Capítulo 1

Macromoléculas biológicas: proteínas, DNA y RNA

Inmaculada Yruela y Alvaro Sebastián

El descubrimiento de las estructuras químicas del DNA y las proteínas, y de la relación entre ambas no fue tarea fácil en la historia de la biología. Desde que Mendel realizara sus experimentos a mitad del siglo XIX y Miescher aislara el DNA del esperma de salmón en 1869, tuvieron que pasar más de 70 años para descubrir la relación de los genes con el metabolismo. En 1941 Beadle y Tatum desarrollaron la hipótesis *"un gen, una enzima"* con sus experimentos sobre el metabolismo del moho *Neurospora* [2]. En 1944 Avery confirmó unos experimentos realizados por Griffith años antes que demostraban que el DNA era el material genético, hasta entonces se creía que las proteínas cargaban la información hereditaria. Unos años después Sanger [13] consiguió secuenciar parte de la secuencia de la insulina y Watson y Crick resolvieron el enigma de la estructura del DNA a partir de las imágenes de difracción de rayos X de Franklin [14]. Sin embargo se seguía sin conocer cuál era el código que permitía traducir DNA a proteínas. Hasta que casi 10 años más tarde, en 1961, Crick y Brenner demostraron que un codón consistía en 3 pares de bases de DNA [6], a la vez que Matthei y Nirenberg consiguieron dar con la clave del código genético [10] con la ayuda de la polinucleótido fosforilasa descubierta anteriormente por Severo Ochoa [7]. El código fue completado poco después por Nirenberg, Leder [9] y Khorana [8]. El descubrimiento de que 3 nucleótidos de DNA codificaban para la síntesis de 1 aminoácido fue un descubrimiento que revolucionó la ciencia. Actualmente comparando una secuencia de RNA y la secuencia de la proteína codificada sería muy sencillo descifrar el código genético y ver que 3 nucleótidos codifican 1 aminoácido, pero hace 50 años no existían dichas secuencias y llegaba a ser una tarea muy compleja, incluso filosófica. Finalmente, el método clásico de secuenciación de DNA que ha permitido conocer la secuencia de miles de genes y proteinas fue desarrollado por Sanger en 1975 [12] hasta la llegada hace pocos años de las nuevas tecnologías de secuenciación que permiten secuenciar genomas eucariotas completos en unas horas.

1.1. Genes y proteínas

Un *gen* es una secuencia de nucleótidos en la molécula de DNA (o RNA, en el caso de algunos virus) que contiene la información necesaria para que a través de una cascada de procesos biológicos se culmine la síntesis de una macromolécula con función celular específica, habitualmente proteínas pero también mRNA, rRNA y tRNA[1]. En el genoma humano se estiman pocos más de 20000 genes, siendo su secuencia codificante de proteínas sólo un 1.5 % de la longitud total del genoma [5]. Si pensamos que hasta hace pocos años todos los estudios se centraban en genes y proteínas, podemos concluir que conocemos muy poco acerca de nuestro genoma.

Volviendo al dogma fundamental de la biología *"un gen, una proteína"*, podemos describir los 3 procesos biológicos que permiten la síntesis de una proteína a partir de la información contenida en un gen en organismos eucariotas:

- *Transcripción*: Consiste en la síntesis de RNA a partir de una cadena molde de DNA genómico mediante la enzima RNA polimerasa, el RNA sintentizado se denomina mRNA o RNA mensajero.
- *Splicing* (o ajuste): El mRNA tras la transcripción es cortado y empalmado de forma que se eliminan secuencias llamadas intrones y se conservan secuencias llamadas exones (el *splicing* no sucede en procariotas).
- *Traducción*: Es el proceso durante el cual un mRNA maduro se usa de molde para sintetizar proteínas. Se lleva a cabo en los ribosomas, donde por cada 3 nucleótidos en el RNA (codón) se añade un aminoácido a la proteína que se sintetiza.

Los tres procesos están esquematizados en la Figura 1.1. Como se puede observar, en el proceso de *splicing* el pre-RNA pierde parte de la secuencia original del DNA (los *intrones*), conservando los *exones*, y en la traducción parte de los extremos del mRNA no se traducen a proteína (UTR o región no traducida). La complejidad de estos procesos, así como la variedad de combinaciones posibles nos hace distinguir en las bases de datos diferentes tipos de secuencias:

- *DNA genómico*: DNA tal y como se encuentra en el cromosoma o material genético objeto de estudio y que se replica de una generación a la siguiente.
- *Gen*: Consiste habitualmente, como ya se ha explicado, en una pequeña parte del DNA genómico que tiene la particularidad de poder ser transcrito a mRNA para sintetizar posteriormente una proteína.
- *Promotor*: Secuencia de DNA que no se transcribe y que precede a un gen facilitando el acoplamiento de factores de transcripción y otras proteínas fundamentales para el inicio de la transcripción.
- *cDNA* (o *DNA complementario*): Es la secuencia de mRNA maduro (sin intrones) en forma de DNA (en la secuencia sólo cambia la letra U por la T). Estas cadenas de DNA se sintetizan a partir del RNA mediante las enzimas transcriptasa reversa y DNA polimerasa. Se suele utilizar para la secuenciación y clonación de genes.
- *EST* (o *marcador de sequencia expresada*): Es una sub-secuencia de un cDNA. Se pueden usar para identificar genes que se transcriben y en secuenciación y descubrimiento de genes.

[1] Definición modificada de Wikipedia. `http://es.wikipedia.org/wiki/Gen`

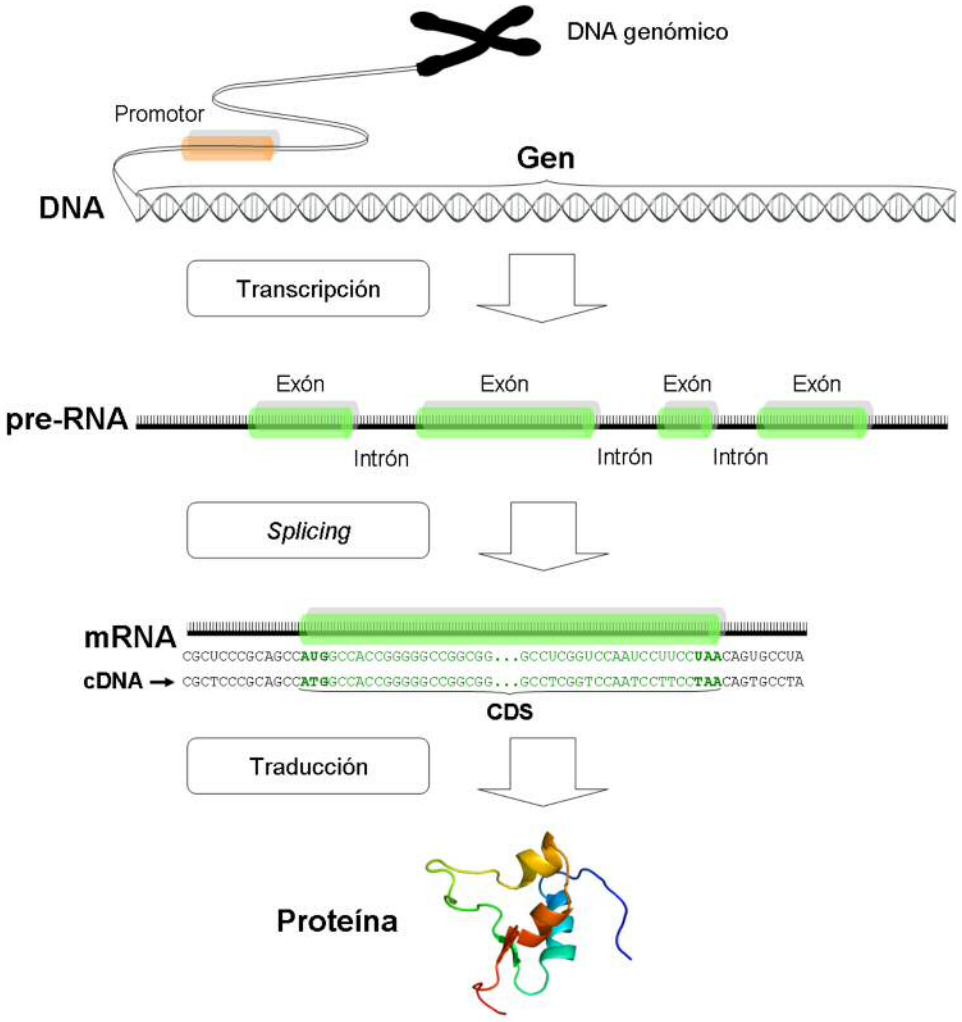

Figura 1.1: Esquema de los procesos de transcripción, *splicing* y traducción que conducen a la síntesis de una proteína a partir de una secuencia génica de DNA.

- *CDS* (*secuencia codificante*): Consiste en la parte de un gen que codifica una proteína. Es la secuencia de cDNA formada por exones y comprendida entre las UTRs o regiones no traducidas.
- *ORF* (o *marco abierto de lectura*): A veces no se conoce dónde comienzan y terminan las UTRs y no es posible asignar una CDS a un gen. En estos casos se predicen una o varias secuencias ORFs entre los supuestos inicios y finales de la traducción. Frecuentemente se confunde ORF con CDS, sin embargo la CDS real formará parte de una de las ORFs predichas.
- *Proteína*: Una proteína puede estar formada por varias cadenas peptídicas, pero vulgarmente se define como proteína a la secuencia de aminoácidos de un péptido fruto de la traducción del mRNA maduro.

1.2. Estructura primaria de DNA y RNA

El *DNA o ácido desoxirribounucleico* y el *RNA o ácido ribonucleico* son polímeros compuestos por unidades repetidas llamadas nucleótidos. Cada *nucleótido* está compuesto de una base nitrogenada, un azúcar de 5 carbonos que puede ser ribosa o 2-desoxiribosa y un grupo fosfato (Figura 1.2A). En los extremos del nucleótido quedan un grupo fosfato unido a la posición 5' del azúcar y un grupo OH en posición 3' del azúcar (Figura 1.2A), dichos extremos determinan la dirección de la secuencia, generalmente en biología se trabaja con las secuencias en dirección $5' \rightarrow 3'$.

