



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

1	Introduzione	3
2	Background	4
2.1	Background biologico	5
2.1.1	Organizzazione della vita: dagli atomi alle cellule	5
2.1.2	Concetti fondamentali in biologia	8
2.1.3	Dogma centrale della biologia	9
2.1.4	Dai geni alle proteine	11
2.1.5	Proteine: le macromolecole più importanti della vita	14
2.2	Background informatico	18
2.2.1	Bioinformatica	18
2.2.2	Soft computing	19
2.2.3	Intelligenza Artificiale	20
2.2.4	Machine Learning	21
2.2.5	Reti neurali artificiali (ANN)	22
3	Protein Folding	23
3.1	Postulato di Anfinsen	24
3.1.1	Esperimento di Anfinsen	25
3.1.2	Denaturazione	25
3.2	Struttura delle proteine	27
3.2.1	Legami e interazioni molecolari	27
3.2.2	Geometria dei legami	36
3.2.3	Energetica del ripiegamento	36
3.3	Ripiegamento assistito	36
3.3.1	Misfolding e malattie	38
3.3.2	Controllo qualità e proteasomi	39
3.3.3	Unfolded protein response	41
3.4	Eccezioni al postulato di Anfinsen	41

3.4.1	Intrinsically disordered proteins	42
3.4.2	Fold switching proteins	42
3.4.3	box: Filosofia della scienza	42
3.5	Il problema del Protein Folding	42
4	Predizione della struttura di proteine	44
	Bibliografia	45

Capitolo 1

Introduzione

Capitolo 2

Background

Cos'è la vita? Da dove viene? - Fino al 18° 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. Il cambiamento avvenne nel 19° secolo. Un'importante svolta fu il lavoro di Louis Pasteur che stabilì un collegamento fra processi vitali e reazioni chimiche: la conversione di zucchero in alcool (fermentazione) era un risultato della crescita 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»^[1]). Non si conosceva la loro natura chimica, si scoprì successivamente che tutti gli enzimi sono *proteine* (dal greco «primario», «che occupa la prima posizione»^[2]). Queste proteine agivano da catalizzatori: acceleravano le reazioni chimiche all'interno delle cellule e nei tessuti 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^[3]. 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'inconcorazione delle proteine a *macromolecole più importanti della vita* si può legare ad un'altra svolta nel pensiero scientifico avvenuta nella seconda metà del 20° secolo: la rivoluzione genetica. Le proteine sono ben più che "macchine molecolari": sono i prodotti primari dei geni, responsabili, fra altri, dell'espressione dell'informazione genetica. È sullo sfondo di questa rivoluzione che l'informatica si è inserita all'interno del mondo della biologia.

2.1 Background biologico

2.1.1 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 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^[4].

Prima di parlare di cellule è opportuno richiamare l'attenzione sulle strutture biologiche. L'organizzazione biologica si basa su una gerarchia di livelli strutturali¹, ognuno dei quali poggia su un gradino sottostante:



Tutta la materia è costituita da 94 elementi chimici in natura (tralasciando quelli non stabili). La materia organica 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: $1\text{Da} = 1.7 \times 10^{-24}\text{g}$. Un elettrone pesa 0.0005Da . Gli elettroni più esterni sono chiamati *elettroni di valenza* e determinano il comportamento chimico di un atomo.

Lo scheletro dei composti organici 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 un livello nella gerarchia strutturale 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 gusci protettivi, gli acidi nucleici permettono di codificare l'informazione genica e le proteine sono alla base delle funzioni vitali.

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.

¹Questa sezione di background biologico si basa in larga parte sui personali *Appunti del corso Elementi di Biologia e Neuroscienze, prof. Mario Pirchio, Unipi CdL Filosofia, 2021*, frequentato nell'a.a. 2020/21 come esame a libera scelta.

