DEM/PDB](#). Brevemente, la puntuación del factor Q calcula la resolubilidad de los átomos midiendo la similitud de los valores del mapa alrededor de cada átomo en relación con una función tipo Gauss para un átomo bien resuelto. Una puntuación Q de 1 significa que la similitud es perfecta, mientras que un valor cercano a 0 indica una similitud baja. Si el átomo no está bien situado en el mapa, puede darse un valor Q negativo. Por lo tanto, los valores del factor Q en los informes oscilan entre -1 y +1.

### III.3.2. Parámetros estereoquímicos

Dado que todos los modelos estructurales contienen cierto grado de error y que algunos de los parámetros globales de modelización pueden ser controvertidos, podemos analizar la geometría, la estereoquímica y otras propiedades estructurales del modelo para evaluar los modelos estructurales. Estos parámetros comparan una estructura dada con lo que ya se sabe sobre ese tipo de molécula a partir de nuestro conocimiento de las estructuras de alta resolución. Esto significa que las estructuras del espacio estructural actual definen lo que es «normal» en la estructura de una proteína. La ventaja de estos análisis y parámetros derivados es que no tienen en cuenta el proceso que conduce al modelo, sino sólo el producto final y su fiabilidad. La principal desventaja es que el espacio estructural actual se centra en proteínas con función conocida y de interés biomédico o biotecnológico.

Uno de los métodos más comunes y potentes para evaluar la estereoquímica de una proteína es el diagrama de Ramachandran, que se definió en 1963 y sigue utilizándose.

Otro análisis muy utilizado (disponible para todas las estructuras PDB) es el de los **ángulos de torsión de la cadena lateral**, medidos normalmente como **valores atípicos de la cadena lateral**. Las cadenas laterales de los aminoácidos también tienen algunas conformaciones preferidas. Al igual que el diagrama de Ramachandran, el diagrama de los ángulos de torsión  $\chi_1$ - $\chi_2$  puede indicar problemas con un modelo de proteína si los valores de los ángulos están fuera de los valores de alta densidad.

Los malos contactos o choques indican un modelo deficiente. Es obvio que dos átomos no pueden estar en el mismo lugar (o muy cerca). Podemos definir esto como una situación en la que dos átomos no unidos tienen una distancia entre centros menor que la suma de sus radios de van der Walls.

## III.4. Visualización de estructuras de proteínas

### III.4.1. Formatos de ficheros de estructuras proteicas

Los datos estructurales experimentales de diferentes métodos se almacenan en diferentes formatos de archivo. Por ejemplo, los datos cristalográficos en bruto suelen almacenarse como archivos \*.ccp4, pero los mapas de densidad de Cryo-EM o rayos X pueden almacenarse en archivos \*.mrc o \*.mtz. Otros formatos de archivo complejos, como el Extensible Markup Language \*.xml, proporcionan un marco para estructurar información compleja y documentos como estructuras de proteínas.

Junto con la creación del Banco de Datos de Proteínas, se desarrolló un formato sencillo y estandarizado. El formato Brookhaven o PDB consiste en registros de líneas en un formato fijo que describen coordenadas atómicas, características químicas y bioquímicas, detalles experimentales de la determinación de la estructura y algunas características estructurales como asignaciones de estructuras secundarias, enlaces de hidrógeno o sitios activos. La versión actual se denomina PDBx/mmCIF) también incorpora el formato de archivo de información cristalográfica ampliado (mmCIF), que permite la representación de estructuras de gran tamaño, química compleja y métodos experimentales nuevos e híbridos. Así, los archivos \*.pdb y \*.cif pueden considerarse idénticos.