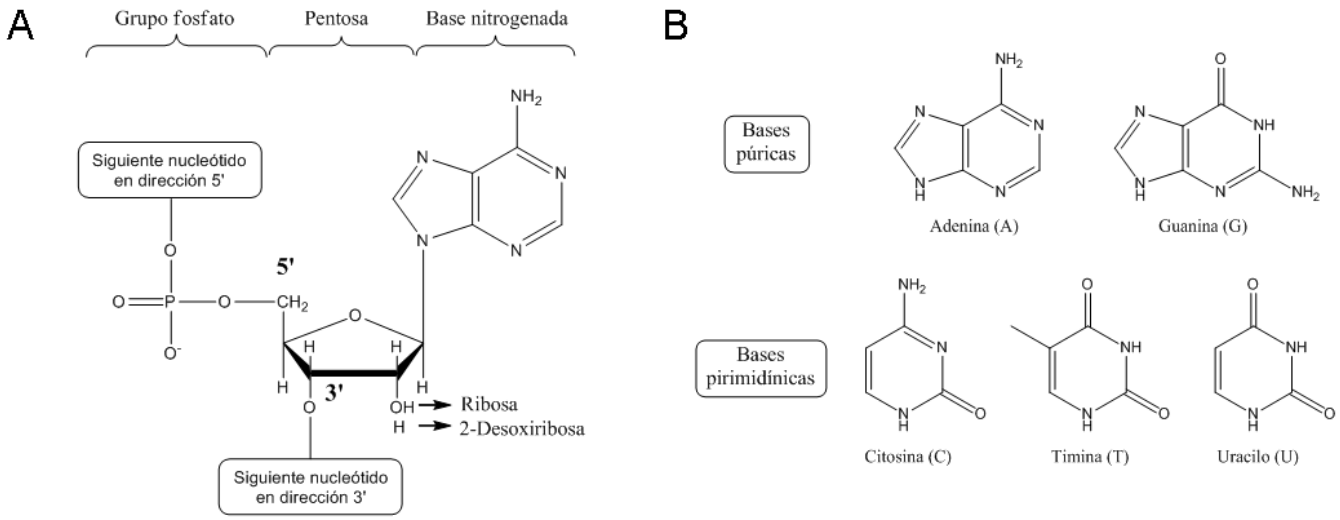

Figura 1.2: A: Estructura de un nucleótido. B: Letra, nombre y fórmula de las bases nitrogenadas.

Las *bases nitrogenadas* que conforman los nucleótidos pueden ser de dos tipos: púricas y pirimidínicas según sea el anillo de purina o el de pirimidina el núcleo de su estructura. Tanto en DNA como en RNA están presentes las bases púricas llamadas *adenina* (A) y *guanina* (G) y la base pirimidínica *citosina* (C) (Figura 1.2B). Existe otra base pirimidínica denominada *timina* (T) que sólo conforma el DNA y que su equivalente en el RNA es el *uracilo* (Figura 1.2B). La particularidad de estas bases es que pueden formar enlaces de hidrógeno intermoleculares entre adenina y timina y entre guanina y citosina. Estas interacciones intermoleculares permiten al DNA formar una estructura de doble hélice formada por dos cadenas sencillas enroscadas en direcciones opuestas (antiparalelas,

una en dirección $5' \rightarrow 3'$ y otra $3' \rightarrow 5''$) donde las bases nitrogenadas se encuentran en el interior formando enlaces de hidrógeno entre pares A-T y G-C. Debido a la estructura de doble cadena, cuando hacemos referencia al tamaño de una doble hélice de DNA solemos hablar de *pares de bases*, así el genoma humano tiene un tamaño aproximado de 3200 millones de pares de bases. El RNA también puede formar apareamientos de bases pero tiene generalmente estructura monocatenaria. Cada nucleótido según la base nitrogenada que lo compone se designa con la letra asignada a dicha base (Figura 1.2B) y la estructura polimérica se representa con una succesión de letras en sentido $5' \rightarrow 3'$ denominada *estructura primaria*, ver ejemplo en la Figura 1.3. También existen letras para designar combinaciones de varias bases en una determinada posición de la secuencia[2], a utilizar por ejemplo cuando existe un 50 % de probabilidad de una u otra base en dicha posición.

```
5'-ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAG
CCGCAGCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGG
GGAACGAGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTG
GAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGC
GTGGCATTGTGGAACAATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAAC-3'
```

Figura 1.3: Ejemplo de secuencia de la ORF (cDNA) del gen precursor de la insulina humana (INS).

1.3. El código genético

El código genético es el conjunto de reglas que permite a las células traducir la información del material genético (DNA o mRNA) en aminoácidos constituyentes de las proteínas. Como ya se ha mencionado, los genes se dividen en secuencias de 3 nucleótidos llamadas codones y cada codón codifica para un aminoácido en la traducción a proteína del mRNA. El código genético representado en la Tabla 1.1 es universal, aunque tiene excepciones en diferentes organismos y orgánulos celulares. Tres características fundamentales del código genético son:

- *Es un código degenerado*: varios codones pueden codificar el mismo aminoácido, por ejemplo el aminoácido serina es codificado por 6 codones, glicina por 4, sin embargo metionina sólo es codificado por 1 codón. La degeneración del código permite que determinadas mutaciones en el DNA no causen modificaciones en las proteínas codificadas.
- *Codones de inicio y parada*: según el organismo determinados codones indican el comienzo y final de la traducción. El codón de inicio más habitual es AUG que codifica metionina, mientras que los que la finalizan son: UAA, UAG, UGA.
- *Marco de lectura*: una secuencia de DNA o mRNA debe ser leída de 3 en 3 nucleótidos en el orden correcto, dicho orden se denomina marco de lectura. Una doble hélice de DNA contine 6 posibles marcos de lectura (3 para cada cadena) por ello es muy importante conocer cuál es el que dará como resultado la secuencia proteica adecuada.

En el ejemplo de la Figura 1.4 se pueden observar los codones, el marco de lectura y los aminoácidos que codifican en el gen de la insulina humana. La conversión de secuencias de DNA a proteína en

[2]Códigos IUPAC para nucléotidos y aminoácidos. `http://www.bioinformatics.org/sms/iupac.html`

los tres marcos de lectura y de proteína a DNA se puede realizar con numerosas herramientas disponibles online[3].

PRIMERA BASE	SEGUNDA BASE: U	SEGUNDA BASE: C	SEGUNDA BASE: A	SEGUNDA BASE: G	TERCERA BASE
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA FIN	UGA FIN	A
	UUG Leu	UGG Ser	UAG FIN	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CGG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	AGG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GGG Ala	GAG Glu	GGG Gly	G

Tabla 1.1: Código genético representado por tripletes de bases y los aminoácidos que codifican.

```
atggccctgtggatgcgcctcctgcccctgctggcgctgctggccctctggggacctgacccagcc
 M   A   L   W   M   R   L   L   P   L   L   A   L   L   A   L   W   G   P   D   P   A
gcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacctagtgtgcggg
 A   A   F   V   N   Q   H   L   C   G   S   H   L   V   E   A   L   Y   L   V   C   G
gaacgaggcttcttctacacacccaagacccgccgggaggcagaggacctgcaggtggggcaggtg
 E   R   G   F   F   Y   T   P   K   T   R   R   E   A   E   D   L   Q   V   G   Q   V
gagctgggcgggggccctggtgcaggcagcctgcagcccttggccctggaggggtccctgcagaag
 E   L   G   G   G   P   G   A   G   S   L   Q   P   L   A   L   E   G   S   L   Q   K
cgtggcattgtggaacaatgctgtaccagcatctgctccctctaccagctggagaactactgcaac
 R   G   I   V   E   Q   C   C   T   S   I   C   S   L   Y   Q   L   E   N   Y   C   N
```

Figura 1.4: Traducción del ORF de la insulina humana mostrada en la Figura 1.3 a su secuencia proteica de la Figura 1.11 mostrando el aminoácido codificado por cada codón.

[3]Herramientas de traducción de secuencias en la página web del EMBL-EBI. `http://www.ebi.ac.uk/Tools/st/`

1.4. Aminoácidos y enlace peptídico

1.4.1. Aminoácidos

Las *proteínas* están compuestas por uno o más polipéptidos. Cada *polipéptido* es un polímero formado por aminoácidos unidos entre sí mediante enlaces peptídicos. Un *aminoácido* es un compuesto químico que cuenta con un *grupo amino* (-NH_2) de carácter básico y un *grupo carboxilo* (-$COOH$) de carácter ácido unidos a un mismo carbono llamado *carbono alfa* (Cα) que además une un hidrógeno y una *cadena lateral* variable (-R) que distingue cada uno de los 20 aminoácidos presentes en los seres vivos (Figura 1.5). La glicina es el aminoácido más simple y pequeño donde el grupo R es un hidrógeno (H).

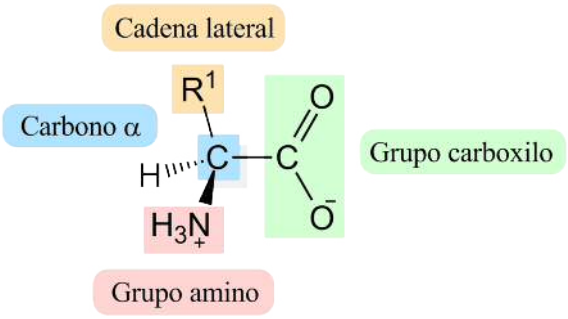

Figura 1.5: Estructura general de los α-aminoácidos que forman las proteínas. El carbono α (Cα) tiene cuatro sustituyentes: un átomo de hidrógeno (H), un grupo carboxilo (-COOH), un grupo amino (-NH_2), y un grupo R que es diferente en cada uno de los aminoácidos. Los grupos amino y carboxilo se representan en su forma ionizada predominante a pH neutro.

Los *aminoácidos* tienen un papel central como elementos constitutivos de las proteínas pero también juegan un importante papel como intermediarios en el metabolismo celular. *Los humanos producimos 10 de los 20 aminoácidos que forman las proteínas*, los otros 10 deben ser adquiridos a través de la alimentación. Esto es debido a que no tenemos los enzimas requeridos para la biosíntesis de todos los aminoácidos. Los 10 aminoácidos que podemos producir son alanina, asparagina, ácido aspártico (aspartato), cisteina, ácido glutamico (glutamato), glutamina, glicina, prolina, serina y tirosina. La tirosina se produce a partir de la fenilalanina. Si la dieta es deficiente en fenilalanina debe ser adquirida también. Las plantas son capaces de sintetizar todos los aminoácidos.

Los aminoácidos se suelen clasificar en grupos atendiendo a las *propiedades fisicoquímicas de las cadenas laterales*. Es importante señalar, que cualquier clasificación ordena y organiza los elementos de un conjunto según unas reglas, de forma que facilite y simplifique el estudio. Por tanto, podemos encontrarnos clasificaciones que difieren ligeramente en el criterio de clasificación y agrupan los aminoácidos de distintas maneras (ej. polaridad, pH, tipo de grupo en la cadena lateral, etc.). Lo importante, por tanto, es conocer la estructura y propiedades de cada uno de ellos.

