



# UNIVERSITÀ DI PISA

Corso di Laurea Triennale in Informatica (L-31)

## TESI DI LAUREA

# Protein Folding: dai metodi classici per la predizione della struttura di proteine alla rivoluzione di AlphaFold

### Relatore

Prof. Paolo Milazzo

### Correlatore

Prof. Mario Pirchio

### Candidato

Ludovico Venturi

ANNO ACCADEMICO 2020/2021

# Indice

<b>1</b>	<b>Background</b>	<b>3</b>
1.1	Background biologico . . . . .	4
1.1.1	Organizzazione della vita: dagli atomi alle cellule . . . . .	4
1.1.2	Concetti fondamentali in biologia . . . . .	7
1.1.3	Dogma centrale della biologia . . . . .	8
1.1.4	Dai geni alle proteine . . . . .	12
1.1.5	Proteine: le macromolecole più importanti della vita . . . . .	14
1.2	Background informatico . . . . .	18
1.2.1	Bioinformatica . . . . .	18
1.2.2	Soft computing . . . . .	19
1.2.3	Intelligenza Artificiale . . . . .	20
1.2.4	Machine Learning . . . . .	21
1.2.5	Reti neurali artificiali (ANN) . . . . .	22
<b>2</b>	<b>Protein Folding</b>	<b>24</b>
2.1	Postulato di Anfinsen . . . . .	25
2.1.1	Esperimento di Anfinsen . . . . .	26
2.1.2	Denaturazione . . . . .	27
2.2	Struttura delle proteine . . . . .	28
2.2.1	Legami e interazioni molecolari . . . . .	29
2.2.2	Livelli strutturali . . . . .	31
2.2.3	Evoluzione e classificazione . . . . .	38
2.3	Dinamica del ripiegamento . . . . .	40
2.3.1	Geometria ed energetica del ripiegamento . . . . .	40
2.3.2	Ripiegamento assistito . . . . .	44
2.3.3	Misfolding, prioni e malattie . . . . .	45
2.3.4	Controllo qualità e apoptosis . . . . .	47
2.4	Sfide al dogma di Anfinsen: IDP e fold switching . . . . .	49
2.4.1	Considerazioni epistemologiche . . . . .	51

2.5 Il problema del Protein Folding . . . . .	54
<b>Bibliografia</b>	<b>56</b>

# Capitolo 1

## Background

*Cos'è la vita? Da dove viene?* - Fino al XVIII secolo per rispondere a tale quesito si faceva riferimento alla fede nel vitalismo: l'esistenza di una forza vitale non subordinata a leggi della chimica e della fisica. Importanti svolte furono gli esperimenti, prima di Redi poi di Spallanzani, per dimostrare l'infondatezza della teoria della *generazione spontanea*, secondo la quale la vita poteva generarsi da materia non vivente. Un'importante passo in avanti, in concomitanza con l'affermarsi della *teoria cellulare*, fu il lavoro di Pasteur che stabilì un collegamento fra processi vitali e reazioni chimiche: per la conversione di zucchero in alcool (fermentazione) era necessaria la presenza di microorganismi.

Successivamente vi sono i lavori di Berthelot e Buchner (premio Nobel per la Chimica 1907), il quale dimostrò che era possibile ottenere la fermentazione in assenza di microorganismi, usando solamente sostanze estratte da essi. Queste sostanze furono chiamate *enzimi* (dal ted. Enzym, letteralmente «dentro il lievito»<sup>[1]</sup>). Non si conosceva la loro natura chimica, si scoprì successivamente che tutti gli enzimi sono *proteine* (dal greco «primario», «che occupa la prima posizione»<sup>[2]</sup>). Queste proteine agivano da catalizzatori: acceleravano le reazioni chimiche all'interno delle cellule senza cambiare la loro natura, quindi senza consumarsi, e senza entrare nei prodotti finali della reazione.

La scoperta degli enzimi portò ad un cambio di paradigma nel pensiero scientifico riguardo le origini della vita: veniva ora considerata come la conseguenza di numerosi processi chimici resi possibili dalle proteine<sup>[3]</sup>. I fondamenti del pensiero biologico si spostarono dal vitalismo al meccanicismo secondo il quale tutti i fenomeni naturali, vita compresa, sono governati dalle stesse leggi, sia per sostanze organiche che inorganiche.

L'inconorazione delle proteine a *macromolecole più importanti della vita* si può legare ad un'altra svolta nel pensiero scientifico avvenuta nella seconda metà del XX secolo: la rivoluzione genetica. Le proteine sono ben più che "macchine molecolari": sono i prodotti primari dei geni e sono coinvolte nell'espressione dell'informazione genetica. È sullo sfondo di questa rivoluzione che l'informatica si è inserita all'interno del mondo della biologia.

## 1.1 Background biologico

### 1.1.1 Organizzazione della vita: dagli atomi alle cellule

Nonostante le grandi differenze in dimensione, dieta, riproduzione, morfologia, comportamento, vi è un tratto comune a tutti gli organismi viventi: sono composti di cellule. Tutte le cellule sono caratterizzate da una stupefacente somiglianza chimica poiché utilizzano molecole simili e hanno ereditato tutte le stesse *intuizioni*<sup>1</sup> genetiche. Si pensa quindi vi sia un antenato comune a tutti i viventi: una cellula vissuta circa 3,5 miliardi di anni fa che conteneva un prototipo del macchinario universale della vita sulla Terra oggi<sup>[4]</sup>.

Prima di parlare di cellule è opportuno richiamare l'attenzione sulle strutture biologiche. L'organizzazione biologica si basa su una gerarchia di livelli strutturali<sup>2</sup>, ognuno dei quali poggia su un gradino sottostante:



Tutta la materia è costituita da 94 elementi chimici in natura<sup>[6]</sup> (tralasciando quindi gli altri 24 elementi sintetici). La materia vivente è composta per il 96% da atomi di C, O, N, H (carbonio, ossigeno, azoto, idrogeno). Un atomo ha un nucleo composto da neutroni e protoni circondato da una nube di elettroni in rapido movimento. Il Dalton (Da) è l'unità della massa atomica, corrisponde al peso di un protone o neutrone:  $1\text{Da} = 1.66 \times 10^{-24}\text{g}$ . Un elettrone pesa  $0.0005\text{Da}$ . Gli elettroni più esterni sono chiamati *elettroni di valenza* e determinano il comportamento chimico di un atomo.

Lo scheletro delle molecole organiche è formato da catene carboniose, lunghe catene di atomi di carbonio legati fra loro da legami covalenti (il tipo di legame chimico più forte). Salendo di complessità si arriva alle macromolecole biologiche, fondamentali per le cellule: carboidrati, lipidi, acidi nucleici e proteine. I carboidrati sono combustibili cellulari e materiale da costruzione, i lipidi sono sia depositi di energia che i principali costituenti delle membrane cellulari, gli acidi nucleici permettono di codificare l'informazione genica e le proteine sono alla base delle funzioni vitali.

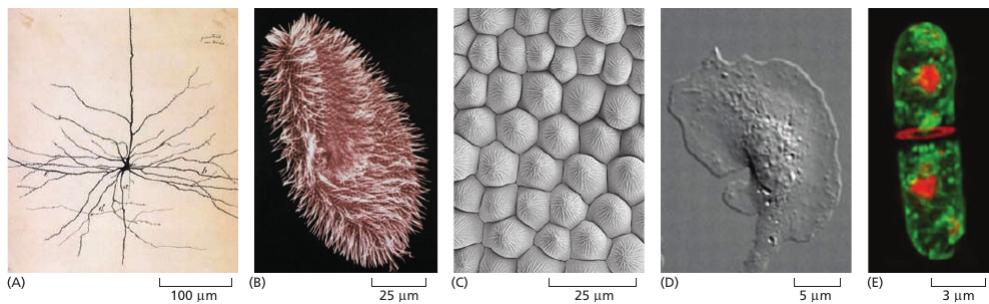
---

<sup>1</sup>Il termine *intuizione* è qui usato creativamente per indicare le soluzioni genetiche sviluppatesi e sopravvissute ad oggi. Non si intende attribuire intelligenza, pensiero o volontà all'evoluzione.

<sup>2</sup>Questa sezione di background biologico si basa in larga parte su N. A. Campbell, J. B. Reece, L. A. Urry et al., *Biologia E Genetica*. Pearson, 2012.

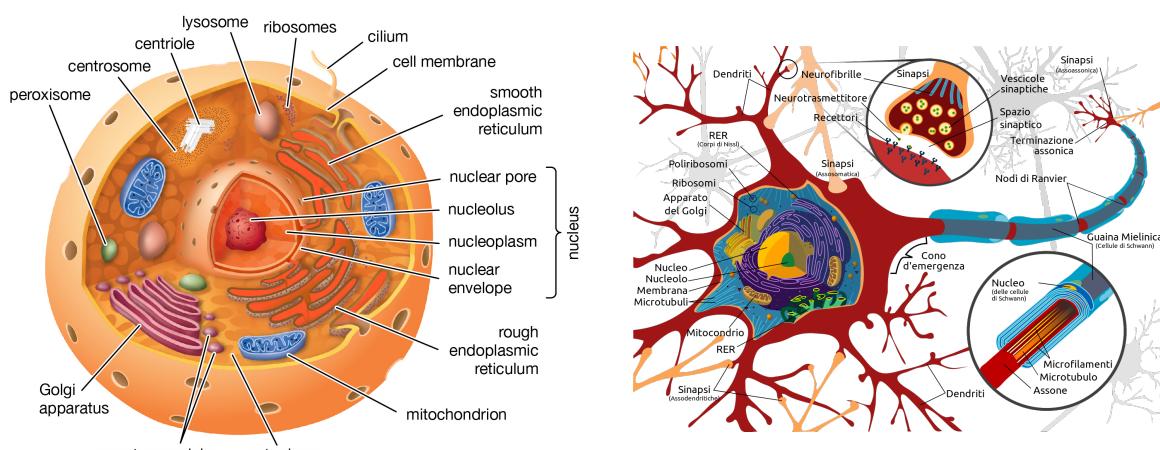
La cellula è la più piccola unità in grado di vivere. Per *vivente* si intende un essere dotato di: organizzazione interna, metabolismo, omeostasi, interazione con l'ambiente, adattamento, crescita e riproduzione.

Le cellule hanno dimensioni che variano dai *micrometri* ( $\mu\text{m}$ ) ai *centimetri* delle uova di rana, gallina o struzzo ai *metri* di neuroni con lunghi assoni. In figura 1.1 si possono notare le diverse dimensioni di alcune cellule.



*Figura 1.1:* (A) disegno di un neurone. (B) Paramecium. (C) superficie di un petalo di fiore di bocca di leone. (D) Macrofago. (E) Un lievito di fissione viene catturato nell'atto di divisione cellulare. Fonte: [4]

È possibile dividere gli esseri viventi in due domini<sup>3</sup>: *procarioti* ed *eucarioti*. Il primo include i due regni Bacteria e Archaea. Sono caratterizzati da cellule piccole, circa  $1\mu\text{m}$ . Il secondo dominio include cinque regni: animali, piante, funghi, protisti e cromisti. Gli organismi eucarioti dispongono di cellule più grandi (circa  $10\text{-}100\ \mu\text{m}$ ) dotate di compartimenti interni che separano le funzioni cellulari. La strutture tipiche di una cellula animale e di un neurone sono mostrate nelle seguenti figure:



*Figura 1.3: Neurone. Fonte [8]*

*Figura 1.2: Cellula animale. Fonte: [7]*

---

<sup>3</sup>Tale classificazione è soggetta a frequenti cambiamenti.

Una cellula eucariote animale è formata innanzitutto dalla membrana cellulare, un involucro costituito da un doppio strato fosfolipidico che permette alla cellula di avere il suo "spazio vitale" in quanto la separa dall'ambiente (spesso acquoso) circostante. È attraversata da piccoli pori che le permettono lo scambio di sostanze con l'esterno. Tutto ciò che si trova all'interno della cellula è immerso nel citoplasma, gel acquoso contenente grandi e piccole molecole. Il citosol è la parte del citoplasma non contenuta all'interno delle membrane intracellulari. Il volume totale delle cellule è composto da acqua per il 70% circa. Vi è poi il citoscheletro che dà forma strutturale e permette in alcuni casi movimenti direzionati.

Il primo organello di grande importanza è il reticolo endoplasmatico, formato da tubuli e cisterne e in comunicazione con l'involucro nucleare. È rugoso quando sono presenti ribosomi (sintetizzatori di proteine). È il componente della fabbrica cellulare che si occupa di attività e sintesi di molecole fondamentali per la sopravvivenza della cellula (sintesi di steroidi, metabolismo del glucosio, eliminazione di sostanze nocive). L'apparato del Golgi produce vescicole che si fondono poi con la membrana cellulare: è una centrale di smistamento per confezionare sostanze da esportare. I lisosomi sono il centro di degradazione e riciclo della cellula. Il mitocondrio è la centrale energetica della cellula, dove avviene la respirazione cellulare: utilizza ossigeno per bruciare molecole organiche degradate nel citoplasma come *piruvato* e *acetil-coenzima A* al fine di produrre energia che verrà immagazzinata sottoforma di ATP.

Infine è presente il nucleo, custode del DNA. È formato dall'involucro nucleare, cromatina e nucleolo. Il DNA nel nucleo è associato a delle proteine con cui forma un materiale fibroso chiamato cromatina, mostrandosi "sfilacciato" in modo da poter essere letto. Quando la cellula si riproduce la cromatina si condensa in strutture compatte e singole: i cromosomi. Il nucleolo non è provvisto di membrana e serve per la sintesi di RNA ribosomiale, cioè l'RNA che uscendo dai pori dell'involucro nucleare andrà nel citoplasma a formare i ribosomi. L'involucro nucleare possiede dei pori nucleari attraverso i quali possono transitare RNA e proteine, ma non DNA.

Il ciclo di vita delle cellule si basa su 4 fasi: crescita iniziale, sintesi del DNA, ulteriore crescita e mitosi (divisione cellulare). Le cellule dei mammiferi possono impiegare anche dei giorni per completare un ciclo di mitosi, mentre i lieviti solamente 90 minuti. Per questa ragione il lievito da fornaio (*Saccharomyces cerevisiae*) è molto utilizzato in citologia e genetica: è uno degli organismi eucarioti modello<sup>[4]</sup> ed il suo genoma è stato il primo ad essere sequenziato completamente tra gli eucarioti<sup>[9]</sup>.

Le cellule hanno una durata di vita molto variabile, ad esempio alcuni organismi unicellulari come le spore possono vivere anche decenni, così come i nostri neuroni, mentre i globuli bianchi non sopravvivono oltre pochi giorni.

Gli strumenti utilizzati per indagare nel mondo microscopico riescono a mostrare dettagli che vanno dal limite di  $200\text{nm}$  del microscopio ottico (limite imposto dalla natura ondulatoria della luce) alla precisione di  $1\text{nm}$  del microscopio a trasmissione elettronica (che usa fasci di elettroni invece di fasci di luce ma di contro necessita di campioni molto fini):

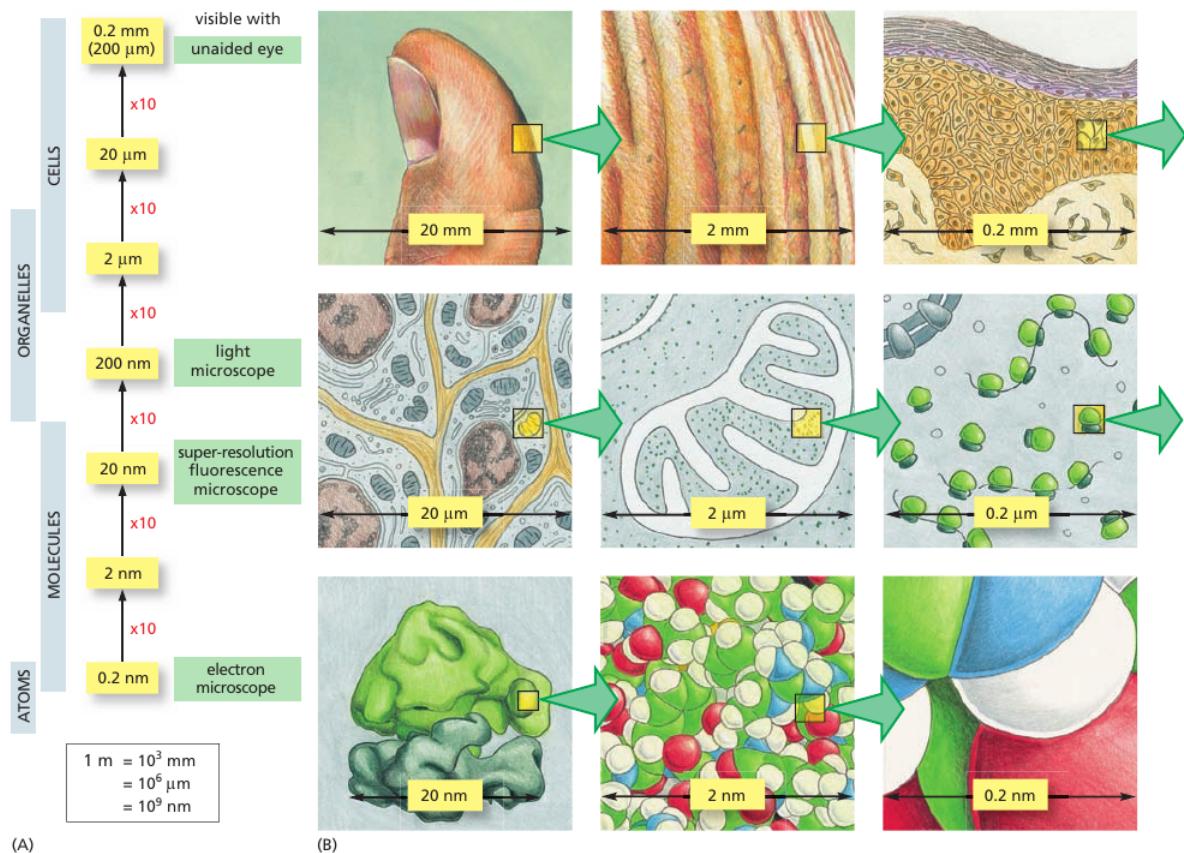


Figura 1.4: (A) Il grafico elenca le dimensioni dei livelli strutturali biologici, le unità di misura relative e gli strumenti necessari per visualizzarli. (B) Uno stesso dettaglio a varie scale di grandezza: pollice, pelle, cellule, mitocondrio, ribosomi, insieme di atomi che formano parte di una proteina. I dettagli molecolari sono oltre la potenza del microscopio elettronico. Fonte: [4]

### 1.1.2 Concetti fondamentali in biologia

- *Proprietà emergenti*

Ad ogni livello di indagine, ovvero passando da un livello della gerarchia strutturale al superiore, si palesano nuove proprietà non riconducibili ai livelli più semplici: le proprietà emergenti. Una singola molecola d'acqua non è né solida né liquida.

- *Teoria cellulare*

Le cellule rappresentano le unità strutturali e funzionali degli organismi.

- *Geni*

Il perpetuarsi della vita è possibile grazie alla trasmissione dei geni.

- *Forma e funzione*

Forma e funzione sono correlate a tutti i livelli biologici. Se le ali degli uccelli non fossero così come sono essi non potrebbero volare, se i mitocondri non avessero numero creste produrrebbero minori quantità di ATP, se i neuroni non avessero lunghi assoni non riuscirebbero a comunicare efficientemente e se i *paramecium* non avessero le loro ciglia non potrebbero muoversi come sommersibili (vedi figura 1.1B).

- *Evoluzione*

L'evoluzione rappresenta il tema centrale ed unificante della biologia, come si è già accennato sopra. Gli organismi sono sistemi aperti che interagiscono continuamente con l'ambiente, dotati di variabilità individuale e finalizzati alla competizione per la sopravvivenza.

- *Diversità e unità*

Vi sono da 5 a 30 milioni di specie differenti eppure scendendo sempre di più nella struttura degli organismi si osserva una similitudine quasi sconcertante. Un esempio che ci riguarda è la somiglianza fra le ciglia di *paramecium* e le ciglia di una cellula epiteliale delle vie aeree degli esseri umani: presentano la stessa sezione trasversale. Il codice genetico (le triplett) sono universali, gli amminoacidi si condificano nello stesso modo per tutti gli organismi. Diversità e unità della vita sulla Terra sono due facce della stessa medaglia. Il sequenziamento dei genomi e il loro confronto, basato su approcci informatici, ha rivelato una conservazione evoluzionistica, un'eredità comune: è possibile infatti scambiare geni omologhi codificanti proteine del ciclo di divisione cellulare fra uomini e lievito<sup>[4]</sup>: una cellula di lievito ha quindi tutto il macchinario molecolare necessario per leggere, interpretare e utilizzare il nostro codice genetico per la produzione di proteine umane funzionanti. Sono osservazioni simili che hanno guidato la direzione di alcune tecniche informatiche, anche per la predizione della struttura di proteine (come si vedrà successivamente).

### 1.1.3 Dogma centrale della biologia

Nel 1958 il premio Nobel Francis Crick introdusse il *dogma* centrale della biologia, che allo stato attuale si può considerare come l'insieme dei principali meccanismi alla base dell'espressione genica.

Il dogma descrive il flusso di informazione genetica: essa è conservata negli acidi nucleici DNA (RNA per alcuni virus) che possono essere duplicati, il DNA viene poi trascritto

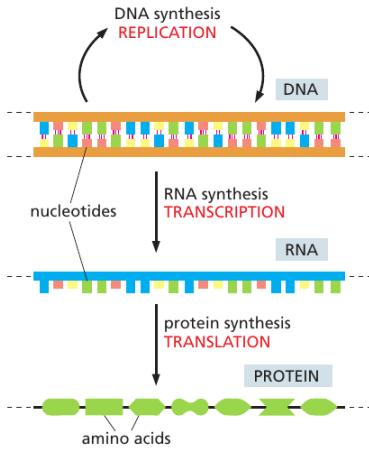


Figura 1.5: Dogma centrale in biologia. Fonte [4]

sottoforma di RNA e se codificante questo è poi tradotto in proteine, concepite come la forma operativa e terminale delle informazioni contenute nel genoma<sup>[10]</sup>.