| Atom Number         | Type Residue | Chain | Res. Number | X      | Y      | Z      | Occupancy | B-factor | Atom Type |
|---------------------|--------------|-------|-------------|--------|--------|--------|-----------|----------|-----------|
| ATOM 1 N GLY A -1   |              |       |             | 21.168 | 11.913 | -0.709 | 1.00      | 45.70    | N         |
| ATOM 2 CA GLY A -1  |              |       |             | 20.234 | 11.449 | -1.719 | 1.00      | 44.85    | C         |
| ATOM 3 C GLY A -1   |              |       |             | 19.218 | 12.497 | -2.132 | 1.00      | 43.19    | C         |
| ATOM 4 O GLY A -1   |              |       |             | 19.405 | 13.687 | -1.883 | 1.00      | 44.65    | O         |
| ATOM 5 N GLY A 0    |              |       |             | 18.132 | 12.049 | -2.765 | 1.00      | 39.19    | N         |
| ATOM 6 CA GLY A 0   |              |       |             | 17.108 | 12.958 | -3.227 | 1.00      | 38.34    | C         |
| ATOM 7 C GLY A 0    |              |       |             | 16.079 | 13.289 | -2.161 | 1.00      | 38.87    | C         |
| ATOM 8 O GLY A 0    |              |       |             | 15.905 | 12.565 | -1.183 | 1.00      | 37.51    | O         |
| ATOM 9 N MET A 1    |              |       |             | 15.391 | 14.408 | -2.365 | 1.00      | 34.06    | N         |
| ATOM 10 CA MET A 1  |              |       |             | 14.351 | 14.874 | -1.451 | 1.00      | 32.26    | C         |
| ATOM 11 C MET A 1   |              |       |             | 13.002 | 14.566 | -2.090 | 1.00      | 31.06    | C         |
| ATOM 12 O MET A 1   |              |       |             | 12.513 | 15.324 | -2.932 | 1.00      | 32.52    | O         |
| ATOM 13 CB MET A 1  |              |       |             | 14.513 | 16.362 | -1.161 | 1.00      | 33.01    | C         |
| ATOM 14 CG MET A 1  |              |       |             | 15.911 | 16.749 | -0.701 | 1.00      | 34.74    | C         |
| ATOM 15 SD MET A 1  |              |       |             | 15.957 | 18.359 | 0.105  | 1.00      | 42.63    | S         |
| ATOM 16 CE MET A 1  |              |       |             | 17.455 | 19.056 | -0.582 | 1.00      | 40.77    | C         |
| ATOM 17 N PHE A 2   |              |       |             | 12.400 | 13.452 | -1.688 | 1.00      | 29.52    | N         |
| ATOM 18 CA PHE A 2  |              |       |             | 11.169 | 12.965 | -2.293 | 1.00      | 30.07    | C         |
| ATOM 19 C PHE A 2   |              |       |             | 9.963  | 13.332 | -1.439 | 1.00      | 30.31    | C         |
| ATOM 20 O PHE A 2   |              |       |             | 10.016 | 13.272 | -0.207 | 1.00      | 30.48    | O         |
| ATOM 21 CB PHE A 2  |              |       |             | 11.225 | 11.450 | -2.495 | 1.00      | 32.37    | C         |
| ATOM 22 CG PHE A 2  |              |       |             | 12.168 | 11.024 | -3.580 | 1.00      | 32.36    | C         |
| ATOM 23 CD1 PHE A 2 |              |       |             | 11.753 | 10.995 | -4.900 | 1.00      | 33.00    | C         |
| ATOM 24 CD2 PHE A 2 |              |       |             | 13.471 | 10.662 | -3.282 | 1.00      | 32.69    | C         |
| ATOM 25 CE1 PHE A 2 |              |       |             | 12.617 | 10.609 | -5.905 | 1.00      | 33.82    | C         |
| ATOM 26 CE2 PHE A 2 |              |       |             | 14.341 | 10.275 | -4.283 | 1.00      | 35.07    | C         |
| ATOM 27 CZ PHE A 2  |              |       |             | 13.914 | 10.247 | -5.596 | 1.00      | 38.10    | C         |

Figura III.7: Coordenadas en un fichero PDB.

### III.4.2. Ocupancia y factor B

Excepto por la repetición del tipo de átomo en la columna de la derecha, las últimas columnas del archivo PDB son la **ocupancia** y el **factor de temperatura** o el **factor B**.

Los cristales macromoleculares están formados por muchas moléculas individuales empaquetadas en una disposición simétrica. En algunos cristales hay ligeras diferencias entre las moléculas individuales. Por ejemplo, una cadena lateral en la superficie puede oscilar entre varias conformaciones, o un sustrato puede unirse en dos orientaciones en un sitio activo, o un ion metálico puede detectarse unido sólo a algunas de las moléculas. Cuando los investigadores construyen el modelo atómico de estas porciones, pueden utilizar la ocupación para estimar la cantidad de cada conformación observada en el cristal. Por tanto, por definición, la suma de los **valores de ocupación** de cada átomo debe ser 1. Normalmente, vemos un único registro para un átomo, con un valor de ocupación de 1, lo que indica que el átomo se encuentra en todas las moléculas en el mismo lugar del cristal. Sin embargo, si un ion metálico se une sólo a la mitad de las moléculas del cristal, el investigador ve una imagen débil del ion en el mapa de densidad electrónica y puede asignar una ocupación de 0,5 para este átomo en el archivo de estructura PDB. Para cada átomo, se incluyen dos (o más) registros de átomos con ocupaciones tales como 0,5 y 0,5, o 0,4 y 0,6, u otras fracciones de ocupaciones que suman un total de 1.