Los 20 aminoácidos mayoritarios se pueden clasificar de acuerdo a la estructura y propiedades físico-químicas de su cadena lateral en 4 grupos (Figura 1.6):

- *Neutros apolares*: cadenas con residuos poco polares que repelen a las moléculas de agua (hidrofóbicas), son: glicina, alanina, valina, leucina, isoleucina, metionina, prolina, fenilalanina y triptófano.

- *Neutros polares*: cadenas con residuos polares e hidrófilos que pueden formar interaciones débiles con moléculas de agua, son: serina, treonina, cisteína, tirosina, asparagina y glutamina.
- *Ácidos*: cadenas con residuos ácidos cargados negativamente a pH fisiológico, capaces de protonar moléculas de agua y disminuir el pH, son: ácido aspártico y ácido glutámico.
- *Básicos*: cadenas con residuos básicos cargados positivamente a pH fisiológico, capaces de 'robar' protones a las moléculas de agua y aumentar el pH, son: lisina, arginina e histidina.

También se puede distinguir otro grupo, los *aromáticos*, que son aminoácidos ya clasificados en las categorías anteriores que contienen en su cadena lateral un anillo aromático: fenilalanina, tirosina y triptófano.

Las propiedades físico-químicas de los aminoácidos no sólo determinan su actividad biológica sino también el *plegamiento de la proteína* en una estructura tridimensional y la *estabilidad de la estructura* resultante. Algunas propiedades también ayudan a la *caracterización bioquímica* de las proteínas. Por ejemplo, los aminoácidos aromáticos son los responsables de la *absorbancia a 280 nm*, típica de las proteínas. Con la medida de la absorbancia a esta longitud de onda obtenida en un espectrofotómetro se puede determinar, de forma no destructiva, la concentración de una disolución proteíca. A continuación haremos un repaso de las propiedades particulares de cada uno de los 20 aminoácidos esenciales ordenados alfabéticamente.

- *Alanina* (Ala, A): es una molécula hidrofóbica ambivalente, es decir, puede encontrarse tanto en el interior como en el exterior de una proteína. El Cα de la alanina es ópticamente activo, pero en proteínas sólo se encuentra el isómero L. Tiene como cadena lateral un grupo metilo (-CH_3) y se utiliza como referencia para determinar las propiedades de los demás aminoácidos, que pueden considerarse derivados de él. Al ser apolar no participa en ningún mecanismo catalítico por lo que su función es meramente estructural. La alanina es un α-aminoácido análogo al α-cetopiruvato, un intermediario en el metabolismo de los azúcares, por lo que alanina y piruvato son intercambiables por una reacción de transaminación.
- *Arginina* (Arg, R): tiene un grupo guanidinio que capta protones y está cargado positivamente a pH neutro. Puede unir anión fosfato y a menudo se encuentra en los centros activos de las proteínas que unen sustratos fosforilados. Como catión, la arginina, así como la lisina, tiene un papel importante en el mantenimiento del balance de cargas de una proteína. La arginina tiene además un papel importante en el metabolismo del nitrógeno. En el código genético hay 6 codones para arginina. Sin embargo, aunque este elevado número de codones generalmente se asocia con una alta frecuencia del aminoácido en la secuencia de proteínas, la arginina es uno de los aminoácidos menos frecuentes. Esta discrepancia entre su frecuencia y el número de codones es mayor para arginina que para otros aminoácidos.
- *Asparagina* (Arg, N): es la amida del ácido aspártico. El grupo amida no soporta una carga formal bajo ninguna condición de pH biológicamente relevante, por lo que se mantiene neutro a pHs fisiológicos. El grupo amida se hidroliza fácilmente convirténdose la asparagina en ácido aspártico. Esta reacción se considera que está relacionada con las bases del envejecimiento. La asparagina suele formar puentes de hidrógeno dado que el grupo amida puede aceptar y donar dos hidrógenos. Se suele encontrar tanto en la superficie como en el interior de las proteínas. La asparagina es un sitio frecuente de unión de carbohidratos en glicoproteínas.
- *Ácido aspártico* (o aspartato, Asp, D): es uno de los dos ácidos, junto al ácido glutámico (o glutamato), que tienen una función importante en los centros activos de los enzimas, además de

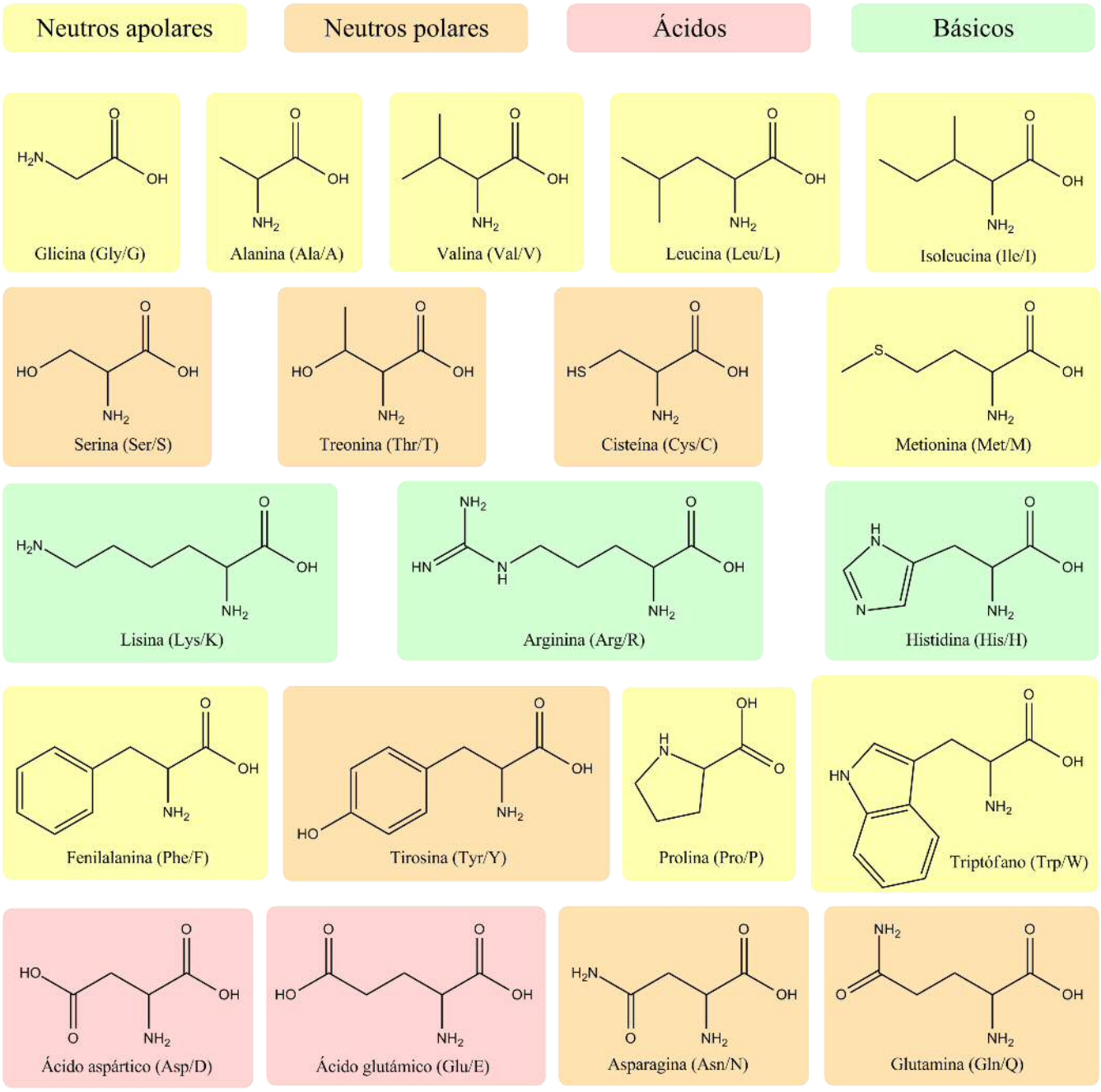

Figura 1.6: Fórmulas y nomenclatura de los 20 aminoácidos.

mantener la solubilidad y el carácter iónico de las proteínas. El pK_a del grupo β-carboxílico del ácido aspártico es $\approx 4{,}0$. Participa en mecanismos de catálisis ácido/base. El ácido aspártico es homólogo al oxaloacetato, de igual modo que el piruvato lo es de la alanina. Así, el ácido aspártico y el oxaloacetato son interconvertibles por una simple reacción de transaminación.

- *Cisteína* (Cys, C): es uno de los dos aminoácidos que contienen átomos de azufre; el otro es la metionina (Met, M). La cisteína difiere de la serina en un átomo de azufre (S) en el grupo tiol que sustituye al oxígeno (O) del alcohol. Si no está ionizada no es tan polar como la serina o la treonina. Es de resaltar que la cisteína y la serina son mucho más diferentes en sus propiedades fisico-químicas que lo que sugiere su similitud molecular. Por ejemplo, el protón del grupo tiol (-SH) de la cisteína es mucho más ácido que el protón del grupo hidroxilo (-OH) de la serina, lo que hace que el primero sea más reactivo que el segundo. La cisteína se puede oxidar y condensarse con otra cisteína formando un *puente disulfuro* (-S-S-) (Figura 1.7) que tiene carácter covalente y sirve para entrecuzar dos regiones de una proteína o dos subunidades de un complejo proteíco y mantener una conformación estable. El ambiente dentro de la célula es demasiado reductor para formar puentes disulfuros pero en el espacio extracelular los puentes disulfuro se pueden formar y estabilizar muchas proteínas.

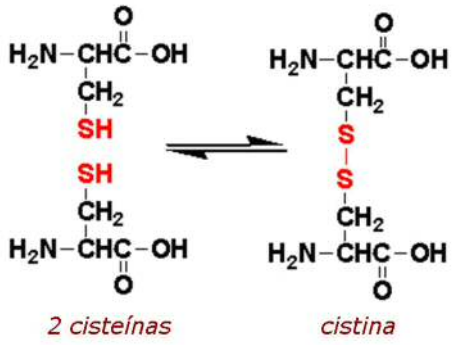

Figura 1.7: Puente disulfuro entre dos residuos de cisteína. La oxidación de dos cisteínas próximas y con orientación adecuada ocurre de forma espontánea dando lugar a un enlace covalente entre dos átomos de azufre (S). En ocasiones dos cisteínas unidas por un puente disulfuro se denomina cistina.