Per avere una miglior panoramica del funzionamento di questo principio è importante approfondire la struttura del DNA (*acido desossiribonucleico*). Il DNA è una molecola composta da due catene complementari che si avvolgono l'una intorno all'altra tramite legami idrogeno formando una doppia elica. Le catene sono chiamate filamenti e sono antiparalleli. Dal punto di vista chimico è un polimero di nucleotidi, dove ogni nucleotide è composto da una base azotata, uno zuccherio pentoso (*ribosio* nell'RNA e *desossiribosio* nel DNA) e un gruppo fosfato (vedi figura 1.7). Per ogni giro dell'elica vi sono 10 coppie di basi. La struttura a doppia elica consente un'agevole meccanismo di replicazione del DNA, coadiuvato dagli enzimi DNA polimerasi, primasi e DNA ligasi. Gli accoppiamenti seguono delle regole precise: GC, AT/AU, da una parte deve esserci una pirimidina (C, T) e dall'altra una purina (A,G):

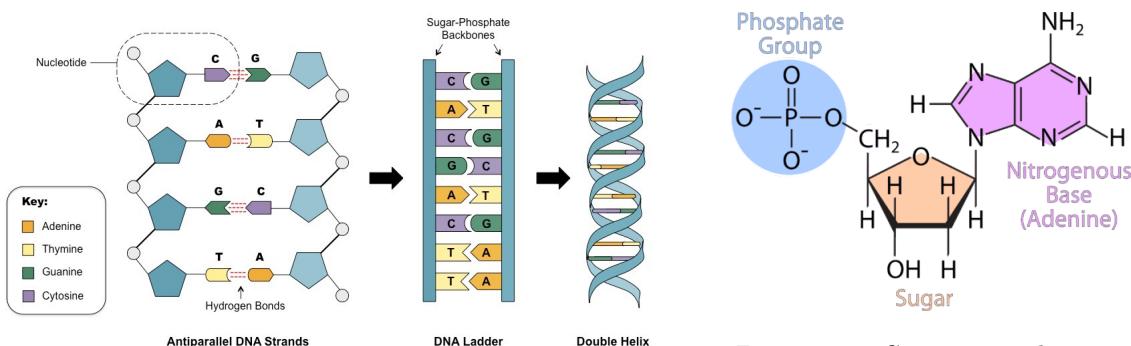
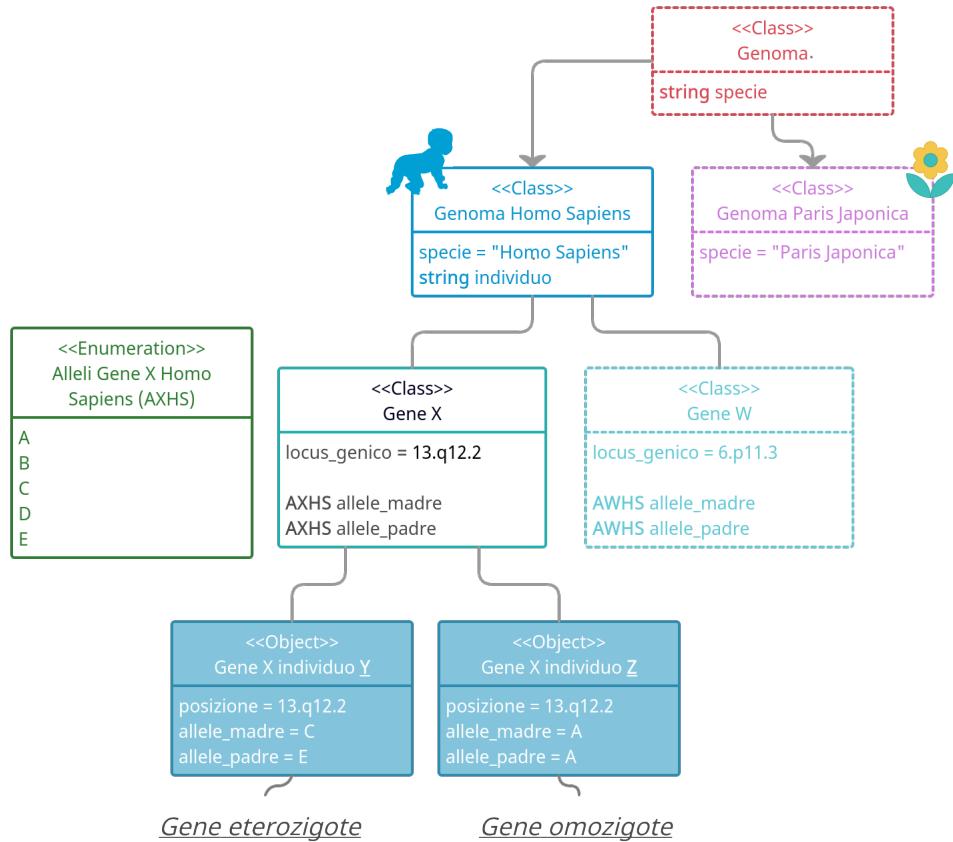


Figura 1.6: struttura del DNA. Fonte: [11]

Figura 1.7: Componenti di un nucleotide con Adenina per base azotata. Fonte [12]

Il *gene* è l'unità elementare dell'informazione genetica e corrisponde al segmento di

DNA (raramente di RNA) in grado di codificare la sequenza primaria di una proteina.



*Figura 1.8: Schema UML di genoma e alleli in stile programmazione a oggetti. AXHS nella classe Gene X indica il tipo enumerazione definito accanto (diminutivo di Alleli Gene X Homo Sapiens). AWHS nella classe gene W indica invece il tipo enumerazione che elenca i possibili alleli del gene W (non riportato). Ogni oggetto istanziato della classe Gene X avrà lo stesso locus genico e due possibili alleli. Creato su [createley.com](http://createley.com)*

Geni che controllano un medesimo carattere (per esempio, il colore dei capelli) sono disposti sui cromosomi in *loci* (plurale di locus genico, posizione) identici. I cromosomi omologhi sono cromosomi morfologicamente identici che presentano in loci corrispondenti gli stessi geni con le stesse informazioni. Ogni gene è presente in doppia coppia nelle cellule diploidi e può risultare pertanto omozigote od eterozigote. Il termine omozigote si riferisce a un gene in cui l'informazione riportata dall'allele materno è identica a quella paterna, mentre in geni eterozigoti il contributo dell'allele materno e paterno è diverso: in questo caso la determinazione fenotipica è legata ai concetti di dominanza e recessività genetica. Lo stesso gene nella stessa specie può esistere in varie forme, con leggere differenze nella sequenza nucleotidica: si sta parlando degli *alleli* del gene. Più precisamente l'insieme

delle possibili versioni (1, 2 o più) dello stesso gene corrisponde ai suoi alleli, un allele è quindi una delle versioni dello stesso gene nello stesso locus su un cromosoma.

Il *genoma* indica il patrimonio complessivo del DNA di una cellula (compreso il DNA di altri organelli come mitocondri o cloroplasti). L'insieme di tutti i geni di un individuo determinano il suo *genotipo* (quindi solo le regioni codificant); relativamente a un gene il genotipo può anche indicare il corredo di alleli che l'organismo si trova a possedere (nell'uomo al massimo 2). Il *fenotipo* indica invece l'insieme delle caratteristiche morfologiche e funzionali di un organismo, quali risultano dall'espressione del suo genotipo e dalle influenze ambientali. In un organismo, nonostante tutte le cellule condividano gli stessi geni, cellule diverse possono esprimere geni differenti (*espressione genica*).

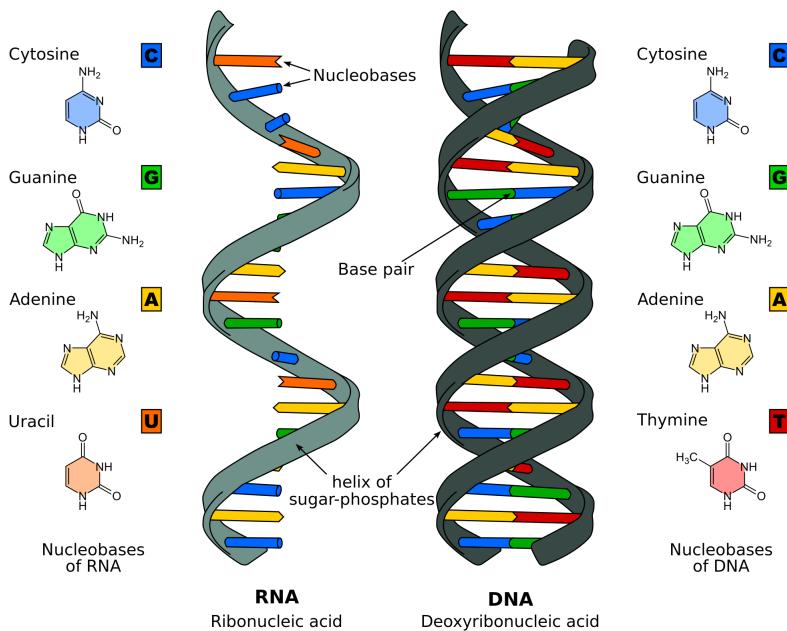


Figura 1.9: Differenze fra RNA e DNA Fonte: [13]

L'RNA (*acido ribonucleico*) esiste in varie forme. Le differenze con il DNA sono mostrate nella figura 1.9, si può notare che vi è un singolo filamento e che la base azotata timina è assente e al suo posto si trova la base uracile (U). Essendo ad un unico filamento può formare legami a idrogeno con sé stessa e assumere forme tridimensionali vantaggiose. Esistono vari tipi di RNA:

- mRNA, messaggero, contiene l'informazione per la sintesi delle proteine
- tRNA, di trasporto, necessario per la traduzione nei ribosomi
- rRNA, ribosomiale, entra nella struttura dei ribosomi
- snRNA, hnRNA

L'RNA catalitico o *ribozima*, enzima ad RNA, è una molecola di RNA in grado di catalizzare una reazione chimica similmente agli enzimi.

Il DNA dell'uomo contiene  $3^9$  coppie di nucleotidi (3.3Gb, *gigabasepairs*) ha circa 21000 geni codificanti e pesa 3.56pg<sup>4</sup>: se il genoma umano venisse esteso in lunghezza sarebbe lungo 2,2 metri dato che ogni nucleotide è lungo 0.34nm. Il batterio più semplice (*Nasutella lutea*) ha un genoma di 112Kb<sup>[14]</sup> (circa 76μm in lunghezza) contenente 137 geni codificanti mentre il genoma maggiore ad oggi riportato è quello della pianta *Paris Japonica* con 148.8Gb<sup>[15]</sup>, 50 volte quello dell'uomo (circa 100m in lunghezza), tanto per avere una visione quantitativa della diversità genetica tra gli organismi.

Species	T2 phage	<i>Escherichia coli</i>	<i>Drosophila melanogaster</i>	<i>Homo sapiens</i>	<i>Paris japonica</i>
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	Virus	Bacteria	Fruit fly	Human	Canopy Plant



Figura 1.10: Dimensioni del genoma di diverse specie a confronto. Fonte: [16]

Figura 1.11: Fiore di *Paris Japonica*. Fonte [17]

### 1.1.4 Dai geni alle proteine

Il codice genetico lavora a sequenze di codici di 3 lettere (es. "GAA" = Glutammato), questo perché si hanno a disposizione 4 lettere (le basi azotate) e si devono codificare i 20 diversi amminoacidi. Con 2 lettere avrei  $4^2$  possibilità che non sono sufficienti a descrivere 20 informazioni diverse, si utilizzano pertanto 3 lettere anche se ciò causa ridondanza nei codici. Un amminoacido è quindi codificato da una tripla: si parla di *codice a triplett*.

Il primo passo consiste nella *trascrizione*. Un filamento di DNA fa da stampo per la creazione di mRNA, il tutto esclusivamente tramite *complementarità di forma*. Il DNA non viene aperto come una zip ma l'apertura, la trascrizione (compiuta dall'RNA polimerasi, soggetta a errori anche frequenti) e la chiusura della doppia elica avvengono di pari passo. Vi è un terminatore nel DNA per indicare la fine del gene.

Le triplett nucleotidiche dell'mRNA sono dette *codoni* e codificano un amminoacido. I codoni devono essere letti in direzione 5' -> 3'. La molecola di mRNA lascia il nucleo attraverso i pori nucleari. È importante osservare che non tutti i geni codificano proteine (lo

<sup>4</sup>In termini di massa è possibile convertire il numero di paia di basi azotate in *picogrammi*, 1pg=0.978Gb, poiché una coppia di base azotate pesa 650Da. Il peso del genoma umano è calcolabile come segue:  $3.3 \times 10^9 \times 650 \times 1.66 \times 10^{-24} = 3.56 \times 10^{-12} g$

stadio di trascrizione potrebbe risultare quello finale) e che il codice genetico è *universale*, è condiviso dai batteri, piante, animali: per tutti la prolina si codifica in "CCG".

Negli eucarioti è presente un passaggio intermedio: la *maturazione*, o fase di processamento. È composto da due sottofasi:

- *incapsulamento*, viene aggiunta una coda e un cappuccio alle due estremità al fine di proteggere l'mRNA dalla degradazione e per segnalare l'inizio ai ribosomi.
- *splicing*, il DNA possiede lunghe sequenze nucleotidiche non codificanti, gli *introni*. In questa fase vengono rimossi e gli *esoni* (sequenze codificanti) vengono riunite insieme. È in questo modo che è possibile dare origini a sequenze primarie (delle proteine) diverse a partire da un unico gene.

L'ultimo passaggio è la *traduzione*, attraverso la quale la cellula interpreta il messaggio genetico e polimerizza gli amminoacidi per costruire la relativa proteina. Il processo di traduzione è la transizione da un linguaggio a 4 lettere (basi azotate) ad un linguaggio a 20 lettere (amminoacidi). La traduzione viene realizzata dal tRNA, una sorta di adattatore da linguaggio *genetico* a linguaggio *amminoacidico*. Il tRNA è un acido nucleico a forma di L composto da circa 80 basi, a un'estremità vi è l'anticodone (interfaccia con il linguaggio genetico) e all'altra vi è il sito di legame con un singolo amminoacido. Il tRNA trasporta ai ribosomi uno specifico amminoacido contenuto nel citoplasma. Esiste di conseguenza uno specifico tipo di tRNA per ogni codone.

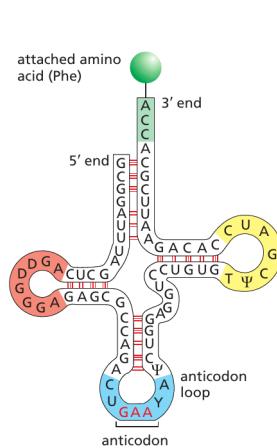


Figura 1.12: tRNA. Fonte [4]

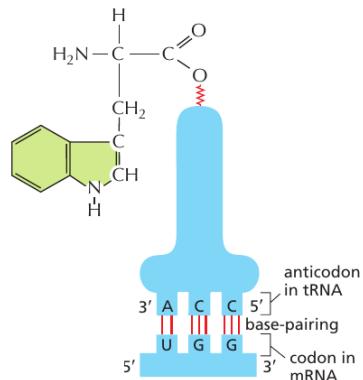


Figura 1.13: Traduzione: l'amminoacido triptofano (Trp) è codificato dal codone UGG nell'mRNA e si lega al tRNA tramite un legame energetico forte. Fonte: [4]

È interessante notare che il tRNA, proprio come le proteine, è caratterizzato dall'avere più strutture: quella primaria, costituita dalla sua sequenza nucleotidica, quella secondaria data dalla sua struttura a quadrifoglio e quella terziaria dovuta alla struttura tridimensionale.

nale a L. La differenza fra la struttura del tRNA e delle proteine sono gli elementi unitari: nel tRNA si tratta di nucleotidi mentre nelle proteine di amminoacidi.

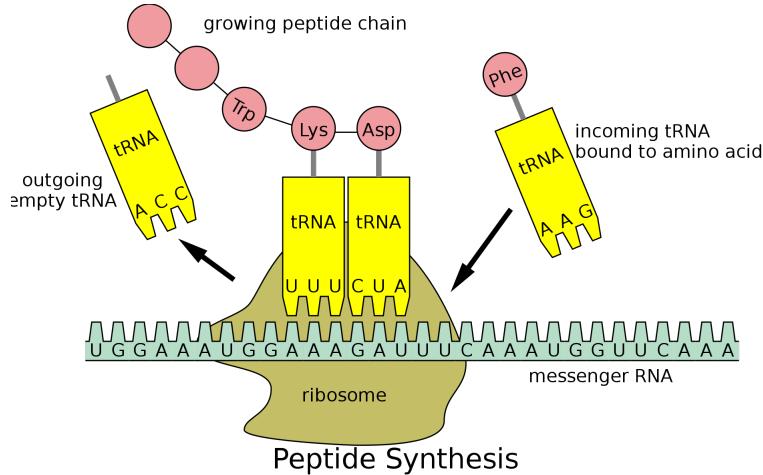


Figura 1.14: Traduzione, sintesi peptidica. Fonte: [18]

La traduzione comincia con il primo codone (AUG, che oltre a segnalare l'inizio codifica anche la metionina, vedi figura 1.15) al quale si incassa nel ribosoma un tRNA avente il corrispondente amminoacido legato. Si formano legami idrogeno fra i nucleotidi. Arriva un secondo tRNA combaciante con il successivo codone. I due amminoacidi si trovano vicini e formano un legame peptidico. L'mRNA scorre così che si crei posto per nuovi tRNA, nel frattempo gli amminoacidi si legano fra loro e cominciano a formare la proteina. Il ripiegamento della proteina comincia già durante la sua biosintesi. Il processo termina quando si arriva ad un codone di stop (es. UAA). Per velocizzare il processo di sintesi ribosomiale questo viene parallelizzato: tanti *poliribosomi* sono associati allo stesso mRNA attuando una rapida sintesi di copie multiple di un polipeptide a partire da un unico mRNA.

codons	AGA	UUA	AGC	GCA	GGG	AUA	UUG	AGU	UCA	CCC	UCC	ACG	GUA	UAA
	AGG	UUG	AGU	CGA	GGC	CUC	CUA	ACA	UCA	CCA	UCC	ACC	GUC	UAG
	GCA	CUA	UUA	GCA	GGG	AUC	CUC	UCC	UCC	CCC	UCC	ACG	UAC	UGA
	CGA	CUC	AAU	CAA	GGG	CUU	UUU	ACG	UCU	CCU	UCU	AGC	GUG	UAA
	GCC	CUU	AAU	UGC	GGG	AAG	AAA	UUC	UCC	CCG	UCC	UCC	UAC	UAG
	CGC	CUU	AAU	UGU	GGG	CAU	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UGA
	GCG	CUU	AAU	GAC	GGG	CAC	AAA	UUC	UCC	CCG	UCC	UCC	UAC	UAA
	CGG	CUU	AAU	AAC	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	UGC	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	UGU	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	GAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU	CAA	GGG	CAC	AAA	UUU	UCC	CCG	UCC	UCC	UAC	UAA
	CGU	CUU	AAU	CAA	GGG	GGG	AAA	UUU	UCC	CCG	UCC	UCC	GUU	UAG
	GCU	CUU	AAU											

superata tale soglia *polipeptide*. Una proteina può essere quindi sia un semplice peptide<sup>5</sup> che un singolo polipeptide o essere formata da più polipeptidi. La sequenza amminoacidica determina la struttura della proteina ed è proprio questo il collegamento fra il messaggio genetico nel DNA e la struttura tridimensionale che è associata alla sua funzione biologica.

Un amminoacido è una molecola organica formata da un atomo di carbonio centrale chiamato  $C_\alpha$  circondato da 4 componenti (vedi fig. 1.17):

1. un atomo di idrogeno
2. un gruppo amminico ( $\alpha - amino$ ), (-NH<sub>2</sub>) in condizioni fisiologiche carico positivamente (-NH<sub>3</sub><sup>+</sup>)
3. un gruppo carbossilico ( $\alpha - carboxyl$ ), (-COOH) carico negativamente (-COO<sup>-</sup>)
4. un gruppo R, gruppo laterale chiamato anche *residuo* che per sineddoche indica l'intero amminoacido una volta che questo si trova all'interno della catena proteica

Vi sono circa 20 amminoacidi proteinogenici diversi (come si può vedere nella figura 1.15 o 1.20). Il gruppo laterale non partecipa alla catena della *backbone* (spina dorsale) della proteina, resa stabile dai legami peptidici: rimane infatti libero di legarsi. È questo il "trucco" che consente alla proteina sia di ripiegarsi su sé stessa che di legarsi ad altre molecole. Gli amminoacidi possono essere polari, non polari, carichi (vedi figura 1.20) e causano differenti ripiegamenti della proteina. Di conseguenza ne influenzano la funzione, si pensi infatti al caso dell'anemia falciforme causata da 1 solo amminoacido di differenza: valina al posto del glutammato. La prima non è polare mentre il secondo è polare carico, ciò causa legami differenti, quindi ripiegamento differente e funzione biologica compromessa.

Gli amminoacidi esistono in 2 configurazioni: L e D. Essi sono infatti molecole *chirali*: le due configurazioni sono l'immagine speculare l'una dell'altra ma non sono sovrapponibili. Nella grande maggioranza degli organismi viventi le proteine sono composte solo da amminoacidi della serie L.