Por otro lado, el **valor de temperatura o factor B** es una medida de nuestra confianza en la localización de átomos individuales, como se ha descrito anteriormente. Si encuentra un átomo con un factor de temperatura alto en la superficie de una proteína, tenga en cuenta que es probable que este átomo se mueva mucho y que las coordenadas dadas en el archivo PDB son sólo una posible instantánea de su ubicación. Así, un conjunto de datos de átomos con una ocupación  $< 1$  puede tener un factor B bajo si esa posición es segura. Esta columna también es utilizada por los modelos derivados computacionalmente para indicar un valor de confianza que puede ser analizado para diversos fines, incluyendo la coloración de la estructura.

La metionina iniciadora siempre recibe el número de residuo de 1. Sin embargo, como a veces los cortes pueden no ser limpios, puede haber alguna cadena residual anterior a la metionina; estos reciben la numeración 0 y -1.

### III.4.3. Aplicaciones de visualización de macromoléculas biológicas

#### III.4.3.1. PyMOL

PyMOL es un sistema de visualización molecular muy potente escrito originalmente por Warren DeLano. Fue lanzado en 2000 y pronto se hizo muy popular. Actualmente se comercializa bajo licencia de Schrödinger pero se puede solicitar una licencia libre para la enseñanza. Además, el código fuente abierto está disponible en GitHub que se puede instalar en Linux o MAC.

PyMOL permite trabajar con diferentes representaciones de estructuras, pero también con datos experimentales brutos en diferentes formatos.

PyMOL está escrito en Python y puede ser usado con menús interactivos y también con línea de comandos. Hay muchos recursos que ayudan con PyMOL, como un Wiki de referencia de documentación o un PyMOLWiki soportado por la comunidad. Además, permite la implementación de nuevas funcionalidades como plugins, como PyMod o DockingPie, entre otros. PyMod está diseñado para actuar como interfaz simple e intuitiva entre PyMOL y varias herramientas bioinformáticas (i.e., PSI-BLAST, Clustal Omega, HMMER, MUSCLE, CAMPO, PSIPRED, y MODELLER). Partiendo de la secuencia de aminoácidos de la proteína diana, PyMod está diseñado para llevar a cabo los principales pasos del proceso de modelado homológico (es decir, la búsqueda de plantillas, la alineación de la secuencia diana-plantilla y la construcción del modelo) con el fin de construir un modelo atómico 3D de una proteína diana (o complejo proteico). La integración con PyMOL facilita un análisis detallado del proceso de modelado.

#### III.4.3.2. UCSF ChimeraX

ChimeraX es un software totalmente de código abierto, desarrollado por la UCSF como versión renovada del antiguo software Chimera, con versiones para Linux, Mac OS y Windows. Pretende ser una herramienta integral de biología estructural, pero es más conocido por sus capacidades para mapas EM. Como cualquier otro software de código abierto, en los últimos años ha adquirido nuevas e interesantes capacidades, como la Realidad Virtual o el modelado AlphaFold2.

### III.4.3.3. Estructuras moleculares en su sitio web: Mol\* y otros

LiteMol Viewer es una potente aplicación web HTML5 para la visualización 3D de moléculas y otros datos relacionados. Se utiliza en un navegador web, lo que elimina la necesidad de software externo y también permite la integración con sitios de terceros como un plugin incrustado.

La misma filosofía se aplica a otros visores de código abierto que se desarrollaron más tarde y que ahora se utilizan más ampliamente, como NGL Viewer y Mol\*, utilizados en los sitios RCSB-PDB y PDBe para la visualización 3D de estructuras. Con Mol\* puede guardar la sesión de trabajo en formatos molj (sin las estructuras reales) o molx (con estructuras incrustadas).

Por último, para los científicos computacionales, también hay muchas bibliotecas que permiten la representación de moléculas en 3D, como la biblioteca 3Dmol Javascript y su envoltorio Python Py3Dmol, que se puede utilizar en Colab, Jupyter, Quarto o cualquier otro cuaderno Python.

Otra aplicación que se utilizaba, pero que ahora están descatalogada es SwissPDBViewer (también conocida como DeepView), desarrollada para trabajar con la aplicación de modelado homológico SWISS-MODEL. Es una aplicación que proporciona una interfaz fácil de usar que permite analizar varias proteínas al mismo tiempo. Actualmente ha caído en desuso ya que la última versión (4.1) es sólo una aplicación de 32 bits.