- *Ácido glutámico* (o glutamato, Glu, E): tiene un grupo metileno adicional en su cadena lateral comparado con el ácido aspártico, y se denomina grupo carboxílico γ. Su pK_a es 4,3, significativamente más alto que el del ácido aspártico. Esto es debido al efecto inductivo del grupo metileno adicional. El ácido glutámico y el α-cetoglutarato, un intermediario en el *ciclo de Krebs*, son interconvertibles por transaminación. Por lo tanto, el ácido glutámico puede entrar en el ciclo de Krebs y ser convertido por la enzima glutamina sintetasa en glutamina, que juega un papel importante en el metabolismo del nitrógeno. También es de resaltar que el ácido glutámico puede fácilmente convertirse en prolina. En un primer paso, el grupo carboxílico γ se reduce a aldehído, produciendo glutamato semialdehído que posteriormente reacciona con el grupo α-amino, eliminando agua y formando una *base de Schiff*. En un segundo paso de reducción, la base de Schiff se reduce dando lugar a la prolina.

- *Glutamina* (Gln, Q): es la amida del ácido glutámico y se encuentra no cargada en todas las condiciones biológicas. El grupo metileno adicional en su cadena lateral, comparado con la as-

paragina le permite en su forma libre o como N-terminal de la cadena proteíca espontáneamente ciclar y formar una estructura de anillo de pirrolidona que se encuentra en el N-terminal de muchas inmunoglobulinas. Este detalle dificulta la determinación de la secuencia de aminoácidos en experimentos de secuenciación.

- *Glicina* (Gly, G): es el más pequeño de los aminoácidos. Su cadena lateral R es un átomo de hidrógeno, por lo que dos de los cuatro sustituyentes del Cα son iguales y no presenta isomería óptica. Es el aminoácido que más flexibilidad proporciona a las proteínas, pues su pequeño tamaño no obstaculiza el movimiento de los aminoácidos que lo flanquean. Es ambivalente, es decir, se puede encontrar tanto en la superficie como en el interior de las proteínas.

- *Histidina* (His, H): tiene un grupo imidazol en su cadena lateral R que es capaz de captar protones y tener carga positiva a pH neutro. Participa generalmente en reacciones catalizadas por enzimas. La forma desprotonada del grupo imidazol tiene un carácter nucleofílico y actúa como base, mientras que la forma protonada actúa como ácido. La histidina tiene un papel en la estabilidad del plegamiento de las proteínas.

- *Isoleucina* (Ile, I): como su nombre indica es un isómero de la leucina, y junto a ésta y la valina es uno de los tres aminoácidos que tienen cadenas laterales R ramificadas. Generalmente se intercambia con leucina y ocasionalmente con la valina. Las cadenas laterales de estos aminoácidos no son reactivas y por lo tanto no participan en uniones covalentes y en la actividad catalítica de los centros activos de los enzimas. Sin embargo, estos aminoácidos confieren cierta rigidez a las proteínas, debido a la presencia de un segundo carbono asimétrico en la cadena lateral, que tiene siempre configuración S. Estos aminoácidos también tienen relevancia en la unión de ligandos en las proteínas y juegan un papel importante en la estabilidad de las mismas. El Cβ de la isoleucina, al igual que el de la treonina, es ópticamente activo y por tanto ambos son centros quirales.

- *Leucina* (Leu, L): al igual que la isoleucina y la valina, tiene una cadena lateral ramificada y un grupo metileno (-CH_2) adicional comparado con la valina. Tiene un carácter hidrofóbico y generalmente se encuentra en el interior de las proteínas.

- *Lisina* (Lys, K): tiene un grupo ϵ-amino positivamente cargado. Básicamente la lisina es una alanina con un sustituyente propilamino en el carbono-β. El grupo ϵ-amino tiene un pK_a significativamente más alto (ca. 10,5) que aquellos en grupos α-amino. El grupo ϵ-amino de la lisina es muy reactivo y a menudo participa en las reacciones catalíticas de los centros activos de los enzimas. La lisina a menudo se encuentra en el interior de las proteínas con sólo el grupo ϵ-amino expuesto al solvente.

- *Metionina* (Met, M): es un tioéter que incluye un átomo de azufre (S) en la cadena lateral R. La cadena lateral es muy hidrofóbica y generalmente se encuentra en el interior de las proteínas. Sus propiedades fisico-químicas están ligadas al átomo de azufre; no tiene un carácter muy nucleofílico aunque a veces reacciona con centros electrofílicos. De esta manera, aunque la metionina se considera un aminoácido apolar el átomo de azufre (S) aparece ocasionalmente implicado en reacciones bioquímicas en algunas enzimas. El átomo de azufre en la metionina, al igual que en la cisteína tiende a oxidarse. El primer paso da lugar a metionina sulfóxido y en un segundo paso irreversible se produce metionina sulfato. Esta reacción se piensa que es la causante del enfisema pulmonar en fumadores.

- *Fenilalanina* (Phe, F): es un derivado de la alanina con un sustituyente fenilo en el Cβ. Es muy hidrofóbico, incluso la fenilalanina libre no es soluble en agua. Como dato curioso es de

mencionar que Marshall Nirenberg y Phil Leder en 1964, cuando realizaban sus experimentos para determinar el triplete del código genético que estaba asociado a cada aminoácido, encontraron que el producto del mensajero sintético polyU, era insoluble. En aquellos momentos no sabían que 'UUU' codifica a fenilalanina, pero poco después de estos experimentos precipitaron el producto formado en su mezcla; ya estaban en el camino de desentrañar el código genético y las bases de la síntesis de proteínas, y conseguir el Premio Nobel de Fisiología y Medicina en 1968.

- *Prolina* (Pro, P): comparte muchas propiedades con los aminoácidos alifáticos. En sentido estricto no es un aminoácido sino un iminoácido, pues la cadena lateral termina uniéndose por su extremo con el grupo α-amino. Por ello la prolina es particularmente rígida y, además, su grupo amino, cuando forma parte de un enlace peptídico no puede actuar como donador en puentes de hidrógeno para estabilizar hélices-α o láminas-β. A menudo se dice que la prolina no puede encontrarse en una hélice-α. Cuando ello ocurre la hélice tendrá una leve curva debido a la falta del enlace del hidrógeno.

- *Serina* (Ser, S): difiere de la alanina en que uno de los hidrógenos del grupo metilo (-CH_3) es sustituido por un grupo hidroxilo (-OH). Es una molécula hidrofílica y suele encontrarse en el centro catalítico de muchos enzimas por la capacidad del grupo –OH de actuar como nucleófilo y atacar a grupos deficientes en electrones de otras moléculas, promoviendo cambios químicos como la hidrólisis.

- *Treonina* (Thr, T): es una molécula hidrofílica y al igual que la serina contiene un grupo hidroxilo (-OH), pero difiere de ella en que tiene un sustituyente metilo (-CH_3) en vez de un hidrógeno en el Cβ. Tanto el Cα como el Cβ de la treonina son ópticamente activos.

- *Triptófano* (Trp, W): es el aminoácido de mayor tamaño. Es un derivado de la alanina, pues tiene un sustituyente indol en el Cβ. El grupo indol absorbe en el ultravioleta (UV) lejano. La nube π del anillo aromático puede actuar como aceptor de puentes de hidrógeno y también puede formar interacciones con grupos cargados positivamente (ej. cadenas laterales de aminoácidos básicos o ligandos catiónicos).

- *Tirosina* (Tyr, Y): es también un aminoácido aromático derivado de la fenilalanina por una hidroxilación en la posición *para*. Aunque la tirosina tiene carácter hidrofóbico debido a su anillo aromático, es significativamente más soluble que la fenilalanina y se considera polar y protonable. El hidroxilo del grupo fenólico es más ácido que el hidroxilo alifático de la serina o la treonina, teniendo un pK_a de ca. 9,8. Al igual que todos los grupos ionizables el pKa preciso depende del ambiente donde se encuentre la proteína. Las tirosinas que se encuentran en la superficie de las proteínas generalmente tienen un pK_a más bajo que las que se encuentran en el interior. La ionización de la tirosina da lugar al anión fenolato que puede ser inestable en el interior hidrofóbico de una proteína. La tirosina absorbe la radiación UV y contribuye a la absorbancia espectral de las proteínas al igual que lo hacen los demás aminoácidos aromáticos (triptófano y fenilalanina). Sin embargo, su contribución es menor, el coeficiente de extinción de la tirosina es ca. 1/5 que el del triptófano a 280 nm. El triptófano es el que más contribuye a la absorbancia en el UV en las proteínas (dependiendo del número de residuos que haya en la secuencia).

- *Valina* (Val, V): es una molécula hidrofóbica y se encuentra en el interior de las proteínas. Al igual que isoleucina confiere cierta rigidez a las proteínas al presentar una ramificación de la cadena lateral del Cβ. Los electrones π del anillo de fenilo favorecen el apilamiento con otras

moléculas aromáticas y a menudo lo hacen en regiones enterradas de las proteínas, añadiendo estabilidad a la estructura. Es de señalar que la valina y la treonina tienen una estructura y volumen similar por lo que es difícil, incluso a alta resolución distinguirlas.

1.4.2. Enlace peptídico

El *enlace peptídico* une dos aminoácidos mediante la condensación del grupo amino y el grupo carboxilo para formar una amida (R_A-NH-CO-R_B) (Figura 1.8). De esta forma, en los extremos de los péptidos quedan un residuo amino y otro carboxilo que serán los únicos que no formen enlaces peptídicos. Por convenio los péptidos se escriben y representan con el extremo N-terminal a la izquierda y el extremo C-terminal a la derecha.

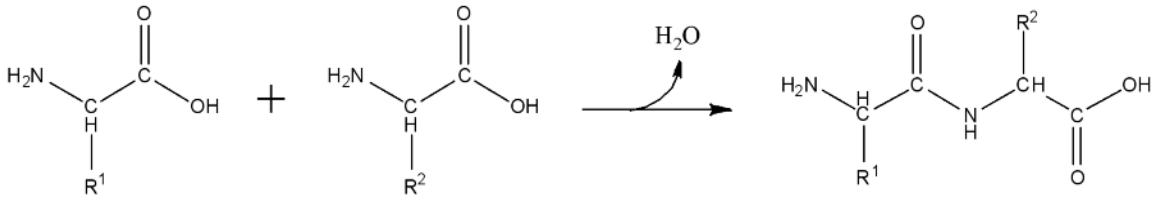

Figura 1.8: Formación de un enlace peptídico entre dos aminoácidos.

Todos los aminoácidos tienen un grupo ácido y un grupo amino ionizables. Sin embargo, para comprender el comportamiento en solución de las proteínas hay que considerar que ambos grupos desparecen cuando se forma el enlace peptídico entre el grupo carboxílico (-COOH) de un aminoácido y el grupo amino (-NH_2) del aminoácido adyacente, y se polimerizan para formar las proteínas en el ribosoma. Sólo los grupos carboxílico y amino que ocupan los extremos de la cadena proteica quedan libres. Pero a pesar de ello, *las proteínas presentan numerosos grupos cargados en las cadenas laterales R que son ionizables*. Estos aminoácidos hacen que sea posible la *solubilidad en agua* de muchas proteínas. Las curvas de titulación de los aminoácidos ayudan a comprender cómo varía su carga en función del pH. Para los aminoácidos ionizables, el *punto isoeléctrico* es la media aritmética, aproximadamente de los pK_a que delimitan el intervalo de pH en que predomina la forma del aminoácido sin carga neta.