Il legame peptidico è il legame che unisce tutti gli amminoacidi di una proteina: unisce il gruppo carbossilico di un amminoacido al gruppo amminico di un altro amminoacido. È un tipo di legame molto stabile, infatti l'emivita della backbone è di 400 anni a 25°C<sup>[4]</sup>. Il legame peptidico comporta l'eliminazione della carica degli ex gruppi *amminico* e *carbossilico*.

Gli unici due residui elettricamente carichi rimasti in una proteina sono quelli alle due estremità (C-terminus ed N-terminus, vedi fig. 1.18). È presente però un fenomeno che permette ai residui di interagire elettrostaticamente: la *risonanza elettronica*. Gli elettroni

---

<sup>5</sup>Esempi di "semplici" peptidi che svolgono funzioni biologiche sono i *neuropeptidi* che agiscono da neurotrasmettitore (ad es. endorfine) e *ormoni* quali l'insulina, il glucagone e l'ossitocina, composta da soli 9 amminoacidi e implicata nelle contrazioni uterine e nella stimolazione dei dotti lattiferi delle mammelle.

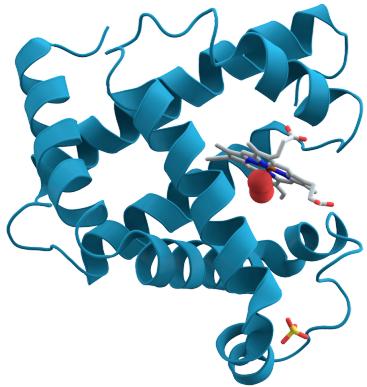


Figura 1.16: Rappresentazione a nastro della struttura tridimensionale della mioglobina. È presente un gruppo hemo al quale è legata una molecola di ossigeno (rossa). Fonte: [19]

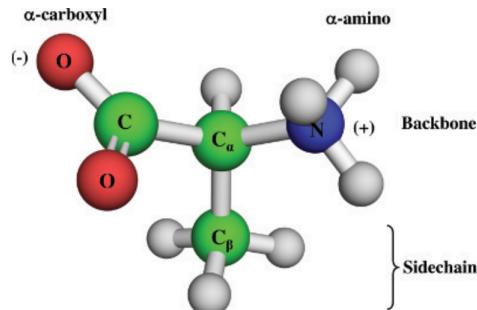


Figura 1.17: Struttura principale degli amminoacidi. Fonte [3]

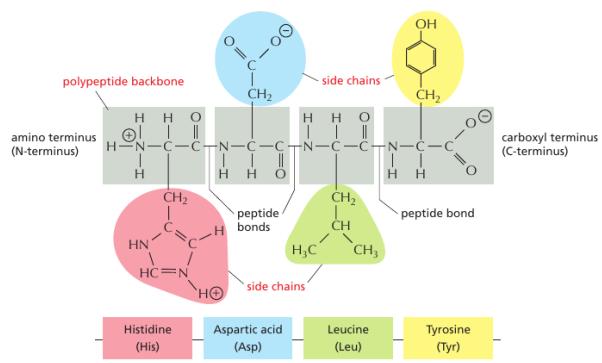


Figura 1.18: Backbone delle proteine. Fonte: [4]

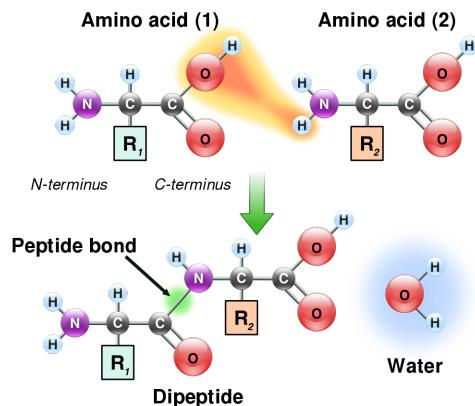


Figura 1.19: Legame peptidico. Fonte [20]

dei legami possono estendersi su più atomi e permettere al residuo di assumere diverse configurazioni elettroniche.

Nonostante gli amminoacidi siano solo 20, la varietà di proteine è elevatissima, in quanto gli amminoacidi si combinano tra loro in sequenze e quantità diverse. Dato un polipeptide di 100 amminoacidi si hanno  $20^{100}$  possibili combinazioni.

È possibile in realtà parlare anche di altri amminoacidi e di derivati. La *selenocisteina* è considerata il 21° amminoacido (così come la *pirrolisina* il 22°). È stata scoperta per la prima volta nel 1986 ed è codificato dal codone UGA, normalmente un codone di stop, che tuttavia in presenza di un particolare segmento di mRNA viene interpretato come elemento costitutivo. La sua struttura è identica a quella della cisteina con una sola differenza: un atomo di selenio al posto di quello di zolfo. Esistono poi una serie di derivati dagli amminoacidi. Si può dire ad esempio che la *tirosina* sia il precursore della

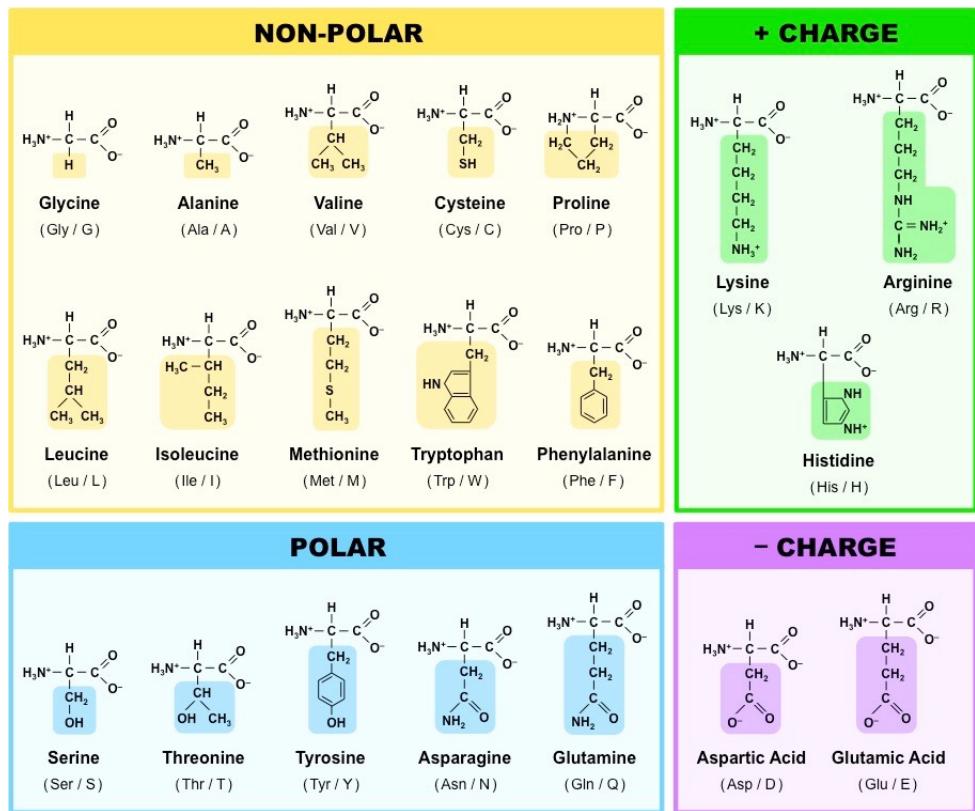


Figura 1.20: I 20 amminoacidi universali. Fonte: [21]

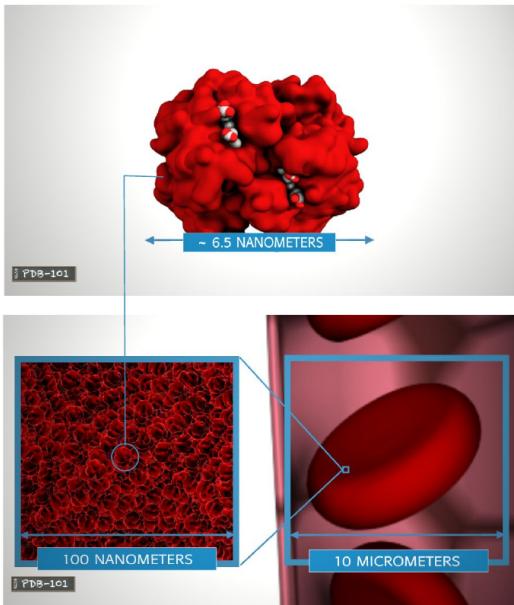
dopamina, melanina e adrenalina, il triptofano di serotonina e melatonina . Tipicamente questi derivati sono modificati dopo la traduzione nei ribosomi: la proteina in formazione viene modificata covalentemente da parte di enzimi e vengono a formarsi questi derivati.

Le proteine sono una classe di macromolecole con funzioni biologiche vitali, consentono infatti il funzionamento di ogni sistema vivente. Riusciamo a pensare, parlare, a digerire il cibo, a muoverci grazie alle proteine. Sono la base della vita cellulare e molecolare.

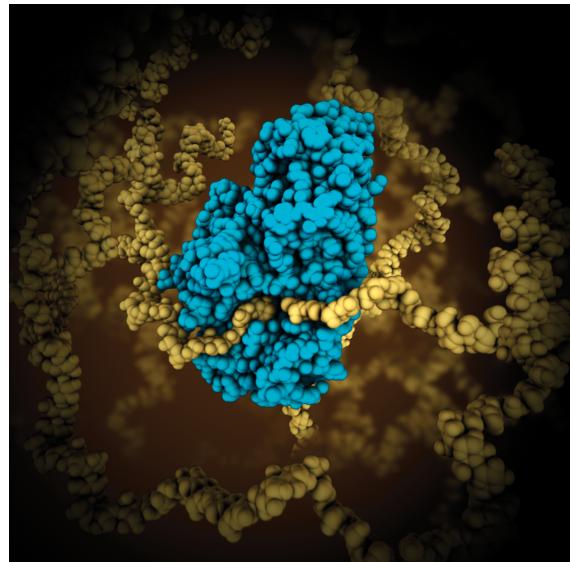
Un tipo fondamentale di proteine sono gli enzimi, come accennato inizialmente. Una loro funzione importante è correlata alla digestione negli animali. Enzimi come le *amilasi* e le *proteasi* sono in grado di ridurre le macromolecole (nella fattispecie amido e proteine) in unità semplici (maltosio e amminoacidi), assorbibili dall'intestino.

Oltre agli enzimi ci sono tante altre proteine importanti. Uno degli esempi più noti è l'emoglobina, proteina animale adibita a trasportare ossigeno dai polmoni ai tessuti così come a riportare CO<sub>2</sub> ai polmoni. Una molecola di emoglobina è composta da 4 polipeptidi e contiene 4 atomi di ferro che le consentono di legare reversibilmente 4 molecole di ossigeno.

Nelle cellule le proteine svolgono, fra le altre, funzioni di supporto strutturale (*collagene*), mobilità (*actina*, *miosina*), protezione (*anticorpi*), regolazione, ormoni (*insulina*),



*Figura 1.21: Emoglobina in diverse scale. Rapresentazione a superficie. Un globulo rosso contiene circa 280 milioni di molecole di emoglobina, per cui può portare più di 1 miliardo di molecole di ossigeno per volta. Fonte: [22]*



*Figura 1.22: Enzima alpha Amilasi in turchese, rappresentazione di tipo space-filling. Si lega a catene di carboidrati (gialle) e le rompe in pezzi più piccoli di glucosio. Fonte [22]*

trasporto, catalisi, magazzino. Nel nostro corpo abbiamo un numero grandissimo di proteine:  $10^{27}$ . Per usare una metafora di Ken Dill<sup>[23]</sup> potremmo dire che se si potesse ingrandire una proteina alla grandezza di un penny (diametro di 19mm) il numero di proteine che una persona ha nel corpo equivale al numero di penny che riempirebbero l'Oceano Pacifico.

Per queste e altre ragioni queste macromolecole sono il target di grandi attività di ricerca e di applicazione biotecnologiche: dal combattere malattie infettive<sup>[24]</sup> al contrastare l'inquinamento ambientale<sup>[25]</sup>.

## 1.2 Background informatico

### 1.2.1 Bioinformatica

La *bioinformatica* ha giocato un ruolo fondamentale durante l'epidemia di COVID-19, in particolare nella realizzazione di vaccini grazie agli avanzamenti nelle tecnologie NGS (Next Generation Sequencing). La bioinformatica è una disciplina dedicata alla risoluzione di problemi biologici a livello molecolare con metodi informatici, per questa ragione viene anche chiamata *biologia computazionale*. Argomenti di interesse di questa disciplina sono:

- allineamento di sequenze genetiche

- predizione genica
- predizione della struttura di proteine
- espressione genica
- interazione proteina-proteina
- interpretazione di dati proveniente da esperimenti biochimici
- organizzazione e archiviazione conoscenze su genomi e proteomi
- modellizzazione di sistemi e reti biologiche

Come si può notare da questa lista una parte importante della bioinformatica si occupa dell'utilizzo di strumenti informatici finalizzati a manipolare, archiviare e confrontare stringhe e sequenze di caratteri. Tuttavia questa disciplina non si ferma all'analisi delle sequenze. Tra le più interessanti applicazioni bioinformatiche odiere vi sono quelle incentrate sull'analisi strutturale<sup>[26]</sup>. Difatti la bioinformatica pone le sue fondamenta nel campo della *structural bioinformatics*: per portare un esempio il database PDB (*Protein Data Bank*) nasce nel 1977 per archiviare coordinate atomiche e legami derivati dagli studi cristallografici sulle proteine<sup>[27]</sup>.

Non va confusa la bioinformatica (o biologia computazionale) con la *computazione bioispirata* (es. algoritmi genetici, reti neurali), con il *biological computing* (ossia computer composti di parti biologiche come DNA, proteine o neuroni) o con la *biological computation* (l'idea che gli organismi eseguano computazioni e che le idee di informazione e computazione possano essere la chiave per comprendere la biologia)<sup>[28]</sup>.

Il Machine Learning (ML) è uno dei paradigmi informatici che più sta influenzando il campo della bioinformatica (come la presente tesi può dimostrare). Questo è dovuto principalmente a due fattori evolutisi in parallelo negli ultimi anni: la crescita esponenziale di dataset biologici disponibili e i progressi informatici del ML. Gli strumenti di ML possono apprendere caratteristiche dei sistemi biologici inferendole direttamente dai dataset. Quando propriamente allenati questi sistemi possono fornire accurate predizioni di caratteristiche astratte, proprio come nel caso di AlphaFold per il problema della predizione della struttura di proteine.

### 1.2.2 Soft computing

Il *soft computing* è un paradigma che si contrappone a quello dell'*hard computing*, ovvero la risoluzione di un problema tramite l'esecuzione di un algoritmo ben definito e decidibile. Il soft computing accantona la precisione od ottimalità e innalza a obiettivo il guadagno nella comprensione del comportamento di un sistema. Il soft computing si basa su due principi:

1. l'apprendimento a partire dai dati

2. l'integrazione di conoscenza umana basata sull'esperienza, strutturata e preesistente, all'interno di modelli matematici computabili

Il ML si avvale delle tecniche del soft computing<sup>[29]</sup> e vi entra pienamente: la stima di performance in ML è infatti l'*accuratezza predittiva*, stimata dall'errore calcolato sul test set.

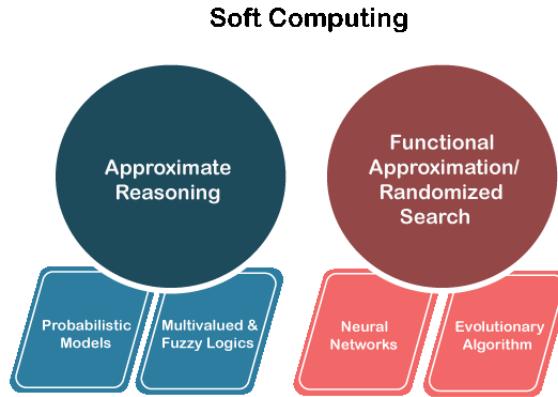


Figura 1.23: Branche del soft computing. Fonte: [30]

## Algoritmi genetici

Gli algoritmi genetici fanno parte del paradigma relativo alle tecniche informatiche *bio-ispirate*, così come le reti neurali. Un algoritmo genetico è un algoritmo euristico utilizzato per tentare di risolvere problemi di ottimizzazione. L'aggettivo "genetico", ispirato al principio della selezione naturale ed evoluzione biologica, deriva dal fatto che, al pari del modello evolutivo darwiniano che trova spiegazioni nella genetica, gli algoritmi genetici attuano dei meccanismi concettualmente simili a quelli dei processi biochimici genetici, come il *crossing over*.

### 1.2.3 Intelligenza Artificiale

Definire cosa sia l'intelligenza non è un compito semplice. Una definizione ampia e utilizzata nel mondo dell'AI è quella data da Kurzweil:

«*L'arte di creare macchine che svolgono funzioni che richiedono intelligenza quando svolte da esseri umani*»<sup>6</sup>

Una definizione di intelligenza proveniente da uno sfondo culturale del tutto diverso è la seguente:

<sup>6</sup>R. Kurzweil, R. Richter, R. Kurzweil et al., *The age of intelligent machines*, 1990

«*The role of intelligence is to determine the positive and negative potential of an event or factor which could have both positive and negative results. It is the role of intelligence, with the full awareness that is provided by education, to judge and accordingly utilize the potential for one's own benefit or well-being*»<sup>7</sup>

Nella sua accezione più semplice, l'Intelligenza Artificiale (AI) si riferisce a sistemi che imitano l'intelligenza umana per eseguire certe attività e che sono in grado di migliorarsi continuamente in base alle informazioni raccolte. L'IA si occupa della costruzione di macchine intelligenti, della comprensione mediante modelli computazionali dei comportamenti e della psicologia di uomini, animali e agenti artificiali e può avere applicazioni innumerevoli nella società. I fondamenti dell'IA sono sin dalla nascita interdisciplinari: filosofia, matematica, economia, neuroscienze, psicologia, informatica, linguistica, cibernetica, statistica, complessità, teoria del controllo, teoria dell'informazione, robotica.

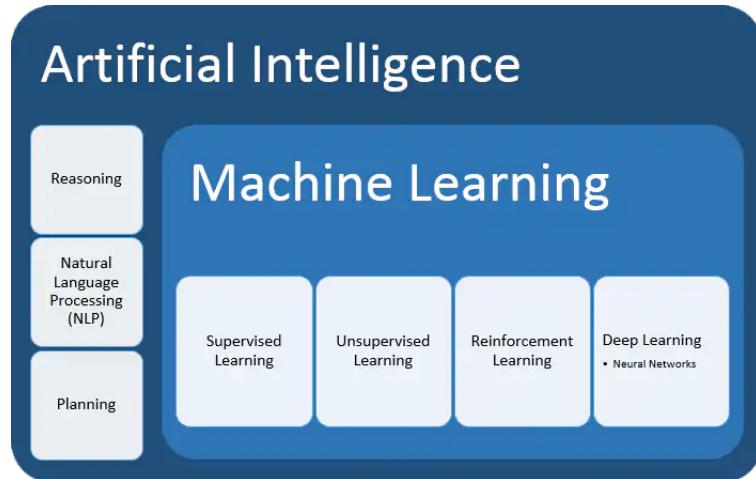


Figura 1.24: Schema riassuntivo dei campi dell'IA. Fonte: [33]

#### 1.2.4 Machine Learning

Il Machine Learning (ML) è un sottoinsieme dell'AI che si occupa di creare sistemi che automaticamente migliorano con l'esperienza, basandosi su rigorosi fondamenti delle scienze computazionali. Utilizza metodi statistici per migliorare la performance di un algoritmo nell'identificare pattern nei dati. Domande fondamentali di questo campo sono del tipo: "come varia la performance di apprendimento al variare del numero di esempi di allenamento presentati?".

<sup>7</sup>H. H. the XIV Dalai Lama, *The heart of the Buddha's path*, 2011

L'apprendimento è al cuore del problema dell'intelligenza sia bologica che artificiale ed è un principio universale comune a tutti gli organismi. Tom M. Mitchell definisce in questo modo l'apprendimento per una macchina:

«*Si dice che un programma apprende dall'esperienza E con riferimento ad alcune classi di compiti T e con misurazione della performance P, se le sue performance nel compito T, come misurato da P, migliorano con l'esperienza E.*»<sup>8</sup>

Il ML si divide in:

- *Supervised Learning*, ad es. SVM (support vector machine), in cui al modello vengono forniti degli esempi nella forma di possibili input e i rispettivi output desiderati e l'obiettivo è quello di estrarre una regola generale che associa l'input all'output corretto; comuni sono i task di classificazione e regressione
- *Unsupervised Learning*, in cui il modello ha lo scopo di trovare una struttura negli input forniti, come un raggruppamento naturale nei dati, senza che gli input vengano etichettati in alcun modo
- *Reinforcement Learning*, il modello interagisce con un ambiente dinamico nel quale cerca di raggiungere un obiettivo (per esempio guidare un veicolo, o imparare a giocare contro un avversario), avendo un insegnante che gli dice solo se ha raggiunto l'obiettivo
- *Deep Learning*, insieme di tecniche basate su reti neurali artificiali organizzate in diversi strati, dove ogni strato calcola i valori per quello successivo; si basa su diversi livelli di rappresentazione, corrispondenti a gerarchie di caratteristiche

Il ML è quindi sì uno strumento molto potente ma è importante comprenderne i limiti. È utile quando non esiste o è difficile da formalizzare la teoria attorno ad un problema, oppure quando i dati da analizzare sono incerti, rumorosi o incompleti.