El *enlace peptídico es un enlace covalente de tipo amida* resultado de una reacción de condensación con la pérdida de una molécula de H_2O. La condensación no es espontánea, al tener lugar en un disolvente acuoso, por lo que requiere aporte de energía. El enlace peptídico puede romperse por *hidrólisis* (adición de una molécula de agua) liberando 8–16 kJ/mol (2–4 kcal/mol) de energía libre. Este proceso es muy lento y es facilitado por los enzimas. La longitud de onda de absorbancia del enlace peptídico es 190–230 nm lo que hace que sea particularmente sensible a la radiación UV. El enlace peptídico presenta cierto carácter de enlace doble que determina gran parte de las propiedades conformacionales de las proteínas. La mayoría de los enlaces peptídicos de las proteínas son de *isomería trans*. La forma *trans* es la de menor energía, sitúa a los sustituyentes voluminosos (Cα con las cadenas laterales) más alejados y por tanto con menor conflicto estérico.

1.4.3. Ángulos de torsión y diagrama de Ramachandran

El enlace peptídico introduce restricciones que limitan las conformaciones que puede adoptar un polipéptido o proteína. La flexibilidad de estos depende del número de sus enlaces que permiten rotaciones de una parte de la molécula respecto al resto. Debido a su carácter de doble enlace la rotación completa no está permitida por lo que la rotación en torno al enlace peptídico (ω) sólo puede tener los valores 0°(conformación *cis*) y 180°(conformación *trans*). Cualquier otro valor sería desestabilizante. Los *ángulos de torsión*, también conocidos como *ángulos de Ramachandran*, describen las rotaciones del esqueleto de la cadena polipeptídica alrededor de los enlaces entre N y Cα (ver Figura 1.9. El *ángulo Phi* (φ) está definido por los átomos C-N-Cα-C (C es el carbono carbonilo) y Cα–C, y el *ángulo Psi* (ψ) está definido por los átomos N-Cα-C-N. Los ángulos φ y ψ de un aminoácido están restringidos por repulsiones estéricas.

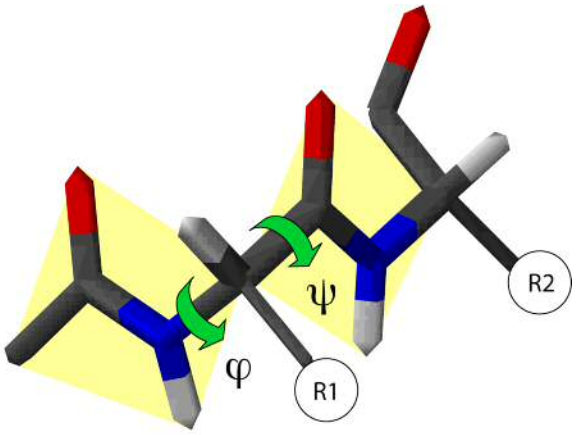

Figura 1.9: Ángulos de torsión en una cadena polipeptídica. Los ángulos de torsión son ángulos dihedros definidos por cuatro puntos en el espacio. Los ángulos Phi (φ) y Psi (ψ) describen la rotación de la cadena alrededor de los enlaces a ambos lados del átomo Cα [4].

El *diagrama de Ramachandran* es una sencilla representación de la distribución de los ángulos de torsión en la estructura de una proteína. Existen mayoritariamente dos zonas de ángulos φ y ψ permitidos, ambas con valores negativos del ángulo φ, y que se diferencian por el valor positivo o negativo del ángulo ψ. Cada punto (φ, ψ) en el diagrama representa a un aminoácido de la cadena polipeptídica (ver ejemplo en Figura 1.10).

El diagrama de Ramachandran permite tener una visión de las regiones de torsión que están permitidas y no permitidas, y por tanto de la flexibilidad de una región para adoptar un determinado plegamiento. No todas las parejas de ángulos son posibles dentro de las estructuras de los péptidos y proteínas, debido a los *efectos estéricos* entre los residuos (cadenas laterales) de los aminoácidos. Por tanto, el conocer los valores de los ángulos de torsión nos sirve como *indicador de la calidad de la estructura 3D de una proteína*. Los ángulos de torsión son uno de los parámetros estructurales más importantes que controlan el plegamiento de una proteína. Si somos capaces de predecir los ángulos

[4]Gráfico de Ramachandran en Wikipedia. `http://es.wikipedia.org/wiki/Gr\penalty\@M\hskip\z@skip\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{a\global\mathchardef\accent@spacefactor\spacefactor}\accent19a\egroup\spacefactor\accent@spacefactor\penalty\@M\hskip\z@skip\setbox\@tempboxa\hbox{a\global\mathchardef\accent@spacefactor\spacefactor}\spacefactor\accent@spacefactorfico_de_Ramachandran`

de Ramachandran de una proteína seríamos capaces de predecir su plegamiento y su estructura secundaria. Esto es así porque el tercer posible ángulo de torsión dentro de una proteína (omega, ω) está prácticamente fijo a 180°, debido al carácter parcial de doble enlace del enlace peptídico, que restringe la rotación alrededor del enlace C–N y coloca dos Cα contiguos y C, O, N e H entre ellos en un plano. Cuando la estructura de rayos X de una proteína no está adecuadamente refinada podemos encontrar ángulos de torsión en regiones no permitidas del diagrama de Ramachandran. En estos casos este tipo de desviaciones nos indica problemas con la estructura resultante de nuestro modelo o nuestra determinación.

El diagrama de Ramachandran también nos puede dar información sobre los diferentes elementos de *estructura secundaria* que componen la estructura 3D. En una estructura proteica se pueden distinguir regiones en las que los Cα de varios residuos consecutivos adoptan ángulos φ y ψ similares, hecho que hace que en estas regiones se encuentren disposiciones periódicas de sus unidades peptídicas. Estos elementos se pueden distinguir por el rango de valores de los ángulos φ and ψ. Las dos disposiciones periódicas tienen forma de hélice (hélice α) o de cadena extendida (lámina β). Las hélices α presentan valores promedio en el rango de φ -57°y ψ -47°, y las láminas β en el rango de φ -130°y ψ +140°. Cuando los ángulos de varios aminoácidos consecutivos se reparten entre las dos regiones no se forma una estructura secundaria. En estos casos se forma un bucle el que el polipéptido puede cambiar de dirección. Para adquirir estabilidad y generar complejas superficies de interacción las proteínas tienen regiones que carecen de la repetición de valores φ y ψ, característicos de las hélices α y láminas β, estas regiones se denominan *bucles*. Los bucles más cortos se denominan *giros reversos* y están formados por sólo cuatro aminoácidos. En los giros abundan las glicinas.

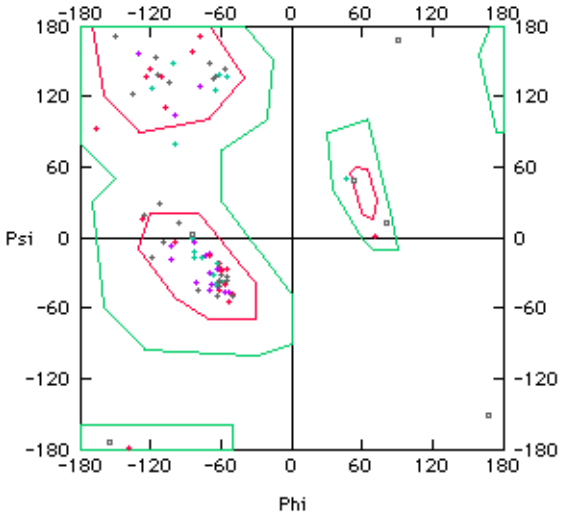

Figura 1.10: Diagrama de Ramachandran del citocromo b5 (PDB 3b5c). En esta estructura refinada sólo los residuos de glicina se encuentran fuera de los regiones permitidas (polígonos con línea verde).

1.5. Niveles estructurales en proteínas

Para describir, analizar y comprender la estructura de las proteínas se han definido varios niveles de complejidad ordenados jerárquicamente. De esta manera las proteínas tienen cuatro niveles

estructurales: i) *estructura primaria*, que corresponde a la secuencia de aminoácidos; ii) *estructura secundaria*, que se refiere a las disposiciones particulares estables de los aminoácidos, que dan lugar a patrones estructurales repetidos; iii) *estructura terciaria*, que describe la disposición tridimensional global de todos los átomos de una proteína; iv) *estructura cuaternaria*, que describe la disposición espacial de las subunidades polipeptídicas que componen una proteína.

1.5.1. Estructura primaria

Se define como *estructura primaria* de una proteína a las secuencias ordenadas de aminoácidos de los péptidos que la constituyen. A los aminoácidos dentro de un péptido o proteína también se les denomina *residuos*. Al igual que en el caso del DNA y el RNA, la secuencia proteica se representa con letras, en este caso con 20, una para cada aminoácido (Figura 1.6). Para facilitar la lectura también es muy común representar los residuos con secuencias de tres letras. El primer aminoácido por la izquierda representará el residuo N-terminal y el último por la derecha el C-terminal. El péptido representado se corresponderá con el péptido codificado por una secuencia de mRNA o cDNA en sentido $5' \rightarrow 3'$ (provenientes de una hebra molde de DNA genómico transcrita en sentido $3' \rightarrow 5''$), ver ejemplo en la Figura 1.11.

NH_2–MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRR
EAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN–$COOH$

Figura 1.11: Ejemplo de secuencia de la pre-insulina humana (INS) tal y como se traduce del cDNA mostrado en la Figura 1.3, antes de sufrir modificaciones post-traduccionales para formar la insulina activa.

Además de la estructura primaria, se definen otros tres niveles estructurales en las proteínas que serán explicados más adelante en el capítulo: estructura secundaria, terciaria y cuaternaria. Los péptidos constituyentes de las proteínas se pliegan formando estructuras tridimensionales y se asocian entre sí para formar proteínas totalmente funcionales. Una proteína no es activa en su forma de cadena peptídica lineal, ésta requiere un determinado plegamiento tridimensional inducido por las propiedades físico-químicas de sus residuos y/o ayudado por proteínas y factores externos para poder ejercer su función celular.

1.5.2. Estructura secundaria

Las proteínas naturales en disolución no son estables en conformaciones extendidas. En un solvente acuoso un polipéptido tiene muchos grupos capaces de formar puentes de hidrógeno y por tanto de manera espontánea se forman elementos de estructura secundaria que se conectan entre sí por lazos (*loops*) o por regiones intrínsecamente desordenadas o dúctiles. De esta manera la *estructura secundaria* se refiere a la conformación local de ciertas regiones de un polipéptido, es decir a los *patrones de plegamiento regulares que adopta la cadena polipeptídica*. Estos patrones son debidos a los *puentes de hidrógeno intramoleculares e intermoleculares* que establecen los grupos amida. La geometría que adopta la proteína está directamente relacionada con la geometría molecular del grupo amida. Sólo unas estructuras secundarias son estables y están ampliamente distribuidas en las proteínas: hélices α y láminas β. La estructura de hélice es la disposición más sencilla que puede

adoptar una cadena polipeptídica si tenemos en cuenta la rigidez de sus enlaces peptídicos y la libertad de rotación de los demás enlaces. Pauling y Corey (1951) predijeron esta estructura en base a imágenes de rayos X llamándola *hélice* α. En esta estructura el esqueleto polipeptídico se encuentra compactamente enrrollado a lo largo del eje imaginario longitudinal de la molécula y de los grupos R de los aminoácidos que sobresalen del esqueleto helicoidal. La hélice está estabilizada por puentes de hidrógeno entre el grupo -NH de un aminoácido y el grupo –C=O del cuarto aminoácido consecutivo. La unidad repetida es el giro de la hélice. En cada giro se disponen 3,6 aminoácidos, esto significa que hay un aminoácido cada 100 grados de rotación (360°/3,6) y que cada residuo se traslada 1,5 Å a lo largo del eje longitudinal de la hélice, dando lugar a una distancia vertical de 5.4 Å entre átomos estructuralmente equivalentes en un giro. En general, en las proteínas globulares se observa que una cuarta parte de los aminoácidos se encuentran formando hélices α aunque la proporción exacta varía en cada proteína. Ver ejemplo de hélice α en la Figura 1.12A

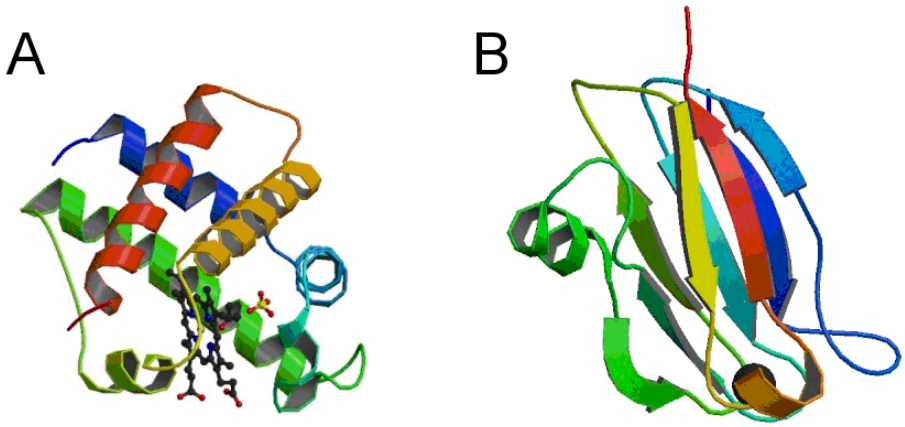

Figura 1.12: A. Estructura 3D de la leghemoglobina de *Glycine max* (soja, pdb 1BIN) a 2,2 Å de resolución. Esta hemo-proteína globular está formada mayoritariamente por hélices α. En la estructura se observa la presencia del grupo hemo como cofactor. B. Estructura 3D de la plastocianina de *Anabaena variabilis* (pdb 2CJ3) a 1.7 Å de resolución. Esta proteína está formada mayoritariamente por láminas β. En la estructura se observa la presencia de un átomo de Cu como cofactor.

Sin embargo, no todos los polipéptidos pueden formar una hélice α estable. Las interacciones entre las cadenas laterales de los aminoácidos pueden estabilizar o desestabilizar la estructura (ej. una cadena con muchos Glu no podrá formar una hélice α a pH 7,0 porque los grupos –COOH cargados negativamente de los Glu adyacentes se repelen con mayor intensidad que la atracción estabilizadora de los puentes de hidrógeno). La presencia de Pro o Gly también restringe la formación de hélices α. Existen cinco tipos de *restricciones que afectan a la estabilidad de una hélice* α: i) la repulsión o atracción electrostática entre aminoácidos consecutivos con grupos R cargados; ii) el volumen de los grupos R adyacentes; iii) las interacciones de las cadenas laterales de aminoácidos separadas 3 o 4 aminoácidos; iv) la presencia de Pro y Gly; v) la interacción entre aminoácidos en los extremos de la hélice y el dipolo eléctrico de la estructura.

La conformación de *lámina* β en la cadena polipeptídica se encuentra extendida en zigzag en lugar de plegarse como una hélice. Los segmentos individuales se mantienen unidos por puentes de hidrógeno intermoleculares entre los grupos amida de dos segmentos separados. En este caso, los puentes de hidrógeno no se forman entre aminoácidos adyacentes como en las hélices α sino entre aminoácidos de segmentos cercanos en la cadena polipeptídica. En algunos casos también pueden estar alejados.

Los segmentos adyacentes de una lámina β pueden tener una orientación paralela o antiparalela. En la orientación paralela, la cadena polipeptídica interacciona con otra en la misma disposición y en las que la dirección N-C es la misma. En este caso la interacción de varias cadenas da lugar a lo que se llama *hoja plegada paralela*. En la orientación antiparalela, las cadenas interaccionan de manera que su polaridad es opuesta. En este caso la estructura que resulta es la *hoja plegada antiparalela*. En el caso de la orientación paralela el periodo de repetición es más corto (6,5 Å), comparado con la antiparalela (7 Å), y los patrones de formación de puentes de hidrógeno son diferentes. En algunas estructuras densamente empaquetadas la presencia de ciertos aminoácidos está limitada. Por ejemplo, los grupos R de los aminoácidos de las superficies de contacto deben ser relativamente pequeños. Ver ejemplo de lámina β en la Figura 1.12B.

Otro elemento de estructura secundaria frecuente son los *giros* β. Estos son elementos de conexión que unen segmentos sucesivos de hélices α o láminas β. A menudo en los giros β son frecuentes los aminoácidos Gly y Pro. Esto es debido a que la Gly es pequeña y flexible, y la Pro adopta una configuración *cis* poco frecuente. Existen dos tipos de giros β, tipo I y tipo II.

Motivos estructurales

Los *motivos estructurales* o también llamados *estructuras suprasecundarias* o *plegamientos* constituyen un nivel superior a la estructura secundaria y se definen como la disposición estable y la conectividad entre distintos elementos de estructura secundaria. En una proteína las hélices α y láminas β pueden estar conectadas entre sí y combinadas de diferentes maneras, aunque en la naturaleza estas posibilidades son limitadas. Los motivos conocidos van de simple a complejo y aparecen a menudo como elementos repetidos o combinaciones. En general podemos distinguir cuatro tipos de estructuras suprasecundarias: i) todo hélices α (ej. cuatro hélices α empaquetadas); ii) hélice α/lámina β donde los segmentos están alternados o entremezclados (ej. barril α/β); iii) hélice α + lámina β donde los segmentos están algo segregados; iv) todo láminas β (ej. barril β).

El análisis del plegamiento de una proteína nos puede revelar relaciones evolutivas que son difíciles de detectar al nivel de la secuencia polipeptídica (**??**). Esto puede ayudar a entender mejor la función, la actividad biológica o el papel en un determinado organismo. A la unidad funcional y evolutiva se le denomina *dominio*, y en general se puede decir que a cada dominio le corresponde una función molecular. Existen varias bases de datos de referencia para definir dominios de proteínas. Por ejemplo, la base de datos Structural Classification Of Proteins[5] (SCOP) organiza las proteínas con estructura conocida y depositadas en la base de datos Protein Data Bank (ver Subsección 1.7.1) en base a criterios estructurales y evolutivos. Otra base de datos es Pfam[6], que define familias de secuencias o dominios en base a alineamientos múltiples para facilitar su localización en otras proteínas. La base de datos SUPERFAMILY[7] contiene la anotación estructural y funcional de todas las proteínas y genomas. SUPERFAMILY asigna dominios en base a las definiciones de SCOP. Para más detalle consultar la **??**.

1.5.3. Estructura terciaria

La *estructura terciaria* de una proteína es la disposición espacial de todos sus átomos y está determinada por la secuencia de aminoácidos que la compone. Las interacciones de las cadenas laterales

[5]Structural Classification Of Proteins. `http://scop2.mrc-lmb.cam.ac.uk`
[6]Pfam Database. `http://pfam.xfam.org`
[7]SUPERFAMILY Database. `http://supfam.org/SUPERFAMILY`

de los residuos de la proteína guían al polipéptido para formar una estructura compacta. Existen cuatro tipos de *interacciones que cooperan para la estabilización de la estructura terciaria de las proteínas*: i) *puentes disulfuro*; ii) *interacciones hidrofóbicas* (*fuerzas de van der Waals*); iii) *puentes de hidrógeno*; iv) *interacciones iónicas*.

Los aminoácidos con cadenas laterales no polares tienden a localizarse en el interior de la proteína, en donde se asocian con otros aminoácidos con cadenas laterales no polares para alcanzar la máxima estabilidad posible. En general los aminoácidos polares tienden a encontrarse en la superficie de las proteínas. Esta organización en general está invertida en algunas proteínas de membrana que forman poros o canales, en donde los aminoácidos con cadenas laterales no polares están en contacto con los lípidos componentes de la bicapa lipídica de la membrana y los aminoácidos polares están en el centro de la molécula formando el poro hidrofílico o canal. Dentro de estas interacciones se encuentran las fuerzas de van der Waals. Los aminoácidos con cadenas laterales que contienen átomos de hidrógeno unidos a átomos de oxígeno o nitrógeno, como los grupos alcohol de serina y treonina, pueden formar puentes de hidrógeno con átomos ricos en electrones, como el oxígeno del grupo carboxilo o el oxígeno del grupo carbonilo del enlace peptídico. La formación de los puentes de hidrógeno entre los grupos polares en la superficie de la proteína y el solvente acuoso que la contiene, incrementa su estabilidad. Los aminoácidos con cadenas laterales que contienen átomos de hidrógeno unidos a átomos de oxígeno o nitrógeno, como los grupos alcohol de la serina y treonina, pueden formar puentes de hidrógeno con átomos ricos en electrones, como el oxígeno del grupo carboxilo o bien el oxígeno del grupo carbonilo del enlace peptídico. La formación de los puentes de hidrógeno entre los grupos polares en la superficie de la proteína y el solvente acuoso que la contiene, aumenta la estabilidad. Por otro lado, los grupos cargados negativamente que se encuentran en las cadenas laterales de algunos aminoácidos como el grupo carboxilo en la cadena lateral del aspartato o glutamato, pueden interaccionar con cadenas laterales cargadas positivamente como el grupo ϵ amino de la lisina.