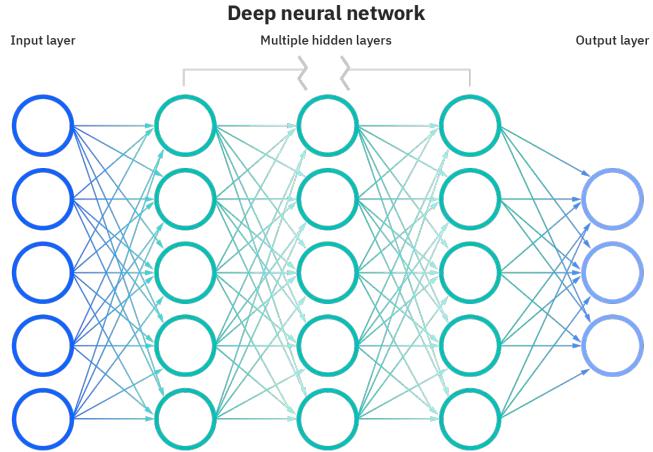
### 1.2.5 Reti neurali artificiali (ANN)

Una rete neurale artificiale (*Artificial Neural Network*) è un modello computazionale composto da neuroni artificiali bio-ispirato alla semplificazione di una rete neurale biologica. È importante notare che l'obiettivo della modellizzazione bio-ispirata non è una comprensione delle reti neurali biologiche, data la semplicità dei modelli utilizzati, ma il tentativo

---

<sup>8</sup>T. Mitchell, *Machine learning*. McGraw hill New York, 1997

di risolvere problemi ingegneristici sfruttando idee derivanti da queste. Nonostante ciò le ANN riflettono tratti di comportamento del cervello umano e consentono di riconoscere pattern e risolvere problemi difficili.



*Figura 1.25: Rete neurale artificiale. Fonte: [35]*

Le ANN sono composte da strati di nodi: uno strato di input, uno o più nascosti e uno di output. Ogni nodo è un neurone artificiale, si connette a tutti i nodi dello strato successivo e ha associato un peso e una soglia. Se l'output di un nodo è sopra la soglia allora il neurone è attivato, trasferendo informazioni al prossimo strato della rete. Con l'allenamento le ANN possono migliorare la loro accuratezza e rivelarsi potenti strumenti. Campi di utilizzo sono, fra gli altri, lo *speech-recognition* e l'*image recognition*.

La parola "deep" in *deep learning* si riferisce alla profondità degli strati in una rete neurale. Una rete neurale artificiale che consiste in almeno 4 strati (inclusi quello di input e output) può essere considerata un algoritmo di *deep learning*<sup>[35]</sup>. Una rete neurale con 2 o 3 strati è una rete neurale semplice.

# Capitolo 2

## Protein Folding

«la forma è l'immagine plastica della funzione»<sup>1</sup>

La correlazione tra forma e funzione si rivela fondamentale nel caso delle proteine. Un canale ionico neuronale permette il passaggio di ioni grazie alla sua forma a canale; una ferritina cattura e immagazzina gli ioni ferro grazie alla sua forma a sfera cava.

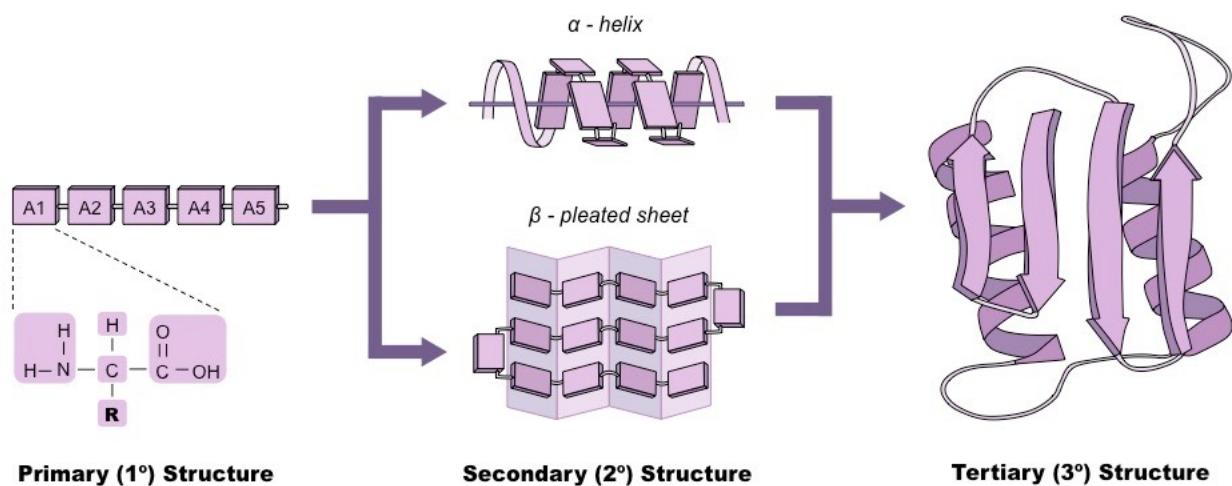


Figura 2.1: Protein folding: dagli amminoacidi alla struttura tridimensionale. Fonte: [37]

Il ripiegamento delle proteine (*protein folding*) è il processo di ripiegamento molecolare attraverso il quale a partire dalla sequenza lineare amminoacidica le proteine ottengono la loro struttura tridimensionale, chiamata forma *nativa*, che permette loro di svolgere la relativa funzione biologica<sup>2</sup>.

<sup>1</sup>A. Ruffini, *Fisiogenia, la biodinamica dello sviluppo ed i fondamentali problemi morfologici dell'embriologia generale*, 1925

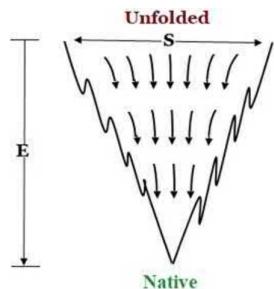
<sup>2</sup>per una trattazione superficiale della mancanza di generalità di questo paradigma si veda la sezione

Il ripiegamento nella forma tridimensionale avviene spontaneamente sia durante la sintesi proteica nei ribosomi sia al termine di questa. Una specifica proteina si ripiegherà nello stesso modo e avrà la stessa struttura finale<sup>3</sup>.

La prima teoria del ripiegamento proteico è stata proposta negli anni venti del XX secolo da Hsien Wu<sup>[38]</sup>, in relazione al processo di denaturazione (vedi sezione 2.1.2). È però Anfinsen, premio Nobel per la chimica, negli anni '60 a compiere un fondamentale passo nella comprensione del processo del ripiegamento proteico<sup>[39]</sup>.

## 2.1 Postulato di Anfinsen

Il postulato di Anfinsen (conosciuto anche come *dogma o ipotesi termodinamica* di Anfinsen) afferma che la struttura nativa delle proteine (almeno quelle globulari) è determinata solamente dalla sequenza di aminoacidi di cui sono costituite. In altri termini: la struttura nativa, in ambiente fisiologico standard, corrisponde a quella struttura unica, stabile e cinematicamente accessibile avente *minima energia libera*.



*Figura 2.2: Un profilo energetico idealizzato dell'energia libera a forma di imbuto. E=energia, S=entropia.*  
Fonte: [40]

Vi sono quindi 3 condizioni:

1. *unicità*, la sequenza non deve possedere altre configurazioni dotate di energia libera comparabile
  2. *stabilità*, piccoli cambiamenti nell'ambiente circostante non possono produrre cambiamenti nella configurazione a energia minima. Ciò può essere descritto come una superficie parabolica di energia libera con lo stato nativo corrispondente al punto
- 
- 2.4. Per una scrittura e lettura più agevole della presente tesi si è preferito accettare la visione del paradigma

<sup>3</sup>ciò non è vero nel 100% dei casi, alcune proteine possono avere più di una conformazione stabile per adempire funzioni diverse (vedi la sezione 2.4) e alcune proteine possono andare incontro a misfolding (vedi la sezione 2.3)

di minimo (visivamente simile ad un imbuto, vedi fig. 2.2); la superficie di energia libera nelle vicinanze dello stato nativo deve essere abbastanza ripida ed elevata

3. *accessibilità cinetica*, il percorso nella superficie di energia libera dallo stato *unfolded* a *folded* deve essere ragionevolmente piano

### 2.1.1 Esperimento di Anfinsen

L'esperimento, compiuto nel 1957<sup>[41]</sup>, consisteva nella denaturazione e rinaturazione della ribonucleasi A, dimostrando che il secondo processo era possibile senza agenti ausiliari. L'enzima in questione è formato da 124 amminoacidi, tra cui 8 cisteine che formano 4 punti disolfuro ( $-CH_2 - S-S - CH_2 -$ , vedi sez. 2.2.1). È stato usato un agente riducente per scindere questi punti e l'urea per denaturare la proteina: questa non mostrava più alcuna attività enzimatica. A questo punto se l'urea era rimossa prima, seguita dall'aggiunta di un agente ossidante per consentire ai punti disolfuro di riformarsi, la ribonucleasi A riacquistava spontaneamente la sua struttura terziaria e il prodotto ottenuto risultava praticamente indistinguibile dalla proteina nativa di partenza, riottenendo piena attività biologica.

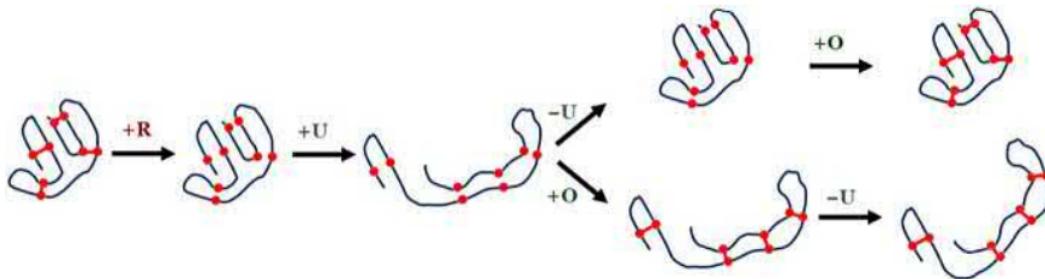


Figura 2.3: Rappresentazione schematica dell'esperimento di Anfinsen.  $R=$ reducing agent,  $U=$ Urea,  $O=$ oxidizing agent, punti rossi=cisteina, linee rosse=ponti disolfuro. Fonte: [40]

I punti disolfuro si riformano nella stessa posizione della proteina nativa nonostante ci siano 105 modi possibili per ricombinarli. Se invece veniva prima aggiunto l'agente ossidante e poi tolta l'urea il prodotto ottenuto era un miscuglio di molte delle possibili 105 configurazioni, raggiungendo solamente l'1% dell'attività enzimatica.

Dai lavori di Anfinsen è possibile trarre due ulteriori importanti conclusioni<sup>[44]</sup>:

- ha dato via alla grande avventura della ricerca nel campo del protein folding *in vitro*<sup>4</sup> piuttosto che all'interno di cellule. La struttura nativa non dipendeva quindi dal fatto che la proteina fosse sintetizzata biologicamente con l'aiuto di ribosomi (ed

<sup>4</sup>La locuzione latina *in vivo* significa *nel vivente*. Se il fenomeno biologico viene riprodotto in una provetta si dice *in vitro* mentre se lo si riproduce tramite una simulazione computazionale si dice *in silico*.

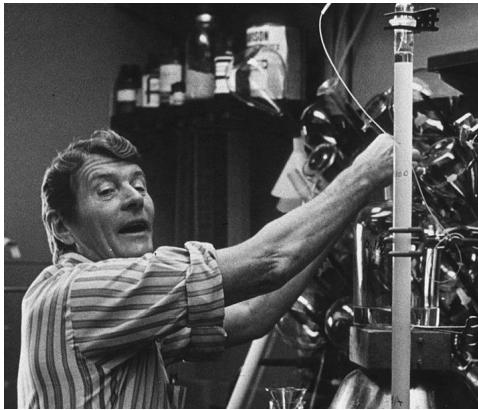


Figura 2.4: C.B. Anfinsen nel suo laboratorio.  
Fonte: [42]

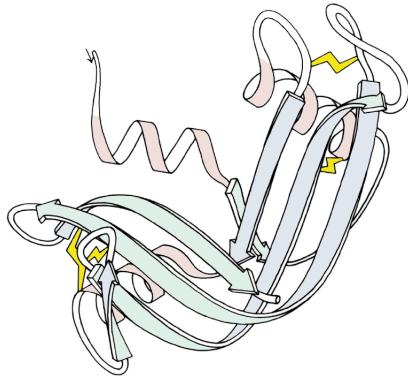


Figura 2.5: Ribonucleasi A, rappresentazione a nastro. In giallo i ponti disolfuro, rosa le  $\alpha$ -eliche, verde e azzurro i  $\beta$ -foglietti. Fonte [43]

eventualmente chaperoni molecolari) o che si ripiegasse nuovamente come molecola isolata all'interno di una provetta

- l'evoluzione può agire in modo da cambiare la sequenza amminoacidica ma l'equilibrio del ripiegamento e la cinetica di una data sequenza sono materia della fisica chimica.

### 2.1.2 Denaturazione

La denaturazione delle proteine è il fenomeno relativo all'alterazione della struttura nativa dovuto a variazioni di temperatura, pH o contatto con determinate sostanze chimiche. La denaturazione è un processo che porta alla perdita di ordine e quindi ad un aumento di entropia. La struttura primaria rimane invariata, data la stabilità dei legami peptidici. A causa della denaturazione le proteine perdono la loro funzione biologica e possono esporre e rendere reattivi alcuni gruppi funzionali che possono causare l'aggregazione di più proteine. Può avvenire che una volta rimosso l'agente denaturante la proteina ritorni allo stato di partenza (*rinnaturazione*) ma spesso il processo è irreversibile.

La proprietà di certe sostanze chimiche (es. urea) di denaturare una molecola proteica si deve alla loro capacità di legare transientemente, attraverso legami deboli, come ad esempio legami idrogeno, i residui amminoacidici costituenti la proteina. Questi legami vengono termodinamicamente preferiti a quelli intramolecolari o intermolecolari con l'acqua. Ciò comporta l'impossibilità per la proteina di mantenere la propria struttura tridimensionale e quindi questa si denatura.

Applicazioni nella vita quotidiana di questo fenomeno sono la cottura dei cibi (basti pensare all'albumina nell'uovo) e la permanente ai capelli (denaturazione dell' $\alpha$ -cheratina, rompendo e riformando ponti disolfuro).

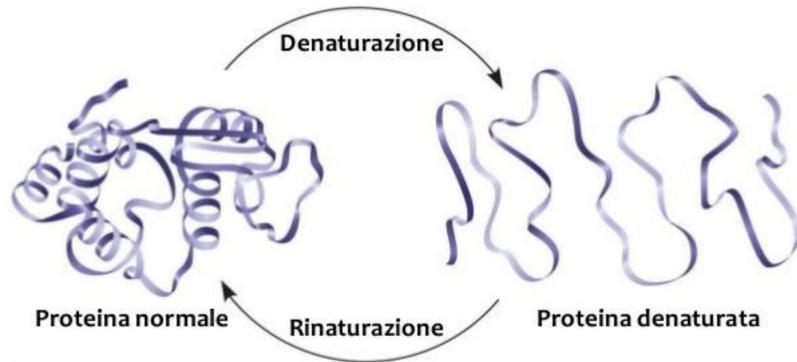


Figura 2.6: Denaturazione e rinaturazione. Fonte: [45]

## 2.2 Struttura delle proteine

Da un punto di vista chimico le proteine sono di gran lunga, tra quelle conosciute, le molecole strutturalmente più complesse e sofisticate funzionalmente. È possibile studiare la loro struttura individuando successivi livelli di organizzazione:

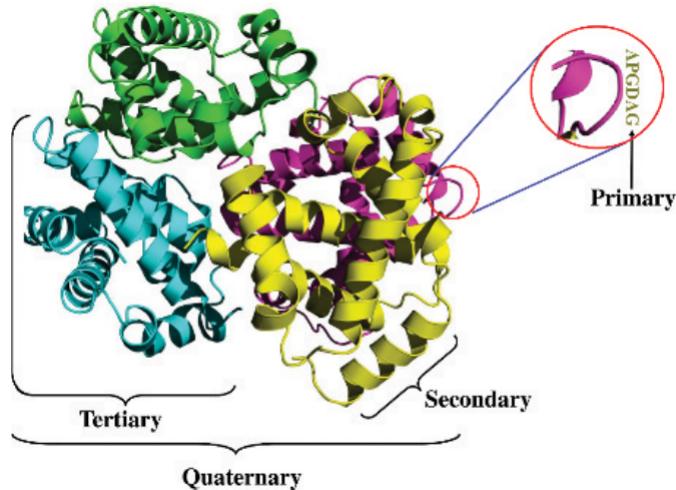


Figura 2.7: Livelli strutturali di una proteina. Fonte: [3]

- *struttura primaria*: la sequenza ordinata degli amminoacidi
- *struttura secondaria*: regioni ripetitive locali stabilizzate da legami idrogeno tra atomi della backbone ( $\alpha$ -eliche e  $\beta$ -foglietti)
- *struttura supersecondaria*: combinazione di strutture secondarie e connessioni (motivi, domini, loop, giri ...)
- *struttura terziaria*: forma tridimensionale di una singola catena polipeptidica, risultante dalle interazioni dei residui

- *struttura quaternaria*: forma finale di proteine "assemblate" da 2 o più catene polipeptidiche già ripiegate

Prima di passare ad analizzare ogni livello della struttura delle proteine è utile un veloce sguardo ai legami chimici e alle interazioni molecolari.

### 2.2.1 Legami e interazioni molecolari

La chimica della vita è di un tipo speciale: è una chimica organica formata da composti carboniosi, in un ambiente acquoso, con temperature "terrestri" e complicata, basata su grandi polimeri. Gli atomi possono risultare incompleti, e grazie a questo formare legami per completarsi. Elementi puri e pienamente completi non trovano spazio nella chimica della vita. Nei viventi solo gli elettroni si spostano<sup>5</sup> per ricercare stabilità, ovvero per permettere agli atomi di completare il loro guscio orbitale più esterno. Ogni atomo può avere tanti *legami* quanti elettroni gli mancano per completare il suo guscio più esterno. Le *interazioni molecolari* sono forze attrattive o repulsive tra molecole e tra atomi non legati. La *forza di legame* è la misura dell'energia necessaria per romperlo (in kJ/mol o kcal/mol). Si elencano ora i principali legami inerenti al ripiegamento delle proteine:

- *legame covalente*: prevede la partecipazione di 2 elettroni di valenza fra più atomi ed è il tipo di legame più forte. Due o più atomi tenuti insieme da legami covalenti formano una molecola. C'è una specifica distanza di legame fra i nuclei degli atomi bilanciata tra forze attrattive e repulsive: se sono troppo vicini c'è repulsione mentre se sono troppo lontani non c'è attrazione.
  - *elettronegatività*: spesso gli elettroni in un legame sono condivisi iniquamente. Questo dipende dall'elettronegatività degli atomi, ad esempio l'ossigeno ha elettronegatività 3.4 mentre l'idrogeno 2.1. Quando la differenza di elettronegatività è compresa tra 0.5 e 1.9 la nube elettronica di legame risulta deformata verso l'atomo più elettronegativo, su cui si origina una carica parziale negativa (indicata con  $\delta^-$ ) mentre l'altro atomo acquisisce una carica parziale positiva di uguale valore assoluto. La molecola, divenuta *polare*, si può immaginare ora come un *dipolo* elettrico.
  - *ponti disolfuro*: i legami (o ponti) disolfuro sono legami covalenti tra due atomi di zolfo con energia di legame di 60kcal/mol. Si formano dall'accoppiamento di due gruppi tiolici (-SH). Essendo legami molto forti costituiscono un elemento architettonico fondamentale nella struttura delle proteine. La cisteina presenta un gruppo -SH nella catena laterale e può quindi formare ponti disolfuro.

---

<sup>5</sup>Protoni e neutroni si separano solo in condizione estreme: nei reattori nucleari, nel sole, per decadimento radioattivo.

Bond Type	Length* (nm)	Strength (kJ/mole)	
		In Vacuum	In Water
Covalent	0.10	377 [90]**	377 [90]
Noncovalent: ionic bond	0.25	335 [80]	12.6 [3]
Noncovalent: hydrogen bond	0.17	16.7 [4]	4.2 [1]
Noncovalent: van der Waals attraction (per atom)	0.35	0.4 [0.1]	0.4 [0.1]

Figura 2.8: Distanza di legame approssimate e forza dei legami chimici. I valori della forza sono riportati in kJ/mol e in [kcal/mol]. Da notare la diminuzione di forza nel legame ionico se in ambiente acquoso. Fonte: [4]

- *legami non covalenti (interazioni molecolari)*
  - *attrazioni elettrostatiche*: le forze d’attrazione agiscono fra gruppi completamente carichi (legame ionico) e fra i gruppi parzialmente carichi delle molecole polari. Decresce con la distanza. Molto deboli in acqua.
    - \* *legame ionico*: l’atomo più elettronegativo strappa completamente un elettrone al suo compagno, si formano due ioni (uno positivo, *catione* e uno negativo *anione*). Si ha quando la differenza di elettronegatività tra i due atomi è maggiore di 1.9.
  - *legame idrogeno*: è una forza dipolo-dipolo che si origina tra molecole contenenti un atomo di idrogeno unito covalentemente a ossigeno, fluoro o azoto. Un atomo di idrogeno elettropositivo è parzialmente condiviso da due atomi elettronegativi; ad es. nell’acqua gli atomi di idrogeno (parzialmente positivi) si trovano fra due atomi di ossigeno (parzialmente negativi). L’idrogeno, legato a uno dei due atomi di ossigeno, permette all’altro di avvicinarsi e di stabilizzare le molecole. Sono legami deboli singolarmente (1/20 della forza di un legame covalente) ma quando se ne formano simultaneamente molti sono abbastanza forti da fornire un legame stretto (l’acqua bollirebbe a -120°C senza legami idrogeno).
  - *interazioni di van der Waals*: nelle molecole apolari gli elettroni si possono accumulare in modo asimmetrico, formando regioni momentaneamente polari che permettono così una temporanea stabilizzazione fra molecole a breve distanza. Due atomi saranno attratti l’uno dall’altro fino a che la distanza fra i loro nuclei è approssimativamente uguale alla somma dei loro raggi di van der Waals (ad es. per il carbonio il raggio è di 0.2nm)

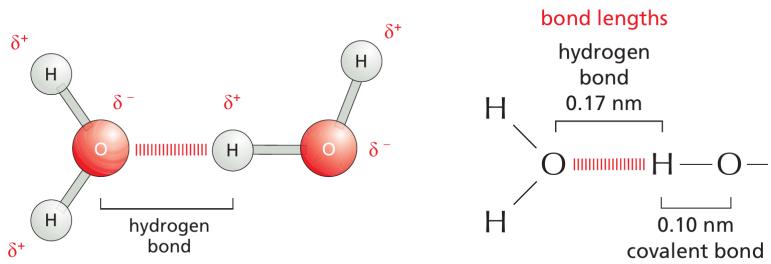


Figura 2.9: Legame idrogeno tra due molecole d'acqua. Fonte: [4]

- *forze idrofobiche*: l'acqua forza insieme i gruppi idrofobici; l'apparente attrazione è in realtà causata da una repulsione dall'acqua, che difende il suo reticolo tenuto insieme da legami idrogeno.