1.5.4. Estructura cuaternaria

Las proteínas también pueden contener múltiples subunidades polipeptídicas. Las *estructuras cuaternarias* de las proteínas comprenden desde dímeros sencillos hasta grandes complejos. Una proteína con varias subunidades se conoce como *multímero* o *proteína multimérica*. Cuando un multímero tiene sólo unas pocas subunidades se denomina *oligómero*. Si un multímero está formado por varias subunidades diferentes, la estructura global puede ser asimétrica y bastante compleja. En otros casos los multímeros pueden tener subunidades idénticas o grupos repetidos de subunidades no idénticas dispuestos simétricamente. Muchas proteínas multiméricas tienen funciones reguladoras. En otros casos, subunidades diferentes pueden realizar funciones separadas aunque relacionadas, tales como la catálisis y la regulación.

1.6. Métodos empíricos para el estudio de macromoléculas

Existen diversos métodos experimentales para estudiar la estructura de las biomoléculas a nivel atómico. Fundamentalmente se utilizan las técnicas de difracción de rayos X, resonancia magnética nuclear y crio-microscopía electrónica de transmisión. La *difracción de rayos X* precisa de la preparación de cristales de proteínas. A pesar de que la obtención de cristales de proteína suele ser un

proceso laborioso, ésta es la técnica preferentemente utilizada para resolver estructuras de proteínas, sobre todo las de gran tamaño. La *resonancia magnética nuclear* (RMN) permite la resolución de proteínas en disolución y de pequeño tamaño, ya que las medidas experimentales son difíciles o imposibles para moléculas grandes. La *crio-microscopía electrónica*, como cualquier método óptico, no tiene suficiente resolución para resolver a nivel atómico una estructura, pero puede proporcionarnos una buena imagen de la disposición subcelular de las moléculas. La *microscopía de electrones* es eficaz para estructuras muy grandes, como supercomplejos de proteínas.

1.6.1. Cristalización

Para iniciar un proyecto de determinación estructural mediante cristalografía y difracción de rayos X primero necesitamos *cristalizar la proteína*. Pero para ello antes debemos purificar la proteína en una relativa gran cantidad (miligramos) y con una *alta pureza* y *homogeneidad*. La presencia de agregados, diferentes oligómeros o cierta inestabilidad en la proteína pueden dificultar su cristalización. Estos detalles son críticos para que la cristalización sea un éxito. Para evaluar estos posibles problemas podemos realizar medidas de *Dynamic Light Scattering* (DLS) para detectar posibles formas oligoméricas o agregados en la muestra a cristalizar o bien medidas de *Differential Scanning Fluorometry* (DSF) que permiten caracterizar la estabilidad de la muestra en diferentes tampones y la presencia de diferentes ligandos o cofactores. Otros métodos espectroscópicos como el *dicroísmo circular* (CD) también pueden dar información a este respecto. Por otro lado, métodos de purificación adicionales mediante *cromatografía en columna* o *ultracentrifugación* pueden ser útiles para una mejor caracterización de la proteína si fuera necesario. Es importante tener en cuenta que la preparación de la muestra para un experimento de cristalografía es mucho más crítica que para un experimento de caracterización bioquímica. En este segundo caso generalmente se exige como criterio de calidad una pureza parcial y que la proteína mantenga su actividad.

La cristalización de una proteína está controlada por las leyes de la termodinámica, por lo tanto es un proceso semejante al de cualquier sal o mineral. En ambos casos se necesita llevar la solución a un estado de *sobresaturación* después del cual la proteína (o la sal) comenzará a cristalizar. Sin embargo, los métodos empleados en uno y otro caso son diferentes. Por ejemplo, en el caso de una sal como $CaSO_4$ podríamos conseguir una solución sobresaturada calentando a 40-45 °C y luego después de dejarla enfriar a temperatura ambiente durante un tiempo precipitarían los cristales en el fondo del recipiente. En el caso de las proteínas el calentamiento no es un método adecuado porque las proteínas rápidamente se desnaturalizan a altas temperaturas (con la excepción de proteínas procedentes de organismos extremófilos). La solubilidad de una proteína no sólo depende de la temperatura, además depende de la concentración, tipo de sal presente en el tampón, pH del tampón, presencia de posibles cofactores. El número de técnicas descritas para cristalizar una proteína es elevado pero sólo unas pocas se usan habitualmente. La más común se basa en alcanzar la sobresaturación mediante *difusión de vapor*. Esto se puede conseguir bien mediante la modalidad de "*gota colgante*", o mediante "*gota posada*". Otra estrategia menos usada es la de *diálisis*.

1.6.2. Difracción de rayos X

La *difracción de rayos X* es una de las técnicas más potentes para la determinación estructural de proteínas, pues proporciona una fotografía tridimensional a escala atómica del material cristalizado,

que incluye su plegamiento, distancias y ángulos de enlace, empaquetamiento cristalino, etc. En definitiva proporciona datos para entender la función biológica de las proteínas.

Un requisito indispensable para abordar un estudio estructural mediante esta técnica es el de *disponer de cristales únicos, homogéneos y de buena calidad*. Cuando un haz de rayos X incide sobre un cristal interacciona con los electrones de los átomos que lo componen, haciéndoles vibrar acopladamente con las variaciones periódicas de su campo eléctrico. De esta manera, los electrones al vibrar se convierten en focos de una nueva radiación X que se emite de forma esférica. Este fenómeno se denomina *dispersión*. La desviación de los rayos X dispersados es mayor cuanto mayor es la densidad electrónica de la materia sobre la que inciden. Los rayos X dispersados por los electrones pueden interferir entre sí dando lugar a cancelaciones, o bien pueden combinarse y reforzarse en determinadas direcciones si están en fase, dando lugar a una dispersión cooperativa conocida como *difracción*. Para que se produzcan todas las difracciones posibles los cristales deben girarse de tal modo que todos los planos virtuales se coloquen en una disposición adecuada para cumplir la *ley de Bragg*. Sólo los haces difractados pueden ser detectados en la película fotográfica o en el detector. Durante la difracción de rayos X el haz incidente produce calor y radicales libres, por lo que es importante enfriar el cristal durante el proceso de difracción, con el fin de evitar daños irreversibles en el cristal y por tanto en la proteína.

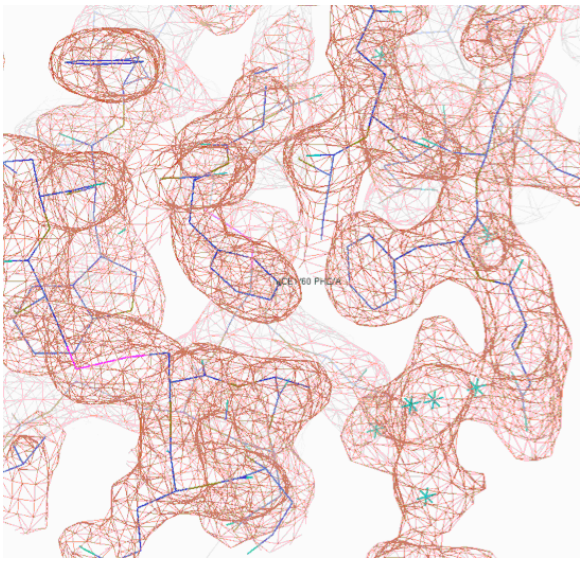

Figura 1.13: Zona del mapa de densidad electrónica de un cristal de proteína[8].

Los rayos X difractados por el cristal suministran una imagen en la que existe una disposición regular de manchas punteadas. El *patrón global de reflexiones y sus intensidades permite la construcción de un mapa de densidad electrónica de la proteína*, que tiene un aspecto similar al de un mapa topográfico. A partir del mapa de densidad electrónica se puede representar el trazado de las uniones entre átomos mediante distintos métodos. Sin embargo, para poder calcular la función de densidad electrónica, y por lo tanto poder saber la localización de los átomos en el interior de la celdilla, necesitamos conocer también el desfase entre las ondas, pero esta información se pierde durante el proceso de medida experimental, ya que no existen técnicas experimentales para medir esos desfases. Para resolver el "*problema de las fases*" Arthur Lindo Patterson (1934) propuso una

[8]Crystallographic X-ray Facility at the Department of Biochemistry, University of Cambridge. `http://www.xray.bioc.cam.ac.uk`

fórmula, basada en métodos de Transformada de Fourier, conocida como la *función de Patterson*, que permite una simplificación de la información contenida en la función de densidad electrónica, pues suprime la información de las fases. La información que proporciona la función de Patterson es un mapa de vectores de posición entre átomos (posiciones relativas). Posteriormente, a lo largo del siglo XX varios autores, entre ellos Jerome Karle y Herbert Hauptmann (Premio Nobel de Química, 1985), propusieron diferentes métodos, métodos directos, para resolver el problema en cristales formados por moléculas de tamaños pequeño y medio. En cristales que contienen moléculas grandes, proteínas o enzimas, el problema de la fase puede resolverse mediante tres métodos, dependiendo del caso: i) *método de Reemplazo Isomorfo Múltiple* (MIR, del inglés, *Multiple Isomorphous Replacement*) basado en el método de Patterson; ii) *método de Difracción Anómala Múltiple* (MAD, del inglés *Multi-wavelength Anomalous Diffraction*); iii) *método de Reemplazo Molecular* (MR, del inglés *Molecular Replacement*), haciendo uso de un modelo estructural de una proteína homóloga, previamente determinada. Por último, el modelo estructural obtenido debe ser validado, es decir, debe ser consistente con los criterios químicos, no presentar impedimentos estéricos, coherencia en las longitudes y ángulos de enlaces, etc. Ver ejemplo de mapa de densidad electrónica en la Figura 1.13.

La cristalografía-difracción de rayos X es un excelente método para determinar las estructuras de las proteínas rígidas que forman cristales ordenados. Las proteínas flexibles o con regiones desordenadas o dúctiles son difíciles o imposibles de estudiar por este método, pues la cristalografía requiere tener muchas moléculas alineadas exactamente en la misma orientación. Las regiones desordenadas o dúctiles son invisibles en los mapas de densidad electrónica pues presentan diversas orientaciones y su densidad electrónica se expande por un espacio grande. La exactitud de la estructura atómica determinada depende de la calidad de los cristales. En cristales perfectos, tenemos mucha más confianza que la estructura atómica resuelta refleje correctamente la estructura de la proteína. Dos medidas importantes de la exactitud de una estructura son su *resolución*, que mide la cantidad de detalle que puede ser visto en los datos experimentales, y el *valor de R*, que mide cómo el modelo atómico se ajusta a los datos experimentales.