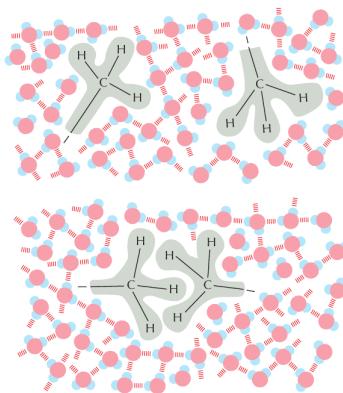


Figura 2.10: Forze idrofobiche. Fonte: [4]

Le sostanze *idrofile* si dissolvono rapidamente nell'acqua poiché le loro molecole formano legami idrogeno con le circostanti molecole d'acqua (nel caso di sostanze polari) o perché queste sono attratte dalle cariche degli ioni (nel caso di sostanze ioniche, es. cloruro di sodio, con ioni  $Na^+$  e  $Cl^-$ ). Le sostanze *idrofobiche* contengono perlopiù legami non polari e sono solitamente insolubili in acqua. Le molecole d'acqua in questo caso non sono attratte ma possono generarsi forze idrofobiche che raggruppano insieme tali sostanze (come nel nucleo idrofobico delle proteine).

## 2.2.2 Livelli strutturali

### Struttura primaria

La struttura primaria delle proteine è la sequenza ordinata degli amminoacidi. La posizione nella sequenza di specifici amminoacidi è un fattore fondamentale per la determinazione di quali porzioni della proteina andranno a legarsi formando globalmente la struttura finale. La nota importante, basata sul dogma di Anfinsen, è che la sequenza amminoacidica di

ogni proteina contiene l'informazione che specifica sia la struttura nativa che la via per raggiungere quello stato. Questo comunque non vuol dire che strutture simili si ripieghino in modo simile.

## Struttura secondaria

La struttura secondaria riguarda le regioni ripetitive locali stabilizzate da legami idrogeno tra atomi della backbone:  $\alpha$ -eliche e  $\beta$ -foglietti.

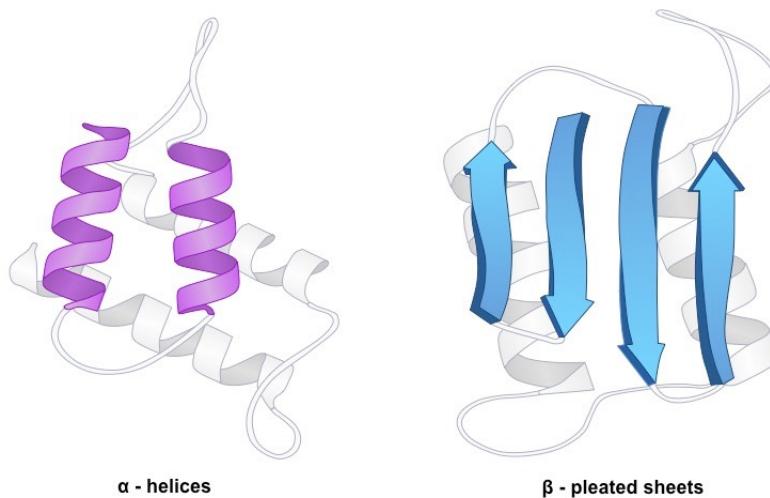


Figura 2.11: Struttura secondaria delle proteine,  $\alpha$ -eliche e  $\beta$ -foglietti. Fonte: [37]

Questo livello di organizzazione è una conseguenza dei legami a idrogeno intramolecolari. All'interno della backbone del polipeptide gli atomi di ossigeno hanno una parziale carica negativa e gli atomi di idrogeno attaccati all'azoto hanno una parziale carica positiva perciò possono formarsi legami idrogeno fra questi atomi. Individualmente sarebbero deboli legami ma poiché sono ripetuti molte volte su di una regione relativamente lunga di una catena polipeptidica possono fare da supporto per una particolare conformazione.

Nella struttura ad  $\alpha$ -elica, la struttura secondaria più comune e teorizzata già negli anni '50 da Linus Pauling, gli amminoacidi sono avvolti in una spirale tenuta insieme da legami idrogeno ogni 4 amminoacidi. Tra l'atomo di idrogeno legato all'azoto di ogni legame peptidico e l'ossigeno del gruppo carbossilico del legame peptidico sovrastante (che si trova a distanza di tre amminoacidi lungo la catena) si instaura un legame a idrogeno. Tuttavia se gli amminoacidi che si succedono lungo un tratto di catena proteica hanno gruppi R voluminosi, come avviene nella prolina, o gruppi R dotati della stessa carica elettrica, come avviene negli amminoacidi lisina e arginina, l' $\alpha$ -elica non può formarsi, a causa delle forze di repulsione che si generano tra i residui. Alcune proteine fibrose, come

l' $\alpha$ -cheratina, la proteina strutturale di capelli, lana e unghie hanno formazioni di  $\alpha$ -eliche sulla maggior parte della loro lunghezza.

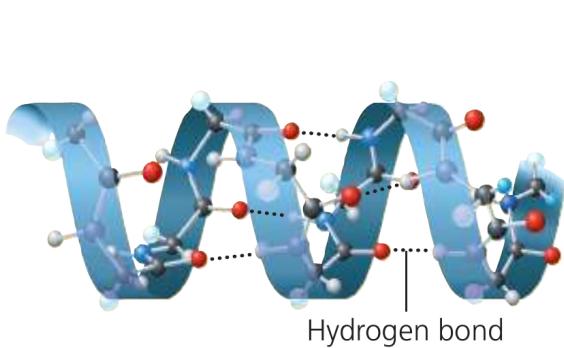


Figura 2.12: Regione di  $\alpha$ -elica. Fonte: [45]

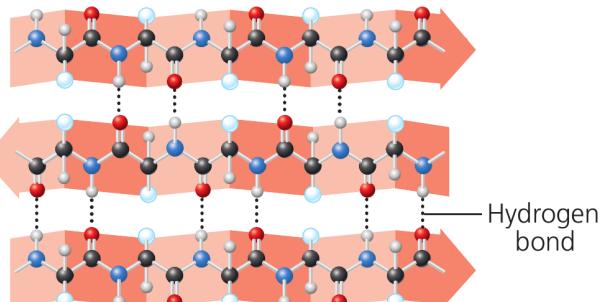


Figura 2.13: Una regione di  $\beta$ -foglietto composto da  $\beta$ -filamenti adiacenti, spesso mostrati come una freccia piegattata o piatta puntata in direzione C-terminus. Fonte [45]

Altre proteine fibrose sono invece dominate dai  $\beta$ -foglietti, come le proteine della seta ( $\beta$ -cheratina) e della tela prodotta dai ragni. In queste conformazioni due o più segmenti della catena polipeptidica giacenti lato su lato (chiamati  $\beta$ -filamenti) sono connessi da tre o più legami idrogeno. Si definisce  $\beta$ -filamento una sequenza peptidica di aminoacidi (tipicamente 5-10) che si dispone linearmente ed è in grado di formare legami idrogeno. Ciascuna delle catene è totalmente estesa e presenta una conformazione a zig-zag, dovuta alla geometria dei legami attorno a ciascun atomo di carbonio e di azoto nella catena.

I gruppi amminici di uno scheletro peptidico formano legame con quelli carbossilici del filamento opposto. In ogni singolo filamento i residui si dispongono perpendicolarmente al piano del foglietto, puntando alternativamente verso l'alto e verso il basso. I  $\beta$ -foglietti tendono a trovarsi all'interno del nucleo della struttura per evitare competizione con le molecole d'acqua per formare legami idrogeno e tendono a favorire residui idrofobici. Si dice che i filamenti sono paralleli quando vanno nella stessa direzione (la freccia che indica la direzione C-terminus è puntata nella stessa direzione).

Nella vita quotidiana, se tiriamo per i due estremi una fibra di lana questa si allunga: si stanno rompendo i legami idrogeno e le eliche si allontanano sempre di più, ma lasciando la presa i legami idrogeno si riformano e le eliche ricompaiono nella struttura. Se invece tiriamo la seta si può osservare che non è elastica: i foglietti di cui è composta la sua struttura non sono smantellabili senza rompere anche i legami covalenti della backbone.

## Struttura supersecondaria

La struttura supersecondaria è riferita alle combinazioni spaziali di strutture secondarie in conformazioni più complesse e alle connessioni che li uniscono<sup>6</sup>. Può essere considerata come esempio di struttura supersecondaria la triplice elica allungata del collagene.

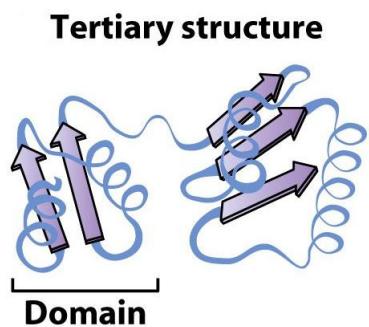


Figura 2.14: Dominio in una proteina. Fonte: [46]

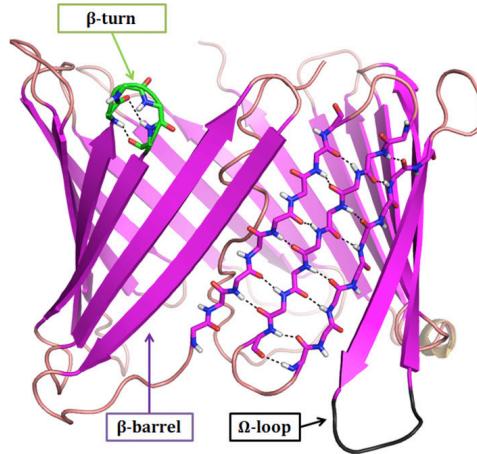


Figura 2.15: Struttura con giri, loop e motivo  $\beta$ -barile. Fonte: [47]

I *motivi* (motifs) e *domini* (domains) sono regioni tridimensionali della catena polipeptidica formate da differenti strutture secondarie adibite a svolgere una determinata funzione per la proteina di cui fanno parte. Tuttavia sono differenti in quanto i motivi non mantengono la loro forma se separati dalla proteina laddove i domini la mantengono. Questo perché i motivi e il resto della proteina sono più vicini e si vengono così a formare legami idrogeno che permettono ai motivi di mantenere la struttura. I domini sono sì legati alla backbone della proteina ma non abbastanza vicini alla restante parte della formazione proteica da stabilire legami, pertanto se vengono separati non perdono la loro struttura e possono mantenere la loro funzione. Una proteina con vari domini può usare questi per interazioni funzionali con differenti molecole.

Più in generale un *motivo strutturale* è una struttura tridimensionale comune che appare in una varietà di molecole differenti ed evoluzionisticamente scollegate. Nel contesto delle sequenze amminoacidiche si definisce *motivo* un pattern amminoacidico conservato in un gruppo di proteine con attività biochimica simile.

In figura 2.16 sono illustrati alcuni motivi comuni nelle strutture proteiche. Il motivo *elica-loop-elica* ad esempio consiste di due  $\alpha$ -eliche collegate da un giro invertito. Un motivo simile è l'*elica-giro-elica* dove al posto di un loop si ha un giro che causa un cambio di direzione più netto. Questa particolare conformazione rende questo motivo in grado di

<sup>6</sup>Non c'è un accordo tra i vari studiosi su di una precisa classificazione di questo livello strutturale.

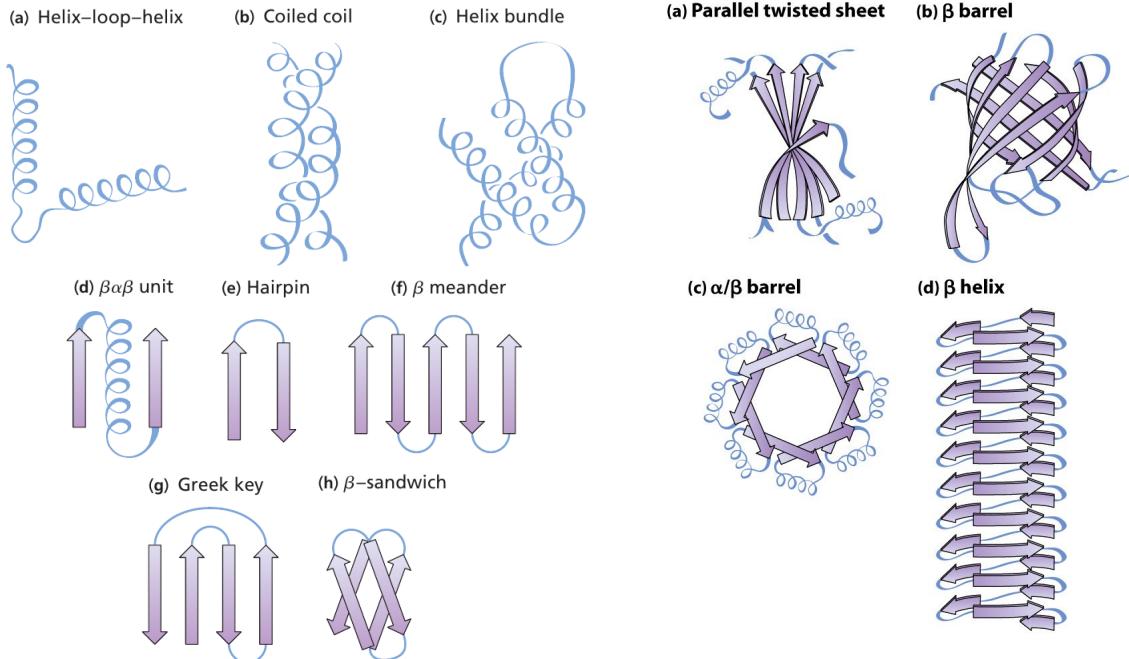


Figura 2.16: Motivi comuni. Fonte [46]

Figura 2.17: Domain folds (ripiegamenti di dominio). Fonte: [46]

legarsi alla scanalatura del DNA e infatti questo motivo si presenta in molte proteine che regolano l'espressione genica.

Il motivo a *simbolo greco* consiste di 4  $\beta$ -filamenti antiparalleli in un  $\beta$ -foglietto dove l'ordine dei foglietti lungo la catena polipeptidica è 4,1,2,3<sup>7</sup>. In figure 2.15 e 2.17 è illustrato il motivo  $\beta$ -barile composto da  $\beta$ -foglietti ripiegati circolarmente a formare una struttura somigliante ad un barile comune in molte proteine di membrana. I *domain folds*, o ripiegamenti di dominio, sono grandi motivi che costituiscono il nucleo di un dominio.

*Giri* e *loop* causano cambi di direzione alla backbone della proteina. I loop sono regioni con una struttura tridimensionale fissa ma non regolare. Si trovano generalmente sulla superficie delle proteine. Non sono strutture casuali e non vanno confuse con regioni disordinate o dispiegate. Hanno principalmente lo scopo di connettere strutture secondarie tra loro. È stato ipotizzato che la posizione degli introni nel DNA possa correlare con la locazione dei loop codificati nella proteina<sup>[48]</sup>.

Nelle strutture secondarie e terziarie si trovano spesso bruschi cambiamenti di direzione nella struttura: i *giri* (turns). Queste nette svolte sono possibili grazie agli amminoacidi prolina e glicina. Il gruppo R della prolina si ripiega verso il gruppo amminico, distorcendo la catena naturalmente. Si forma però uno stretto spazio a causa del giro: l'amminoacido con gruppo R meno voluminoso è ovviamente la glicina ed è per questo che si trovano insieme nei giri.

<sup>7</sup>I numeri indicano l'ordine dei filamenti ovvero la loro posizione nel  $\beta$ -foglietto da destra a sinistra

## Struttura terziaria

La struttura terziaria è la struttura tridimensionale globale risultante dalle interazioni tra i residui successivamente alle conformazioni locali della struttura secondaria ed è quindi la descrizione del risultato del processo di ripiegamento proteico. Un tipo di interazione importante è quella idrofobica che induce i residui non polari (e quindi idrofobici) a raggrupparsi al centro della catena polipeptidica, formando un *nucleo idrofobico*. La forma della proteina può venire rinforzata dai ponti disolfuro, legami covalenti possibili solamente fra due cisteine.

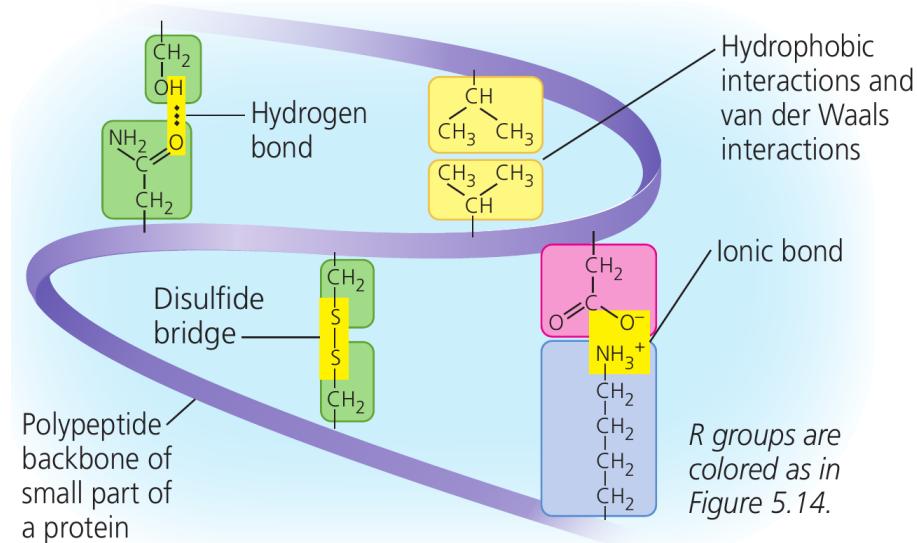


Figura 2.18: I diversi tipi di interazioni che possono contribuire alla struttura terziaria di una proteina.

Fonte: [45]

La glicina assume una speciale posizione tra gli aminoacidi dato che ha il gruppo R più piccolo, un solo atomo di idrogeno (vedi fig. 1.20): può aumentare la flessibilità locale nella struttura (come infatti accade nel caso dei *giri* sopra accennati).

Prima degli anni '80 il protein folding code (bilancio termodinamico delle forze interatomiche, vedi sez. ??) era visto come la somma di molte piccole interazioni (legami idrogeno, interazioni di van der Waals, attrazioni elettrostatiche ...) ma senza nessuna forza dominante<sup>[44]</sup>. Negli anni '80, grazie alla modellazione basata sulla meccanica statistica, è emerso un nuovo paradigma: la componente dominante nel folding code sono le forze idrofobiche, il folding code è distribuito sia localmente che non localmente nella sequenza e le strutture secondarie di una proteina sono una conseguenza della struttura terziaria tanto quanto una causa. Poiché le strutture native sono solamente 5-10kcal/mol più stabili dei loro stati denaturati è chiaro che nessuna forza intermolecolare può essere ignorata, e per questo la questione su quale forza sia quella dominante non è né semplice

né risolta. Tuttavia risulta evidente che solo pochi residui risultano carichi nelle proteine, pertanto le forze elettrostatiche difficilmente possono essere dominanti.

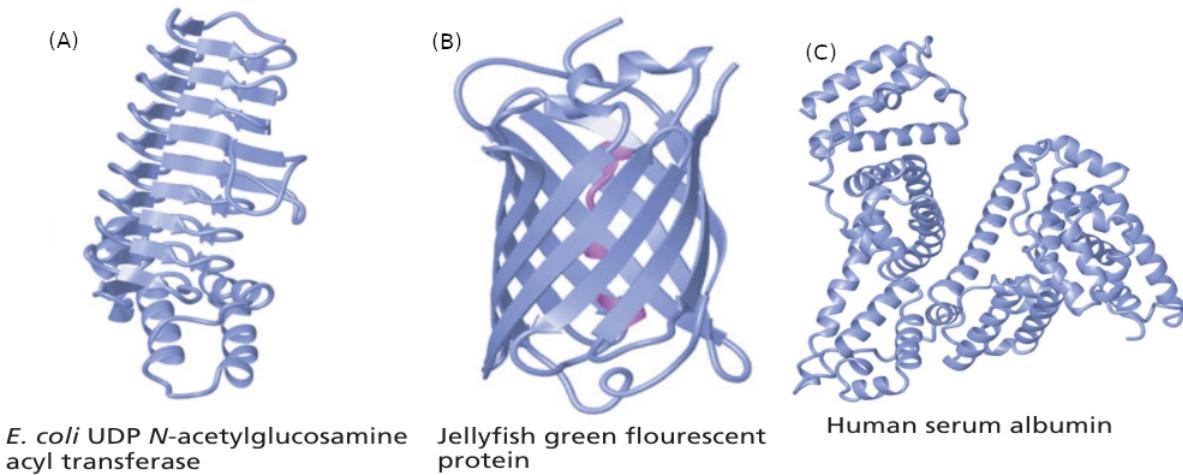


Figura 2.19: Esempi di strutture terziarie in alcune proteine. (A) (classe: all- $\beta$ ) La struttura dell'enzima mostra un classico esempio di  $\beta$ -eliche, struttura abbastanza rara. (B) (classe: all- $\beta$ ) Struttura a  $\beta$ -barile con un' $\alpha$ -elica centrale; i  $\beta$ -filamenti sono anti-paralleli. (C) (classe: all- $\alpha$ ) Albumina del siero umano. Ha molti domini costituiti da  $\alpha$ -eliche a strati e helix bundle. Fonte [46]

## Struttura quaternaria

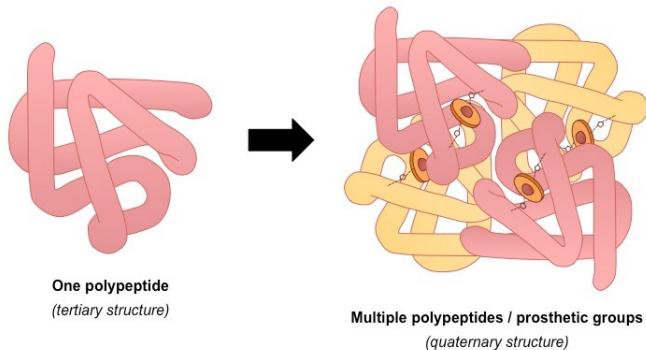


Figura 2.20: Rappresentazione di una struttura quaternaria composta da più polipeptidi e alcuni gruppi prostetici. Fonte [37]

La struttura quaternaria è la forma finale di proteine "assemblate" da 2 o più catene polipeptidiche già ripiegate. Il collagene ne è un esempio poiché è formata da 3 polipeptidi quasi interamente a spirale che si attorcigliano l'uno sull'altro formando un'elica tripla ancora più larga, dando alle lunghe fibre una grande forza (vedi anche la cheratina nella sezione 2.2.3). Un altro esempio è l'emoglobina, proteina globulare formata da 4 subunità polipeptidiche. Le strutture terziarie delle subunità non vengono alterate.

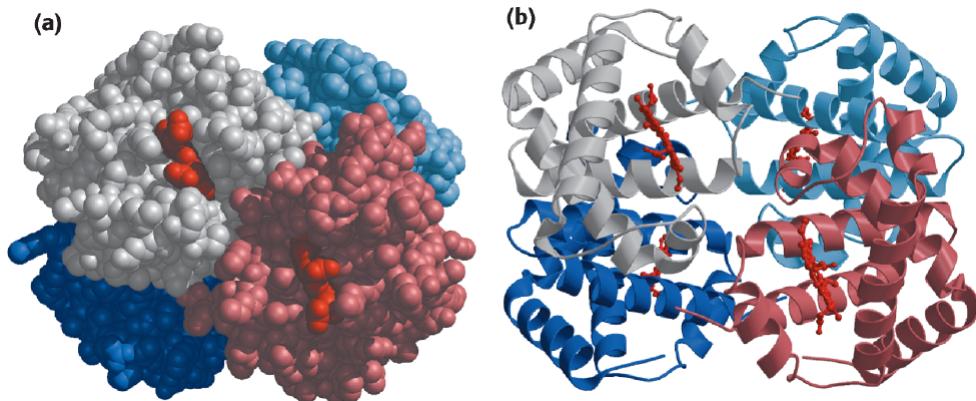


Figura 2.21: Struttura quaternaria della deossiemoglobina, ogni colore rappresenta una diversa subunità. (A) rappresentazione di tipo space-fill. (B) Rappresentazione a nastro. Fonte [46]

La struttura che ne risulta è spesso chiamata *oligomero* e le catene polipeptidiche costituenti sono dette *monomeri*, *protomeri* o *subunità*. Le subunità sono unite mediante legami idrogeno, ionici o forze idrofobiche. I rapporti spaziali tra le subunità sono fissi e la geometria della molecola globale è ben definita. L'unione delle subunità può far emergere proprietà non possedute dai singoli monomeri.

## 2.2.3 Evoluzione e classificazione

### Evoluzione e conservazione

L'evoluzione degli organismi è legata a mutazioni spontanee che avvengono nei loro geni. Le differenze nella struttura primaria sono la “memoria” dei cambiamenti avvenuti a livello genetico nel corso dell’evoluzione. In specie legate da notevole affinità le strutture primarie delle proteine comuni sono simili. Proteine con funzioni analoghe presentano sequenze simili, è molto probabile quindi che queste sequenze si siano evolute a partire da un progenitore comune.

Al contrario di quanto si possa credere la maggior parte dei *motivi* non ha origini evolutive in comune. Motivi simili sono sorti indipendentemente e semplicemente convergono verso una struttura stabile comune. Il fatto che gli stessi motivi si presentino in centinaia di differenti strutture suggerisce l'esistenza di un numero limitato di possibili ripiegamenti nell'universo delle strutture proteiche<sup>[49]</sup>.

### Classificazione

La classificazione delle proteine all'interno dei database può essere basata su somiglianze strutturali e/o di sequenza. A livello biologico, in base ai diversi livelli strutturali assunti, le proteine sono classificate in *fibrose* e *globulari*.

Le proteine fibrose sono caratterizzate dalla prevalenza di strutture secondarie rispetto a livelli di organizzazione superiore. Sono costituite da lunghe catene disposte in lunghi fasci o foglietti. La struttura è estremamente ordinata. Svolgono funzioni di protezione e sostegno. Le proteine fibrose costituiscono prevalentemente: pelle, piume, capelli, corna, unghia, squame (con funzione di protezione) e cartilagine, tendini, ossa (con funzione di sostegno). Contengono per la maggior parte residui idrofobici, pertanto le proteine fibrose risultano insolubili in acqua. Esempi di proteine fibrose sono: cheratina, collagene, elastina, fibroina.

In figura 2.22 è mostrata la struttura di un capello. La proteina dei capelli è l' $\alpha$ -cheratina, una struttura totalmente ad  $\alpha$ -eliche. Un paio di queste eliche si attorcigliano per formare una doppia elica. Queste si combinano poi in strutture di ordine superiore chiamate *protofilamenti* e successivamente *protofibrille*. Circa 4 fogli di protofibrille ( $4 \times 2^3 = 32$  eliche di  $\alpha$ -cheratina in tutto) si combinano per formare un filamento intermedio. Un capello è una schiera di filamenti intermedi di  $\alpha$ -cheratina.

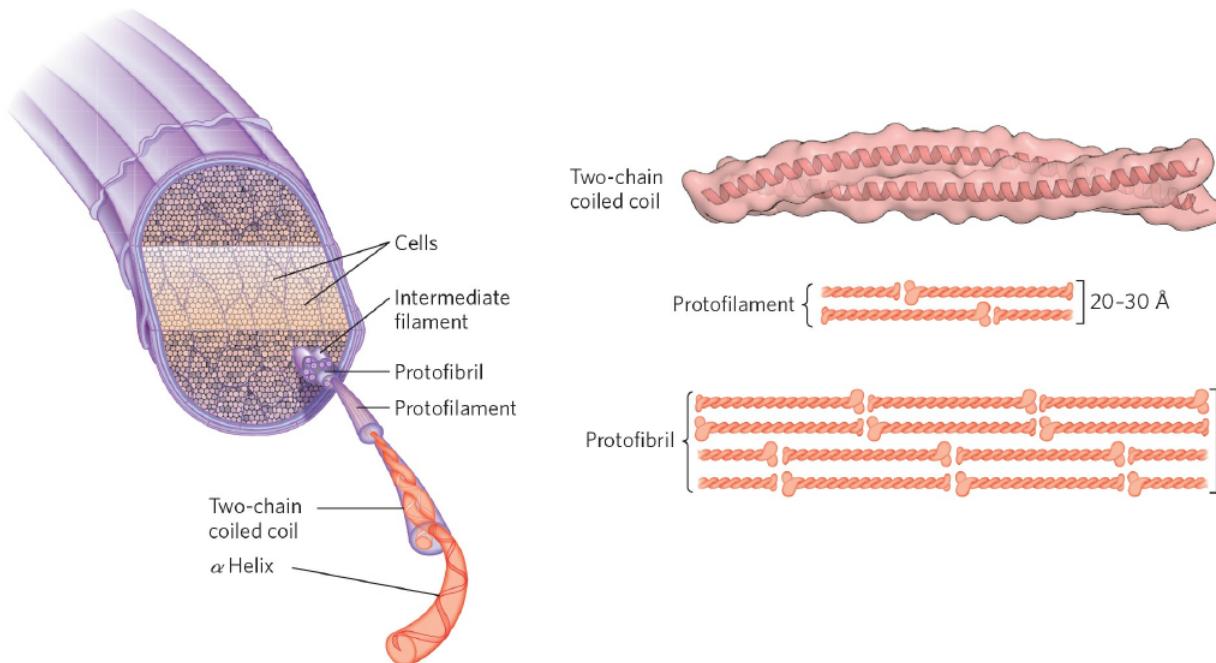


Figura 2.22: Struttura di un capello. Fonte: [50]

Le proteine *globulari* assumono una struttura terziaria e a volte quaternaria. Sono macromolecole compatte di forma all'incirca sferica. Hanno una struttura meno ordinata rispetto a quelle fibrose. Svolgono funzioni di catalisi, trasporto e regolazione di processi cellulari. Categorie di proteine globulari sono: enzimi, trasportatori di ossigeno e lipidi, alcuni ormoni, recettori di membrana e anticorpi. La struttura è caratterizzata da brevi

tratti di  $\alpha$ -elica e struttura  $\beta$ , collegate da tratti non organizzati in struttura secondaria. Sono proteine solubili nel citosol.

## 2.3 Dinamica del ripiegamento

### 2.3.1 Geometria ed energetica del ripiegamento

#### Geometria del ripiegamento

Le tante conformazioni possibili della catena polipeptidica sono possibili grazie alla rotazione di essa attorno all'atomo  $C_\alpha$  di ogni amminoacido.

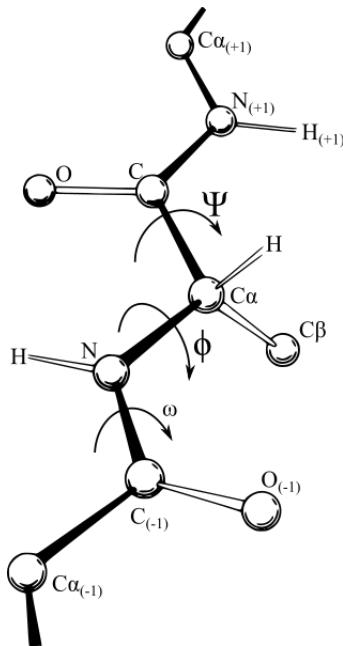


Figura 2.23: Angoli di torsione intorno all'atomo  $C_\alpha$

I tre angoli di torsione principali di un polipeptide sono  $\phi$ ,  $\psi$  ed  $\omega$ ; i legami  $N-C_\alpha$  e  $C_\alpha-C$  sono relativamente liberi nella rotazione (angoli  $\phi$  e  $\psi$ ), mentre il legame  $N-C$  ( $\omega$ ) è fisso dato il carattere di doppio legame parziale del legame peptidico alle temperature fisiologiche.

Data la relativa libertà dei due legami si ha la possibilità di isomeria: le due configurazioni possibili sono *cis* e *trans*. Delle due configurazioni possibili, la *trans* è quella favorita dal punto di vista energetico (minima repulsione sterica), infatti oltre il 99% dei legami peptidici delle proteine naturali hanno configurazione *trans*.

Nel grafico di Ramachandran sono riportati in ordinata i valori di  $\Psi$  ed in ascissa i valori di  $\Phi$ . Ogni puntino (non presenti in figura) rappresenta la posizione di un residuo.

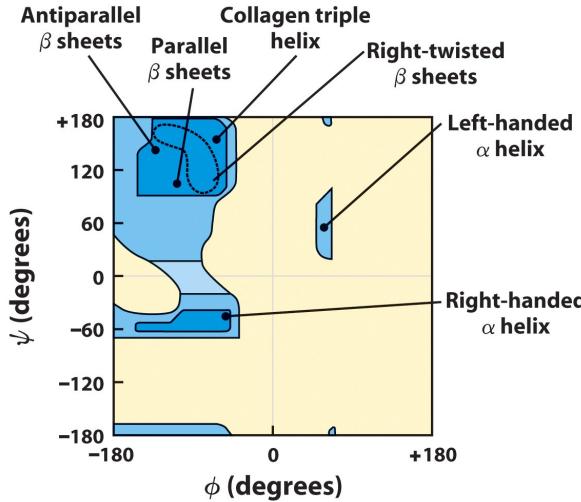


Figura 2.24: Grafico di Ramachandran. Fonte [51]

Il grafico deriva da un modello in cui è simulata la variazione di struttura di piccoli polipeptidi e successivamente si cercano conformazioni stabili. Per ciascuna conformazione sono esaminati i contatti tra atomi, trattati come sfere solide di dimensioni determinate dai raggi di van der Waals. Le conformazioni non consentite sono quelle per le quali sono previste collisioni tra sfere. In figura 2.24 l'area gialla corrisponde a conformazioni instabili in cui atomi della catena sono ad una distanza inferiore alla somma dei raggi di van der Waals. Tali regioni sono stericamente non consentite per tutti gli aminoacidi, fatta eccezione per la glicina, priva di catena laterale. Le regioni blu indicano conformazioni consentite ( $\beta$ -foglietti e  $\alpha$ -eliche destrogiro). Le aree azzurre mostrano le regioni consentite riformulando i calcoli con raggi di van der Waals più corti, ovvero consentendo una prossimità atomica maggiore.

Conformazioni ideali (in cui non vi sono interazioni tra catene laterali):

- $\alpha$ -elica:  $\Phi = -57^\circ$ ,  $\Psi = -47^\circ$
- $\beta$ -parallelo:  $\Phi = -119^\circ$ ,  $\Psi = +113^\circ$
- $\beta$ -antiparallelo:  $\Phi = -139^\circ$ ,  $\Psi = +135^\circ$

In conformazioni diverse, c'è un limite al numero delle combinazioni possibili dei due angoli di rotazione, perché alcune hanno effetti destabilizzanti a causa delle forze di repulsione tra gli O. Vi sono quindi conformazioni più stabili di altre perché favorite da un punto di vista energetico. Conformazioni proibite sono anche quelle in cui si svilupperebbe una forte repulsione tra le catene laterali.

## Energetica del ripiegamento

La termodinamica caratterizza gli stati in natura dalla dipendenza da temperatura, pressione, volume e concentrazione chimica.

L'entropia è la quantità di disordine di un sistema, in altri termini è il numero di possibili stati configurazionali del sistema ( $\Omega$ ).

$$S = K_B \ln \Omega$$

Nei sistemi molecolari cambiamenti di entropia ( $\Delta S$ ) rappresentano cambiamenti nella libertà di movimento di atomi appartenenti sia al soluto che al solvente. L'entalpia è l'energia interna al sistema. Cambiamenti positivi di entalpia nelle macromolecole sono associate alla rottura di interazioni non covalenti favorevoli (si assume che i legami covalenti restino invariati). La formazione di legami covalenti diminuisce l'entalpia del sistema e rilascia calore verso l'ambiente. Formando legami si libera energia mentre per romperli si consuma energia.

$$\Delta H \simeq \Delta E$$

$$E = U + K$$

Dove  $H$  è l'entalpia,  $E$  l'energia interna al sistema,  $U$  l'energia potenziale (circa la somma di tutte le interazioni covalenti e non covalenti, ma non quelle non polari) e  $K$  l'energia cinetica (associata ai movimenti atomici indotti termicamente).

L'energia libera di Gibbs ( $G$ ) è l'*energia utile* sotto temperatura e pressione costante. I processi spontanei raggiungono l'equilibrio decrementando la  $G$  del sistema a un minimo.

L'energia libera di Gibbs è associata all'entropia e all'entalpia dalla seguente relazione:

$$\Delta G = \Delta H - T\Delta S$$

La seconda legge della termodinamica afferma che in un sistema isolato i processi spontanei raggiungono l'equilibrio incrementando l'entropia del sistema.

Il ripiegamento delle proteine è un processo spontaneo. Esibisce una grande varietà di percorsi, meccanismi e velocità che dipendono da parametri come la composizione della proteina e le condizioni di ripiegamento. A prescindere dall'esatto meccanismo, il processo di ripiegamento segue sempre la teoria del profilo energetico (*energy landscape theory*), cioè il ripiegamento è sempre accompagnato da una decrescita del numero di conformazioni che possono essere testate dalla proteina, sfuggendo al paradosso di Levinthal (vedi sez. 2.5).

Una proteina che si ripiega deve procedere da uno stato ad alta energia ed alta entropia a uno stato caratterizzato da bassi valori di energia ed entropia; tale nesso è conosciuto come *imbuto di ripiegamento* (folding funnel, vedi fig. 2.25). Le forze idrofobiche causano la creazione del nucleo idrofobico seppellendo i residui non polari nella struttura nativa, questo vuol dire che l'entropia del solvente acquoso subisce un aumento portando ad una sovraccompensazione dell'entropia, ovvero l'entropia del sistema aumenta.

Le proteine non cercano una conformazione fra tutte le possibili conformazioni fino a trovare quella giusta. Piuttosto si ripiegano in una maniera cooperativa in cui ogni step limita ulteriormente le possibilità di ripiegamento degli step seguenti. Il ripiegamento è completato quando la conformazione con più basso livello di energia libera associata alla proteina è trovata. La superficie rugosa del tunnel riflette il fatto che il processo di ripiegamento passi attraverso molti minimi locali separati da barriere da alta-energia.

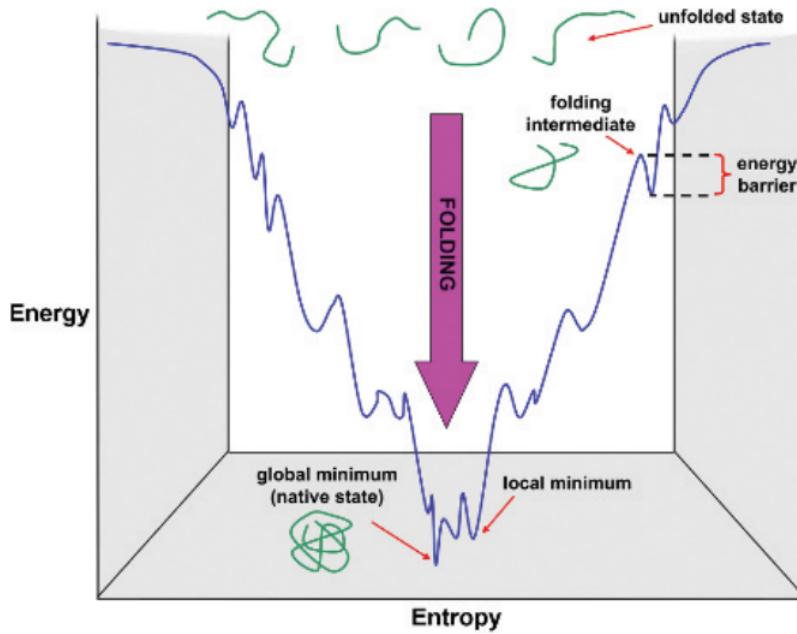


Figura 2.25: Profilo energetico a imbuto del ripiegamento. Fonte [3]

Non c'è un meccanismo di ripiegamento universale, ma una collezione di possibili meccanismi che possono essere usati. La preferenza di una proteina ad usarne uno piuttosto che un altro può dipendere da vari fattori, uno dei quali sembra essere la struttura secondaria. Ad esempio in proteine dominate da  $\alpha$ -eliche il ripiegamento è spesso gerarchico.

La velocità di ripiegamento correla non solo con la dimensione della proteina ma anche con la sua topologia nativa. Le proteine *fast-folding* (ripiegamento su scala temporale di nanosecondi) tendono ad avere grandi proporzioni di elementi secondari locali ( $\alpha$ -eliche e giri) laddove quelle *slow-folding* tendono ad avere proporzioni più grandi di elementi globali ( $\beta$ -foglietti).

I fattori termodinamici che danno luogo all'effetto idrofobico sono complessi e non del tutto conosciuti. L'effetto idrofobico è visto come una combinazione dell'effetto di idratazione (effetto entropico) e di interazioni di van der Waals tra molecole di soluto (effetto entalpico).

### 2.3.2 Ripiegamento assistito

All'interno delle cellule le proteine più piccole si ripiegano indipendentemente, mentre proteine più grandi sono assistite principalmente da complessi chiamati *chaperoni molecolari*. È importante notare che l'assistenza è cinetica in natura: non aggiunge nuove informazioni necessarie alla proteina per ripiegarsi, pertanto il dogma di Anfinsen non viene contraddetto. Ciò che fanno questi complessi è creare un ambiente nel quale le proteine possano ripiegarsi senza "distrazioni" dovute a interazioni con altre entità (ad esempio evitando l'aggregazione con altre proteine) e senza rimanere bloccate in conformazioni intermedie durante il loro percorso di ripiegamento. In poche parole sono misure di protezione della cellula.