1.6.3. Resonancia magnética nuclear

Al igual que la difracción de rayos X, la *resonancia magnética nuclear* (RMN) puede determinar las posiciones de cada uno de los átomos que constituyen una molécula de proteína. La RMN requiere que la proteína se encuentre en disolución, normalmente acuosa. Además, la RMN aporta no solo información estructural, sino también dinámica, porque permite observar ligeros cambios de conformación. Es una técnica muy adecuada para el *estudio de proteínas con regiones flexibles* o para *proteínas intrínsicamente desordenadas o dúctiles* donde la difracción de rayos X tiene una gran limitación. Pero no todo es posible con RMN ya que tiene una limitación según el tamaño de la proteína. En la actualidad *el tamaño máximo de las proteínas analizables por RMN está en torno a los 60-80 kDa*, aunque este límite va aumentando a medida que se producen avances en la técnica. Como en el caso de la difracción de rayos X, hay que preparar la muestra en unas condiciones óptimas para un su estudio. Es decir, hay que tener en cuenta la concentración de proteína, pureza, características del disolvente, el pH, la temperatura, etc. Las condiciones deben ser las que mejor conserven la estructura nativa de la proteína y que a su vez permitan la adquisición de espectros de RMN.

La técnica de RNM permite detectar la energía que absorben los núcleos de los átomos que entran

en resonancia al ser irradiados con radiofrecuencias en el seno de un campo magnético. Cada tipo de núcleo resonante que se puede encontrar en una proteína (1H, 13C, 15N, 31P) absorbe energía en una región de radiofrecuencias específica. Sin embargo, no todos los átomos del mismo tipo (por ejemplo, 1H) que hay en una molécula absorben energía exactamente a la misma frecuencia de radiación; existen pequeñas variaciones que dependen de la densidad electrónica del entorno químico que lo rodea. Por tanto, cada núcleo 1H contenido en una proteína resuena a un valor intrínseco de radiofrecuencia, según el entorno en que se encuentra. Teniendo en cuenta este principio se ha desarrollado la técnica de *RMN bidimensional*, idónea para resolver la estructura tridimensional de proteínas. Esta técnica permite identificar núcleos 1H muy próximos (a menos de 5 Å), debido a un fenómeno denominado *efecto Overhauser nuclear* (NOE). El NOE consiste en que el acoplamiento de los campos magnéticos entre núcleos próximos conlleva a una variación de la radiofrecuencia con las que resuenan. La RMN bidimensional proporciona lo que se denomina un espectro bidimensional, gráficamente una nube de puntos entre dos ejes. Los puntos que están fuera de la diagonal permiten identificar pares de protones separados por menos de 5 Å. La interpretación de un espectro bidimensional RMN en una estructura tridimensional es un proceso complicado que precisa de equipos informáticos con adecuados programas gráficos. El programa de ordenador precisa de información sobre los parámetros químicos, tales como la estructura primaria de la proteína, longitudes y ángulos de enlaces, radios de van der Waals, etc. Entonces, el ordenador genera una familia de estructuras que representan una gama de conformaciones consistentes con las proximidades atómicas detectadas por NOE.

1.7. Herramientas bioinformáticas básicas para el estudio de macromoléculas

1.7.1. Protein Data Bank

El repositorio "*Protein Data Bank*" (PDB)[9] contiene los ficheros de las coordenadas atómicas e información relevante adicional que describe a las proteínas cuya estructura 3D se ha determinado por cristalografía de rayos X, espectroscopia RMN o crio-microscopía electrónica de transmisión. Estos ficheros contienen un listado de los átomos de cada proteína y su localización en el espacio. El formato típico de un *fichero PDB* incluye un texto al inicio, que resume la descripción de la proteína y los detalles de la resolución de su estructura, seguido por la secuencia y una larga lista de los *átomos* y sus *coordenadas*. El archivo también contiene información sobre las condiciones experimentales que han sido usadas para determinar estas coordenadas atómicas.

1.7.2. Visualización de estructuras en 3D

Las estructuras 3D de biomoléculas pueden ser visualizadas con diferentes programas informáticos disponibles, tanto libres como de licencia comercial. Entre ellos podemos destacar PyMOL[10], Vi-

[9]Protein Data Bank. `http://www.rcsb.org`
[10]PyMOL. `http://www.pymol.org`

sual Molecular Dynamics[11] (VMD), UCSF Chimera[12], DeepView[13] (old Swiss-PdbViewer), Jmol[14], RasMol[15] o Protein Explorer[16] entre otros. Estos programas permiten visualizar una estructura molecular si conocemos sus coordenadas atómicas (generalmente en un archivo de formato texto). Existen dos tipos de programas de visualización molecular por ordenador, los que funcionan independientemente de las páginas web y los que funcionan como *plugins* adaptados a las páginas web y que siempre requieren páginas web para funcionar. A continuación comentaremos las aplicaciones de dos programas: PyMol y Swiss-Pdb Viewer, una lista más exhaustiva puede consultarse en Wikipedia[17].

PyMol

PyMOL[10] se puede instalar en nuestro ordenador con licencia o en su versión académica libre. Puede trabajar con una amplia variedad de formatos diferentes desde ficheros PDB a ficheros multi-SDF o mapas de densidad electrónica. PyMOL es fácil de usar y permite representar las estructuras en una gran variedad de formas (ej. esferas, superficie, vistas volumétricas, líneas, enlaces poniendo énfasis en la conectividad, elementos de estructura secundaria, entre otros). También permite la superposición de estructuras. Ejemplo en Figura 1.14.

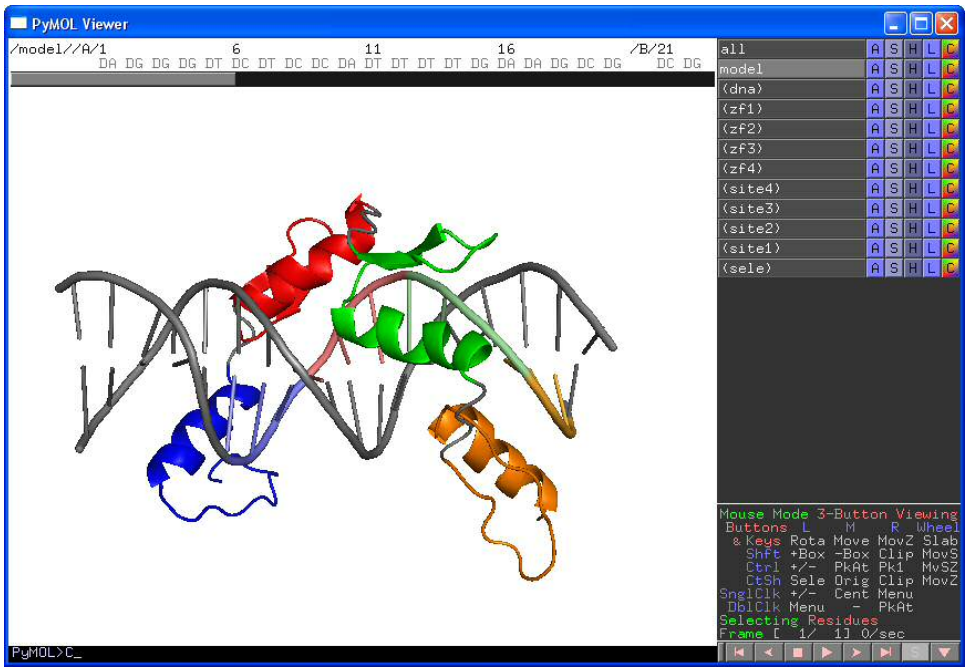

Figura 1.14: Ejemplo de visualización de un complejo proteína-DNA con PyMOL.

[11] Visual Molecular Dynamics. `http://www.ks.uiuc.edu/Research/vmd`
[12] UCSF Chimera. `http://www.cgl.ucsf.edu/chimera`
[13] DeepView. `http://spdbv.vital-it.ch`
[14] Jmol. `http://jmol.sourceforge.net`
[15] RasMol. `http://rasmol.org/`
[16] Protein Explorer. `http://proteinexplorer.org`
[17] Software for protein structure visualization. `http://en.wikipedia.org/wiki/Software_for_protein_structure_visualization`

DeepView

DeepView[13] se puede instalar en el ordenador con licencia y es una aplicación que proporciona una interfaz de fácil manejo para visualizar y analizar proteínas al mismo tiempo. La proteínas se pueden superponer con el fin de deducir alineamientos estructurales y comparar sus sitios activos u otras regiones relevantes. Mutaciones de aminácidos, puentes de hidrógeno, ángulos y distancias entre átomos son fáciles de determinar mediante su interfaz gráfica. DeepView está conectado con SWISS-MODEL[18], un servidor de modelado estructural automático. El uso combinado de estos dos programas facilita el trabajo para generar modelos estructurales. DeepView también lee mapas de densidad electrónica y proporciona varias herramientas para construir mapas de densidad. Además integra varias herramientas de modelado y mutación de aminoácidos.

[18]SWISS-MODEL. `http://swissmodel.expasy.org`

1.8. Bibliografía

[1] S. Al Karadagh. Introduction to protein structure and structural bioinformatics, 2014.

[2] G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in neurospora. *Proc Natl Acad Sci U S A*, 27(11):499–506, 1941.

[3] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Science, 2 edition edition, 1999.

[4] C. G.-M. Calera, Gomez-Moreno, and J. S. Sanz. *Estructura de Proteinas*. Ariel, Editorial S.A., 2003.

[5] I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.

[6] F. H. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192:1227–32, 1961.

[7] M. Grunberg-Manago, P. J. Oritz, and S. Ochoa. Enzymatic synthesis of nucleic acidlike polynucleotides. *Science*, 122(3176):907–10, 1955.

[8] H. G. Khorana. Polynucleotide synthesis and the genetic code. *Fed Proc*, 24(6):1473–87, 1965.

[9] P. Leder and M. W. Nirenberg. Rna codewords and protein synthesis, 3. on the nucleotide sequence of a cysteine and a leucine rna codeword. *Proc Natl Acad Sci U S A*, 52:1521–9, 1964.

[10] J. H. Matthaei, O. W. Jones, R. G. Martin, and M. W. Nirenberg. Characteristics and composition of rna coding units. *Proc Natl Acad Sci U S A*, 48:666–77, 1962.

[11] A. McPherson. *Advances in Protein Chemistry and Structural Biology*. Academic Press, 2010.

[12] F. Sanger and A. R. Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *J Mol Biol*, 94(3):441–8, 1975.

[13] F. Sanger and H. Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. i. the identification of lower peptides from partial hydrolysates. *Biochem J*, 49(4):463–81, 1951.

[14] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953.