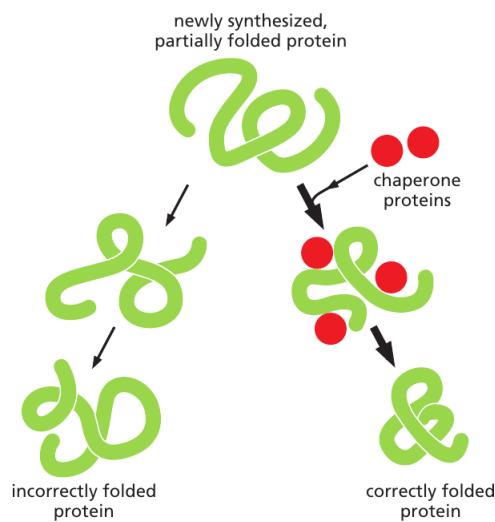
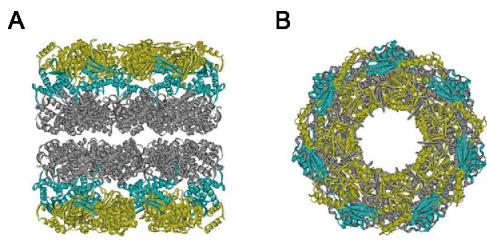


Figura 2.26: Schema della funzione dei chaperoni molecolari. Fonte: [4]

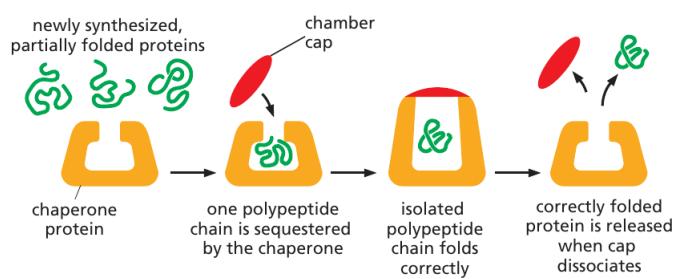
Più in dettaglio i chaperoni molecolari svolgono le seguenti funzioni:

1. assistono il corretto ripiegamento delle catene polipeptidiche (lunghe) appena sintetizzate
2. dirigono l'assemblaggio di complessi multienzimatici
3. donano una "seconda chance" a proteine danneggiate favorendone la rinaturazione
4. partecipano nella parziale denaturazione durante il trasporto di proteine attraverso membrane di mitocondri o cloroplasti

Tutti i compartimenti cellulari delle cellule eucariotiche (nucleo, citosol, reticolo endoplasmatico, mitocondri e cloroplasti) hanno il proprio set di chaperoni che assicura un corretto ripiegamento delle proteine. I chaperoni molecolari comprendono diverse famiglie di proteine altamente conservative, tra cui le Hsp (Heat shock protein), proteine espresse in grande quantità sotto condizioni di alto stress, per contrastarne l'effetto denaturante. Queste ultime sono state classificate in base al loro peso molecolare, ad es. Hsp60 dove "60" indica 60kDa. Le Hsp60 vengono chiamate anche *chaperonine* e sono una famiglia di chaperoni molecolari a doppio anello che agiscono da "camera di isolamento" per il ripiegamento di altre proteine<sup>[52]</sup>, famosa è la chaperonina procariotica GroEL (vedi fig. 2.27), che può essere assunta come modello di riferimento delle chaperonine.



*Figura 2.27: Strutture dei complessi GroEL e GroEL-GroES. (B) si può osservare la tipica forma ad anello.*  
Fonte: [53]



*Figura 2.28: Rappresentazione schematica della funzione della camera di isolamento nelle chaperonine. Fonte [4]*

Sebbene i mitocondri (e i cloroplasti) abbiano il loro genoma e creino le loro proteine, la maggior parte delle proteine che questi organelli usano sono codificate dai geni nel nucleo e importati dal citosol. Ogni proteina viene quindi parzialmente denaturata per effettuare il trasporto. I chaperoni molecolari all'interno di questi organelli aiutano a tirare le proteine attraverso le due membrane e a ripiegarle una volta all'interno<sup>[4]</sup>.

### 2.3.3 Misfolding, prioni e malattie

#### Misfolding

Il *misfolding* è il fenomeno dell'errato ripiegamento di una proteina, ovvero quando una proteina non può raggiungere il suo stato nativo. Ciò può accadere per mutazioni alla sua sequenza amminoacidica (anche per un solo amminoacido differente come nel caso dell'anemia falciforme) o per fattori esterni. Le proteine mal ripiegate tipicamente contengono  $\beta$ -foglietti organizzati in una struttura denominata cross- $\beta$ , disposizione molto stabile e insolubile, resistente alla proteolisi. Il mal ripiegamento di alcune proteine può

innescare ulteriori mal ripiegamenti e la conseguente accumulazione di proteine mal ripiegate in aggregati (od oligomeri) che possono guadagnare tossicità attraverso le interazioni intermolecolari. L'incremento dei livelli di proteine aggregate può portare alla formazione di *amiloidi*, strutture fibrillari formate da deposizioni di materiale proteico insolubile.

## Malattie

L'errato ripiegamento delle proteine è alla base quindi di molte patologie umane, definite malattie da misfolding, categorizzabili in due gruppi:

- malattia causata dalla perdita o degradazione della proteina o dall'errato trasporto intracellulare
- malattie causate dall'accumulo, intra od extra-cellulare, di proteine aggregate (ad esempio le malattie da prione)

Molti tipi di tumore diventano chemio-resistenti perché iper-esprimono alcune Hsp, come la Hsp70 e la Hsp90. Le Hsp sono presenti anche in quantità elevatissime nel cervello dei pazienti con malattia di Alzheimer e morbo di Parkinson. Tuttavia si crede che la loro aumentata espressione non sia lesiva di per sé ma rappresenti piuttosto una risposta difensiva agli elevati livelli di stress che caratterizzano queste patologie. Ci sono molti morbi associati a mutazioni nei geni codificanti i chaperoni. Alterazioni genetiche delle chaperonine possono portare a patologie umane che in genere colpiscono molti organi ed apparati contemporaneamente<sup>[54]</sup>.

## Prioni

I *prioni* (acronimo di "proteinaceous infective **only** particle") sono molecole di natura proteica con la capacità di trasmettere la propria forma mal ripiegata a varianti normali della stessa proteina<sup>8</sup>. Il ruolo ipotizzato di una proteina come agente infettivo è in contrasto con tutti gli altri agenti infettivi conosciuti, come i viroidi, virus, batteri, funghi, parassiti: tutti contengono acidi nucleici (DNA, RNA o entrambi) mentre le proteine sono composte di soli amminoacidi.

I prioni formano amiloidi che si accumulano nei tessuti e sono associati a danni di questi e alla morte cellulare. I prioni sono attualmente considerati i più probabili agenti delle encefalopatie spongiformi trasmissibili (TSE) dei mammiferi. Nel *morbo della mucca pazza* (encefalopatia spongiforme bovina), malattia neurologica degenerativa e irreversibile, vi

---

<sup>8</sup>I prioni sono stati studiati e denominati in questo modo dal premio Nobel per la medicina nel 1997 Stanley Prusiner<sup>[55]</sup>

è il ruolo di un prione a causare mal ripiegamenti di alcune proteine native causando la formazione di strutture amiloidi fatali (al microscopio le dense placche fibrose appaiono come buchi, da qui il caratteristico aspetto "a spugna"). Tutte le malattie da prione sono attualmente inguaribili e letali, con un periodo di incubazione che dura generalmente vari anni.

Gli aggregati di prioni sono stabili e questa stabilità strutturale consente loro di essere immuni alla maggior parte dei trattamenti conosciuti. L'organismo infettato non ha modo di degradarli: a differenza di virus e batteri i prioni rimangono intatti anche in presenza di trattamenti come sterilizzazione, forti dosi di radiazioni ionizzanti, uso di formaldeide, varechina, acqua bollente e a differenza delle altre proteine sono resistenti alla maggior parte delle proteasi.

La proteina di cui sono fatti i prioni, *PrP* (protease-resistant-protein, Pr per **prione**, e P per **proteina**), si trova in tutto il corpo, anche negli individui sani, ed è altamente conservata nei mammiferi. Tuttavia, la *PrP* trovata nel materiale infettante ha una struttura diversa. Nell'uomo la *PrP<sup>c</sup>*(cellulare, forma normale) è codificata da un solo gene, PRNP. La *PrP<sup>sc</sup>*(scrapie, forma patologica) differisce dalla proteina naturale *PrP<sup>c</sup>* per la conformazione tridimensionale: la *PrP<sup>c</sup>* ha una struttura più aperta contenente 3 segmenti ad  $\alpha$ -eliche e pochi  $\beta$ -foglietti; la *PrP<sup>sc</sup>* invece ha una struttura più compatta e stabile e presenta un aumento di  $\beta$ -foglietti.

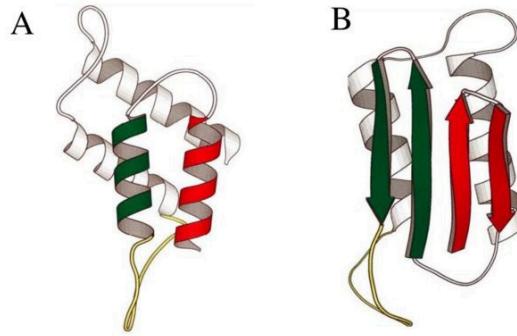


Figura 2.29: (A) Struttura della *PrP<sup>c</sup>*. (B) Struttura della *PrP<sup>sc</sup>*. Fonte: [56]

### 2.3.4 Controllo qualità e apoptosis

#### Controllo qualità

L'uscita delle proteine dal reticolo endoplasmatico (RE) è controllata per assicurare la qualità delle proteine. Sebbene alcune siano appositamente create e destinate a funzionare nel RE la maggior parte delle proteine che entrano nel RE sono destinate ad altri luoghi. Queste vengono impacchettate nelle vescicole di trasporto e gemmano per fondersi

con l'apparato del Golgi. Ma l'uscita dal RE è altamente selettiva: le proteine che falliscono a ripiegarsi nella forma nativa e quelle che non si assemblano correttamente sono attivamente conservative nel RE attraverso i legami con i chaperoni molecolari che risiedono lì. Questi trattengono le proteine nel RE finché non si verifica il corretto ripiegamento o assemblaggio. Se questo non si verifica o fallisce ancora le proteine sono esportate nel citosol dove sono degradate da un *proteasoma*. Le proteine da degradare sono contraddistinte dal loro legame con l'ubiquitina<sup>9</sup>.

Ad esempio gli anticorpi sono composti da 4 catene polipeptidiche che si assemblano in completi anticorpi nel RE. Gli anticorpi parzialmente assemblati sono conservati nel RE finché tutte e 4 le catene non sono pronte. Le molecole di anticorpi che falliscono ad assemblarsi vengono degradate.

Nonostante l'indubbia utilità di questo meccanismo di controllo a volte questo può rivelarsi dannoso per l'organismo. Ad esempio la mutazione predominante che causa la *fibrosi cistica*, comune malattia genetica che comporta seri danni polmonari, produce una proteina di trasporto della membrana plasmatica leggermente mal ripiegata. Tuttavia questa potrebbe funzionare normalmente se raggiungesse la membrana plasmatica ma viene bloccata nel RE e successivamente degradata<sup>[4]</sup> (per usare una metafora si può immaginare la situazione di un condannato alla pena di morte innocente). Le conseguenze sono terribili. La nota da sottolineare è che in questa malattia la mutazione non causa un'inattivazione di una proteina importante ma la proteina attiva è scartata dalle cellule prima che questa possa avere l'opportunità di funzionare.

## Proteasomi

I proteasomi sono complessi di *proteasi* (enzima in grado di catalizzare la rottura del legame peptidico delle proteine) che degradano le proteine mal ripiegate attraverso reazioni di *proteolisi*. Sono presenti nelle cellule di tutti gli eucarioti e procarioti. La struttura e la funzione di questi complessi è altamente conservata.

A causa del ruolo dei proteasomi nella regolazione del ciclo cellulare e dell'apoptosi<sup>10</sup>,

---

<sup>9</sup>Per "la scoperta della degradazione delle proteine mediata da ubiquitina" è stato assegnato il Premio Nobel per la chimica del 2004

<sup>10</sup>Il termine *apoptosi* indica una forma di morte cellulare programmata (un'auto-distruzione). Al contrario della necrosi, che è una forma di morte cellulare risultante da un acuto stress o trauma cellulare, l'apoptosi è portata avanti in modo ordinato e regolato, richiede consumo di energia (ATP) e generalmente porta a un vantaggio durante il ciclo vitale dell'organismo (è infatti chiamata da alcuni morte altruista o morte pulita). Durante il suo sviluppo, ad esempio, l'embrione umano presenta gli abbozzi di mani e piedi "palmati": affinché le dita si differenzino, è necessario che le cellule che costituiscono le membrane interdigitali muoiano

sono oggi un bersaglio rilevante nelle terapie antitumorali. Farmaci inibitori nella terapia antiretrovirale interferiscono con il ciclo replicativo del virus HIV proprio bloccando l'attività dell'enzima della proteasi.

### **Unfolded protein response e Apoptosi**

La dimensione del RE è controllato dalla "richiesta" per il ripiegamento delle proteine. Il meccanismo di controllo nel RE, eseguito dai chaperoni molecolari, può essere sopraffatto. Quando succede le proteine mal ripiegate si accumulano nel RE. Se l'accumulo è abbastanza grande, questo innasca un complesso programma chiamato *unfolded protein response* (UPR). Questo programma incita la cellula a produrre più RE, inclusi più chaperoni molecolari, e altre proteine riguardanti il controllo qualità. L'UPR permette alla cellula di regolare la dimensione del RE per gestire propriamente il volume delle proteine in entrata. In alcuni casi tuttavia anche un RE espanso non riesce a gestire la richiesta e l'UPR indirizza la cellula verso l'*apoptosi*.

Una situazione del genere può avvenire negli adulti in cui insorge il diabete. I tessuti diventano gradualmente resistenti all'effetto dell'insulina. Per compensare questa resistenza le cellule che secernono insulina nel pancreas ne producono ancora di più. Si arriva infine alla situazione in cui il loro RE arriva ad una capacità massima e viene innescato l'UPR e di conseguenza la morte cellulare. Col tempo sempre più cellule secernenti insulina sono eliminate e la richiesta per quelle sopravvissute aumenta rendendole sempre più vulnerabili a questo meccanismo, esacerbando ulteriormente la malattia<sup>[4]</sup>.

## **2.4 Sfide al dogma di Anfinsen: IDP e fold switching**

Il ripiegamento delle proteine in una cellula è un processo molto complesso che riguarda il trasporto di nuove proteine sintetizzate ad appropriati compartimenti cellulari attraverso targeting, misfolding, stati dispiegati temporanei, modifiche post-traduzione, controllo qualità, aggregazione in complessi, facilitazione dei chaperoni molecolari. Come già spiegato, l'aiuto dei chaperoni molecolari non sfida il dogma di Anfinsen, in quanto non influenza la struttura nativa della proteina.

La struttura di alcune proteine è difficile da determinare per una semplice ragione: un crescente numero di ricerche biochimiche ha rivelato che un numero significativo di proteine, o regioni di proteine, non hanno una distinta struttura 3D, non la hanno finché non interagiscono con la molecola target oppure cambiano struttura nativa. La loro flessibilità e struttura indefinita è importante per la loro funzione, che potrebbe richiedere legami con differenti target in tempi diversi. Si stima che queste proteine possano contribuire per il 20-30% al proteoma dei mammiferi.

## Intrinsically disordered proteins

Una proteina intrinsecamente disordinata (IDP) è una proteina, o una regione di essa, a cui manca una struttura terziaria fissa od ordinata. Le IDP sono comunemente riconosciute come regioni mancanti di densità elettronica in strutture di proteine determinate a cristallografia a raggi X. Molte IDP possono adottare una struttura tridimensionale stabile dopo essersi legate ad altre macromolecole, passando per transizioni disordine-ordine e perciò risultare strutturate per un certo periodo di tempo e non strutturate per altro. Ci sono però anche IDP che svolgono la loro funzione senza assumere mai una forma ordinata attraverso la loro esistenza. Nonostante la loro mancanza di struttura stabile le IDP risultano essere una classe importante e grande di proteine.

## Fold switching proteins

Alcune proteine hanno multiple strutture native: possono cambiare la loro forma, rimodelando anche le loro strutture secondarie, in base a fattori esterni. Ad esempio il complesso proteico KaiB cambia ripiegamento durante la giornata agendo da orologio per i cianobatteri. Si possono immaginare le proteine *fold switching* come una sorta di transformer<sup>11</sup> dove in un caso la proteina è come un robot che fa una cosa e in un altro caso, in risposta a cambiamenti ambientali, diventa un'automobile e fa qualcos'altro. Il cambio tra strutture alternative è guidato da interazioni della proteina con piccoli ligandi o altre proteine, da modificazioni chimiche (es. fosforilazione) o da cambiamenti nelle condizioni ambientali (temperatura, pH, potenziale di membrana). Ogni struttura alternativa può o corrispondere al minimo globale di energia libera della proteina in certe condizioni o essere cinematicamente intrappolata in un alto minimo locale di energia libera<sup>[59]</sup>.

Le proteine *fold switching* (FS) differiscono dalle IDP<sup>[58]</sup>:

- le FS richiedono che entrambe le loro conformazioni siano determinate, le IDP sono regioni non determinate
- le IDP sono caratterizzate da sequenze aminoacidiche caratteristiche mentre le FS sono libere da questo vincolo
- le IDP non si ripiegano cooperativamente in isolamento mentre le FS si ripiegano sia cooperativamente che indipendentemente

Per queste ragioni si può affermare che le FS non sono IDP, piuttosto sono un sottinsieme delle proteine globulari le cui strutture stabili cambiano drasticamente in risposta al loro ambiente. I vantaggi di una proteina FS sono legati alla sua bifunzionalità:

---

<sup>11</sup>metafora di Lauren Porter<sup>[57]</sup>, autrice di<sup>[58]</sup>

- può affrontare velocemente richieste biologiche ovviando al bisogno di risorse cellulari aggiuntive per trascrivere e tradurre un'altra proteina. Un esempio è RfaH: funziona sia da fattore di trascrizione che di traduzione
- regolazione di inattività, può essere bloccata in uno stato di attività o inattività finché non viene innescato uno specifico segnale

È stato stimato che una percentuale tra lo 0.5 e il 4% delle proteine nel PDB cambi ripiegamento<sup>[58]</sup>.

#### 2.4.1 Considerazioni epistemologiche

La scoperta delle IDP e delle proteine FS ha creato una spaccatura nel paradigma della struttura rigida delle proteine, secondo il quale la struttura deve essere fissa al fine di compiere la propria funzione biologica. È interessante ripercorrere velocemente la storia di questo paradigma attraverso un breve excursus storico per poterne mettere in luce le relazioni epistemologiche soggiacenti.

Nel 1894 Fischer ha proposto una metafora di *chiave e serratura* per spiegare come fosse possibile un effetto chimico tra un enzima e un glucoside:

«*Per usare una metafora, vorrei dire che l'enzima e il glucoside devono adattarsi l'uno all'altro come una serratura e una chiave per esercitare un effetto chimico l'uno sull'altro*<sup>[60][61]</sup>»

Nel 1936 Mirsky e Pauling hanno raccolto una quantità di informazioni sufficiente per concludere:

«*attribuiamo le specifiche proprietà caratteristiche delle proteine native alle loro uniche e definite configurazioni. Consideriamo le proteine denaturate essere caratterizzate dall'assenza di un'unica definita configurazione*<sup>[62]</sup>».

Né Mirsky, né Pauling né Hsien Wu (probabilmente il primo a proporre il paradigma struttura unica-funzione) citarono il lavoro di Fischer ma la sua metafora avrebbe supportato pienamente le loro tesi. Perciò ancora prima dell'esperimento di Anfinsen e delle risoluzioni atomiche delle strutture, una specifica e ben ordinata forma tridimensionale era stata accettata come prerequisito essenziale per la funzione di una proteina.<sup>[61]</sup>. Le prime strutture proteiche sono state determinate attraverso la cristallografia negli anni '50, dando ancora più credito all'ipotesi che una struttura fissa fosse necessaria per adempiere la funzione biologica. Queste pubblicazioni hanno contribuito a solidificare il dogma centrale della biologia molecolare. Nonostante questo già nel 1950 si parlava di molteplicità di strutture per una proteina, in particolare Karush<sup>[63]</sup> sull'albumina del siero bovino, inferendo che le interazioni proteina-ligando stabilizzassero il membro più adatto da un insieme di strutture in equilibrio, chiamando questo fenomeno *adattabilità configu-*

*razionale*. Successivamente si può ricordare il paradosso di Levinthal negli anni '60, per arrivare negli anni '70 al dogma di Anfinsen, il cui paradigma (sequenza amminoacidica -> struttura tridimensionale -> funzione) è messo in discussione da evidenze di mancanza di generalità. Tuttavia quel sistema di pensiero ha preso piede ed ha pervaso quasi tutti i lavori e teorie successive. Al tempo dell'esperimento di Anfinsen e delle prime risoluzioni atomiche della mioglobina e del lisozima il prerequisito di una forma tridimensionale fissa necessaria per la funzione della proteina era già accettato. La successiva valanga di migliaia di strutture determinate sperimentalmente ha contribuito ad affossare paradigmi di pensiero alternativi.

La scoperta di segmenti intrinsecamente disordinati è stata compiuta nel 1978, quando ancora erano disponibili le strutture di sole 20 proteine. Alcuni segmenti di proteine non fornivano alcuna densità elettronica distinguibile nonostante fossero essenziali per il funzionamento. Una ragione comune è che gli atomi di quel segmento sono disordinati, ovvero la loro posizione cambia. Con l'avvento del NMR, sempre nel 1978, si è arrivati a identificare intere proteine disordinate e dato che questo metodo è più preciso, la riscoperta di disordine nativo ha avuto un impatto significativo.

Fino al 2000<sup>[64]</sup> queste idee non sono apparse nei libri di biochimica, nonostante in 50 anni fossero state pubblicate centinaia di paper sull'importanza della flessibilità e del disordine nelle strutture proteiche.

La parola proteina potrebbe non essere più la parola giusta per identificare il complesso mondo delle macromolecole più importanti della vita<sup>[65]</sup>.

Può risultare interessante confrontare quanto accaduto con le analisi del microbiologo ed epistemologo Ludwik Fleck<sup>[66]</sup>. Nel suo libro del 1935, *Genesi e sviluppo di un fatto scientifico*, affronta il problema della conoscenza scientifica e analizza il caso dell'evoluzione del concetto della malattia sifilide, mostrando come questo si è modificato nel tempo. Senza entrare nel merito di quell'analisi, l'epistemologo ha provato a tracciare delle linee generali su come un fatto scientifico possa svilupparsi.

Analizzando le epoche di un concetto, Fleck si esprime affermando:  
«*Molte teorie, per esempio, hanno due epoche nella loro vita: esse attraversano prima un'epoca classica, in cui tutto si accorda in maniera impressionante, poi una seconda epoca, nel corso della quale si presentano solo delle eccezioni.*»

Non è difficile trovare le somiglianze con il caso della struttura unica per la funzione di una proteina. Inizialmente le eccezioni sono passate inosservate, tutto si accordava perfettamente all'idea di Fischer, Pauling e poi di Anfinsen. Le eccezioni, come quella di Karush, si presentavano ma il paradigma dominante non ne subiva effetti.

Secondo Fleck, per garantire la persistenza dei sistemi d'opinione quando si presentano

eccezioni:

«una contraddizione al sistema appare impensabile. Ciò che non si accorda con il sistema: non viene notato, oppure viene tacito anche se noto, oppure si fa in modo di spiegarlo, con laboriosi sforzi, come non contraddittorio rispetto al sistema: si notano, si descrivono o persino si inventano fatti che corrispondono alla concezione dominante, che cioè ne costituiscono per così dire la realizzazione»

Uno dei concetti più importanti nell'opera di Fleck è il concetto di *collettivo di pensiero e stile di pensiero*:

«Se definiamo il termine collettivo di pensiero come "la comunità degli uomini che hanno fra loro un contatto intellettuale e che si scambiano idee influenzandosi reciprocamente, noi veniamo in possesso, con questo concetto, di ciò che rappresenta lo sviluppo storico di un ambito del pensiero, di un determinato patrimonio di conoscenza e di cultura e quindi, di un determinato stile di pensiero.»

E secondo il microbiologo la condizione dell'individuo (scienziato) nei confronti dello stile di pensiero è di subordinazione inconsapevole:

«Anche se il collettivo consiste di individui, esso non è la loro semplice somma. L'individuo non ha mai - o quasi mai - la coscienza dello stile di pensiero collettivo, che quasi sempre esercita una costrizione incondizionata sul suo pensiero e che è semplicemente impensabile poter contraddirlo.»

«Viene a anche a mettersi in luce che molte idee arrivano a manifestarsi prima che ne risaltino le basi razionali e, anzi, in modo completamente indipendente da queste ultime.» L'idea di Fischer (poi di Pauling) potrebbe avere influenzato il collettivo di pensiero al punto tale che la determinazione delle prime strutture a risoluzione atomica non potesse che dare come risposta una conferma di quelle idee. Esperimento ed esperienza non vivono entrambe nel campo dell'oggettività:

«se l'esperimento può essere interpretato come una pura e semplice domanda e risposta, l'esperienza deve essere invece intesa già come una condizione complessa, frutto di un processo di educazione che si fonda sull'interazione fra chi conosce, ciò che è conosciuto e ciò che deve essere conosciuto.»

La mancata consapevolezza di far parte di un collettivo di pensiero può purtroppo causare l'illusione che esista un nesso logico fra prove e concezioni, ma Fleck ammonisce: «le prove si adattano alle concezioni altrettanto spesso quanto le concezioni si conformano alle prove»

Secondo Fleck la conoscenza è sempre un processo sociale:

«"questo libro è più voluminoso" è incompleta. Sarebbe corretta se si aggiungesse "di quel

*libro*"; [...] la frase «qualcuno conosce qualcosa» richiede un'aggiunta, ad es. "sulla base di un determinante patrimonio di conoscenza" o meglio "come membro di un determinato ambiente culturale" o ancora "in un determinato collettivo di pensiero"»

Mettendo insieme le diverse constatazioni riportate si può fare un'analisi, senza alcuna presunzione di correttezza e come semplice esercizio, del dogma di Anfinsen. Secondo Fleck una formulazione corretta sulle sue scoperte potrebbe essere: "Anfinsen propose, in conformità con le opinioni del suo tempo sul ripiegamento, denaturazione e rinaturazione delle proteine, di vedere nella sequenza amminoacidica la base necessaria e sufficiente per la determinazione della sua struttura tridimensionale nativa. Propone inoltre, sempre sulla base del collettivo di pensiero in cui era immerso, di considerare la struttura nativa di una proteina come quella struttura unica, stabile e cinematicamente accessibile avente minima energia libera". Da questo esercizio si può osservare, essendo noi immersi in un collettivo di pensiero differente, che lo stile di pensiero di Anfinsen era probabilmente ancorato alle idee di Pauling secondo il quale la conformazione stabile delle proteine era una e una soltanto.

## 2.5 Il problema del Protein Folding

Il problema del protein folding è la questione di *come* una sequenza amminoacidica determini la struttura atomica tridimensionale. Il processo del ripiegamento proteico non è così semplice, la maggior parte delle proteine probabilmente passa attraverso strutture intermedie sulla via per raggiungere la struttura nativa, e il semplice osservare la struttura finale non rivela i passaggi del ripiegamento richiesti per raggiungere quella forma. Il problema del protein folding consiste di 3 puzzle strettamente correlati<sup>[44]</sup>:

- *folding code*: la questione termodinamica di quale bilancio delle forze interatomiche determini la struttura della proteina a partire da una data sequenza amminoacidica
- *folding process*: la questione cinematica di quali percorsi alcune proteine usino per ripiegarsi così velocemente
- *protein structure prediction*: si può predire la struttura nativa di una proteina dalla sua sequenza amminoacidica? In altre parole il problema computazionale di come predire la struttura nativa di una proteina dalla sua sequenza amminoacidica

Il problema del protein folding, come si può immaginare, è considerato uno dei problemi più impegnativi degli ultimi 50 anni in biochimica. Un aspetto importante del problema è sottolineato dal *paradosso di Levinthal*. Nel 1968 Cyrus Levinthal si rese conto che, a causa dell'elevato numero di gradi di libertà di un polipeptide non ripiegato, tale molecola presenterebbe un numero astronomico di possibili conformazioni finali. Se la

proteina raggiungesse la sua conformazione finale passando via via attraverso tutte queste configurazioni, sarebbe necessario un tempo ben superiore all'età attualmente stimata dell'universo per raggiungere la configurazione corretta, anche se ogni passaggio richiedesse *picosecondi*. In natura però molte piccole proteine si ripiegano spontaneamente in un tempo dell'ordine dei millisecondi o addirittura dei microsecondi. Il tempo di generazione di *E. coli* può essere di circa venti minuti: ciò significa che tutte le proteine essenziali per tale organismo (e presumibilmente di tutti gli altri) possono essere prodotte da zero in un tempo decisamente ristretto, al massimo nell'ordine dei minuti.

Il processo di ripiegamento non è quindi una ricerca all'interno dell'enorme spazio degli stati configurazionali possibili. La differenza enorme che esiste tra il tempo del ripiegamento prevedibile in teoria e quello osservato in realtà è appunto chiamato paradosso di Levinthal.

Come accennato nella sezione 2.3, la malattia di Alzheimer, la fibrosi cistica e altre malattie neurodegenerative sono associate al mal ripiegamento delle proteine. La conoscenza dei fattori di mal ripiegamento e la comprensione del processo di ripiegamento proteico potrebbero aiutare nello sviluppo di cure per queste malattie. Per queste ragioni è importante anche rispondere alle altre domande del problema e non fermarsi alla predizione della struttura finale, nonostante questa conoscenza fornisca un grande vantaggio per lo sviluppo di nuovi farmaci e il design di nuove proteine.

# Bibliografia

## Libri

- [3] A. Kessel e N. Ben-Tal, *Introduction to proteins: Structure, function and motion*, 2<sup>a</sup> ed. Chapman e Hall/CRC, 2018.
- [4] B. Alberts, D. Bray, K. Hopkin et al., *Essential cell biology*, 5<sup>a</sup> ed. W. W. Norton e Company, 2019.
- [5] N. A. Campbell, J. B. Reece, L. A. Urry, R. Brizzi, T. Niccolò e A. Bartalesi, *Biologia E Genetica*. Pearson, 2012.
- [26] A. D. Baxevanis, G. D. Bader e D. S. Wishart, *Bioinformatics*. John Wiley & Sons, 2020.
- [34] T. Mitchell, *Machine learning*. McGraw hill New York, 1997.
- [40] S. Pal, *Fundamentals of Molecular Structural Biology*. Academic Press, 2019.
- [45] L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, R. B. Orr e N. A. Campbell, *Campbell Biology*. Pearson, 2021.
- [46] L. A. Moran, H. R. Horton, K. G. Scrimgeour, M. D. Perry e D. Rawn, *Principles of biochemistry*. Pearson London, 2012.
- [50] D. L. Nelson, A. L. Lehninger e M. M. Cox, *Lehninger principles of biochemistry*. Macmillan, 2017.
- [66] L. Fleck, *Genesi e Sviluppo di un Fatto Scientifico: Per Una Teoria dello stile e del collettivo di pensiero*. Il mulino, 1983.

## Articoli

- [15] O. Hidalgo, J. Pellicer, M. Christenhusz, H. Schneider, A. R. Leitch e I. J. Leitch, “Is there an upper limit to genome size?” *Trends in Plant Science*, vol. 22, n. 7, pp. 567–573, 2017.

- [24] M. Batool, B. Ahmad e S. Choi, “A structure-based drug discovery paradigm,” *International journal of molecular sciences*, vol. 20, n. 11, p. 2783, 2019.
- [25] B. C. Knott, E. Erickson, M. D. Allen et al., “Characterization and engineering of a two-enzyme system for plastics depolymerization,” *Proceedings of the National Academy of Sciences*, vol. 117, n. 41, pp. 25 476–25 485, 2020.
- [27] F. C. Bernstein, T. F. Koetzle, G. J. Williams et al., “The protein data bank: A computer-based archival file for macromolecular structures,” *Journal of Molecular Biology*, vol. 112, n. 3, pp. 535–542, 1977.
- [28] M. Mitchell, “Biological Computation,” *PDXScholar*, 2010. indirizzo: [https://pdxscholar.library.pdx.edu/compsci\\_fac/2](https://pdxscholar.library.pdx.edu/compsci_fac/2).
- [38] H. Wu e E. Yang, “Studies on denaturation of proteins. XL Effect of hydrogen ion concentration on rate of denaturation of egg albumin by urea. A theory of denaturation,” *Chin J Physiol*, vol. 5, pp. 301–344, 1931.
- [39] C. B. Anfinsen, “The formation and stabilization of protein structure.,” *Biochemical Journal*, vol. 128, n. 4, p. 737, 1972.
- [41] C. B. Anfinsen, E. Haber, M. Sela e F. White Jr, “The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 47, n. 9, p. 1309, 1961.
- [44] K. A. Dill, S. B. Ozkan, M. S. Shell e T. R. Weikl, “The protein folding problem,” *Annu. Rev. Biophys.*, vol. 37, pp. 289–316, 2008.
- [47] J. Murray, N. Laurieri e R. Delgoda, “Chapter 24 - Proteins,” S. Badal e R. Delgoda, cur., pp. 477–494, 2017. doi: <https://doi.org/10.1016/B978-0-12-802104-0.00024-X>.
- [52] N. A. Ranson, H. E. White e H. R. Saibil, “Chaperonins,” *Biochemical Journal*, vol. 333, n. 2, pp. 233–242, 1998.
- [53] R. Iizuka e T. Funatsu, “Chaperonin GroEL uses asymmetric and symmetric reaction cycles in response to the concentration of non-native substrate proteins,” *Biophysics and Physicobiology*, vol. 13, pp. 63–69, 2016.
- [55] S. B. Prusiner, M. R. Scott, S. J. DeArmond e F. E. Cohen, “Prion protein biology,” *cell*, vol. 93, n. 3, pp. 337–348, 1998.
- [56] B. Ruttkay-Nedecky, E. Sedlackova, D. Chudobova et al., “Prion protein and its interactions with metal ions ( $\text{Cu}^{2+}$ ,  $\text{Zn}^{2+}$ , and  $\text{Cd}^{2+}$ ) and metallothionein 3,” *ADMET and DMPK*, vol. 3, n. 3, pp. 287–295, 2015.

- [58] L. L. Porter e L. L. Looger, “Extant fold-switching proteins are widespread,” *Proceedings of the National Academy of Sciences*, vol. 115, n. 23, pp. 5968–5973, 2018.
- [59] A. E. Varela, K. A. England e S. Cavagnero, “Kinetic trapping in protein folding,” *Protein Engineering, Design and Selection*, vol. 32, n. 2, pp. 103–108, 2019.
- [60] E. Fischer, “Einfluss der Configuration auf die Wirkung der Enzyme,” *Berichte der deutschen chemischen Gesellschaft*, vol. 27, n. 3, pp. 2985–2993, 1894.
- [61] A. K. Dunker, J. D. Lawson, C. J. Brown et al., “Intrinsically disordered protein,” *Journal of molecular graphics and modelling*, vol. 19, n. 1, pp. 26–59, 2001.
- [62] A. E. Mirsky e L. Pauling, “On the structure of native, denatured, and coagulated proteins,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 22, n. 7, p. 439, 1936.
- [63] F. Karush, “Heterogeneity of the binding sites of bovine serum albumin1,” *Journal of the American Chemical Society*, vol. 72, n. 6, pp. 2705–2713, 1950.
- [64] C. Bracken, M. M. Young e K. Dunker, “Disorder and flexibility in protein structure and function,” pp. 64–66, 2000.
- [65] G. Parisi, N. Palopoli, S. C. Tosatto, M. S. Fornasari e P. Tompa, ““Protein” no longer means what it used to,” *Current Research in Structural Biology*, vol. 3, pp. 146–152, 2021.

## Risorse Online

- [1] “enzima nell’Enciclopedia Treccani.” (13 gen. 2022), indirizzo: <https://www.treccani.it/enciclopedia/enzima> (visitato il 21/01/2022).
- [2] “proteina in Vocabolario - Treccani.” (13 gen. 2022), indirizzo: <https://www.treccani.it/vocabolario/proteina> (visitato il 22/01/2022).
- [6] “Chemical element - Wikipedia.” (1 gen. 2022), indirizzo: [https://en.wikipedia.org/wiki/Chemical\\_element](https://en.wikipedia.org/wiki/Chemical_element) (visitato il 31/01/2022).
- [7] “eukaryote. Definition, Structure, Facts.” (19 set. 2019), indirizzo: <https://www.britannica.com/science/eukaryote> (visitato il 22/01/2022).
- [8] “Neurone - Wikipedia.” (27 dic. 2021), indirizzo: <https://it.wikipedia.org/wiki/Neurone> (visitato il 23/01/2022).
- [9] “Saccharomyces cerevisiae - Wikipedia.” (25 set. 2021), indirizzo: [https://it.wikipedia.org/wiki/Saccharomyces\\_cerevisiae](https://it.wikipedia.org/wiki/Saccharomyces_cerevisiae) (visitato il 22/01/2022).

- [10] “Dogma centrale della biologia molecolare - Wikipedia.” (16 set. 2021), indirizzo: [https://it.wikipedia.org/wiki/Dogma\\_centrale\\_della\\_biologia\\_molecolare](https://it.wikipedia.org/wiki/Dogma_centrale_della_biologia_molecolare) (visitato il 22/01/2022).
- [11] “DNA Structure. BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/26-structure-of-dna-and-rna/dna-structure.html> (visitato il 22/01/2022).
- [12] S. Bewick, R. Parsons, T. Forsythe, S. Robinson e J. Dupon. “Introductory Chemistry (CK-12).” (1 giu. 2021), indirizzo: [https://chem.libretexts.org/Bookshelves/Introductory\\_Chemistry/Book%3A\\_Introductory\\_Chemistry\\_\(CK-12\)](https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Book%3A_Introductory_Chemistry_(CK-12)) (visitato il 22/01/2022).
- [13] “File: Difference DNA RNA-EN.svg - Wikimedia Commons.” (23 mar. 2010), indirizzo: [https://commons.wikimedia.org/wiki/File:Difference\\_DNA\\_RNA-EN.svg](https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-EN.svg) (visitato il 22/01/2022).
- [14] “Nasuia deltocephalinicola - Wikipedia.” (25 dic. 2021), indirizzo: [https://en.wikipedia.org/wiki/Nasuia\\_deltocophilinicola](https://en.wikipedia.org/wiki/Nasuia_deltocophilinicola) (visitato il 31/01/2022).
- [16] “Genome Size. BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/genome-size.html> (visitato il 31/01/2022).
- [17] “Paris japonica - Wikipedia.” (31 dic. 2021), indirizzo: [https://en.wikipedia.org/wiki/Paris\\_japonica](https://en.wikipedia.org/wiki/Paris_japonica) (visitato il 31/01/2022).
- [18] “Transfer RNA - Wikipedia.” (23 gen. 2022), indirizzo: [https://en.wikipedia.org/wiki/Transfer\\_RNA](https://en.wikipedia.org/wiki/Transfer_RNA) (visitato il 23/01/2022).
- [19] “Protein - Wikipedia.” (21 dic. 2021), indirizzo: <https://en.wikipedia.org/wiki/Protein> (visitato il 23/01/2022).
- [20] “Peptide bond - Wikipedia.” (4 nov. 2021), indirizzo: [https://en.wikipedia.org/wiki/Peptide\\_bond](https://en.wikipedia.org/wiki/Peptide_bond) (visitato il 23/01/2022).
- [21] “Amino Acids. BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/24-proteins/amino-acids.html> (visitato il 23/01/2022).
- [22] “PDB101: Learn: Videos: What is a Protein?” (20 Nov. 2017), indirizzo: <https://www.pdb101.rcsb.org/learn/videos/what-is-a-protein-video> (visitato il 23/01/2022).
- [23] K. Dill. “The protein folding problem: a major conundrum of science: Ken Dill at TEDxSBU.” (23 ott. 2013), indirizzo: <https://www.youtube.com/watch?v=zmc3kovWpNQ> (visitato il 06/01/2022).

- [29] “Apprendimento automatico - Wikipedia.” (1 dic. 2021), indirizzo: [https://it.wikipedia.org/wiki/Apprendimento\\_automatico](https://it.wikipedia.org/wiki/Apprendimento_automatico) (visitato il 23/01/2022).
- [30] “What is soft computing - Javatpoint.” (3 lug. 2021), indirizzo: <https://www.javatpoint.com/what-is-soft-computing> (visitato il 24/01/2022).
- [33] “Machine Learning - IBM.” (29 ago. 2020), indirizzo: <https://www.ibm.com/it-it/analytics/machine-learning> (visitato il 23/01/2022).
- [35] “What are Neural Networks?” (1 Giu. 2021), indirizzo: <https://www.ibm.com/cloud/learn/neural-networks> (visitato il 24/01/2022).
- [37] “Protein Structure .BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/24-proteins/protein-structure.html> (visitato il 27/01/2022).
- [42] “Christian B. Anfinsen - Digital Collections - National Library of Medicine.” (1 gen. 2022), indirizzo: <http://resource.nlm.nih.gov/101408166> (visitato il 25/01/2022).
- [43] “File:RibonucleaseA SS paleRib.png - Wikimedia Commons.” (16 mar. 2012), indirizzo: [https://commons.wikimedia.org/wiki/File:RibonucleaseA\\_SS\\_paleRib.png](https://commons.wikimedia.org/wiki/File:RibonucleaseA_SS_paleRib.png) (visitato il 25/01/2022).
- [48] “Protein structure prediction - Wikipedia.” (30 dic. 2021), indirizzo: [https://en.wikipedia.org/wiki/Protein\\_structure\\_prediction](https://en.wikipedia.org/wiki/Protein_structure_prediction) (visitato il 27/01/2022).
- [49] L. A. Moran. “Levels of Protein Structure.” (13 mar. 2008), indirizzo: <https://sandwalk.blogspot.com/2008/03/levels-of-protein-structure.html> (visitato il 27/01/2022).
- [51] “File:Ramachandran’s Diagram.jpg - Wikipedia.” (21 giu. 2016), indirizzo: [https://it.wikipedia.org/wiki/File:Ramachandran%27s\\_Diagram.jpg](https://it.wikipedia.org/wiki/File:Ramachandran%27s_Diagram.jpg) (visitato il 28/01/2022).
- [54] “Chaperonina - Wikipedia.” (21 ott. 2021), indirizzo: <https://it.wikipedia.org/wiki/Chaperonina> (visitato il 26/01/2022).
- [57] “Lauren Porter and Fold-Switching Proteins.” (14 feb. 2020), indirizzo: <https://www.youtube.com/watch?v=IeX5ebadgiA> (visitato il 28/01/2022).
- [67] “Latest Release Information.” (27 gen. 2022), indirizzo: <https://www.ddbj.nig.ac.jp/latest-releases-e.html> (visitato il 27/01/2022).
- [68] “wwPDB: Deposition Statistics.” (25 gen. 2022), indirizzo: <https://www.wwpdb.org/stats/deposition> (visitato il 27/01/2022).

## Altre fonti

- [31] R. Kurzweil, R. Richter, R. Kurzweil e M. L. Schneider, *The age of intelligent machines*, 1990.
- [32] H. H. the XIV Dalai Lama, *The heart of the Buddha's path*, 2011.
- [36] A. Ruffini, *Fisiogenia, la biodinamica dello sviluppo ed i fondamentali problemi morfologici dell'embriologia generale*, 1925.