Le cellule hanno dimensioni che variano dai $2\mu\text{m}$ ai *centimetri* delle uova di rana, gallina o struzzo ai *metri* di neuroni con lunghi assoni:

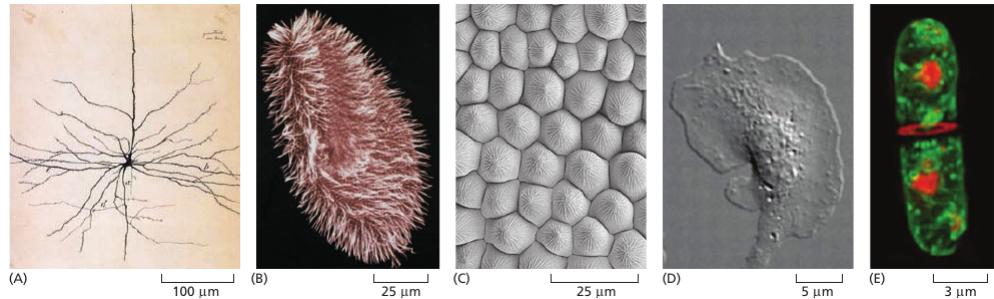


Figura 2.1: (A) disegno di un neurone. (B) Paramecium. (C) superficie di un petalo di fiore di bocca di leone. (D) Macrofago. (E) Un lievito di fissione viene catturato nell'atto di divisione cellulare. Fonte: [4]

È possibile dividere gli esseri viventi in due domini: *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 dividono i processi cellulari.

La strutture tipiche di una cellula animale e di un neurone sono mostrate nelle seguenti figure:

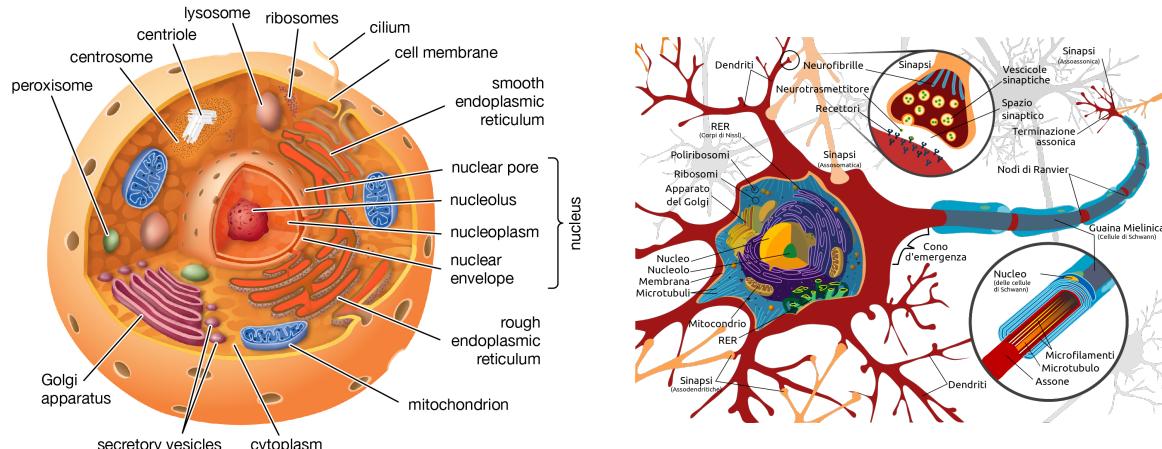


Figura 2.3: Neurone. Fonte [7]

Figura 2.2: Cellula animale. Fonte: [6]

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

cioè 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 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 come zuccheri e grassi 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 tali fibre si ispessiscono divenendo visibili come 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. Dall'involucro nucleare può uscire RNA e proteine ma non il DNA.

Il ciclo di vita delle cellule si basa su 4 fasi: crescita, sintesi del DNA, crescita completa e mitosi (divisione cellulare). Le cellule dei mammiferi impiegano da 18 a 24 ore per completare un ciclo di mitosi, mentre i lieviti solamente 90 minuti. Per questa ragione il lievito da fornaio (*Saccharomyces cerevisiae*) è usato come organismo modello in citologia e genetica: il suo genoma è stato il primo ad essere sequenziato completamente tra gli eucarioti^[8].

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 vengono ricambiati ogni 2 giorni.

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

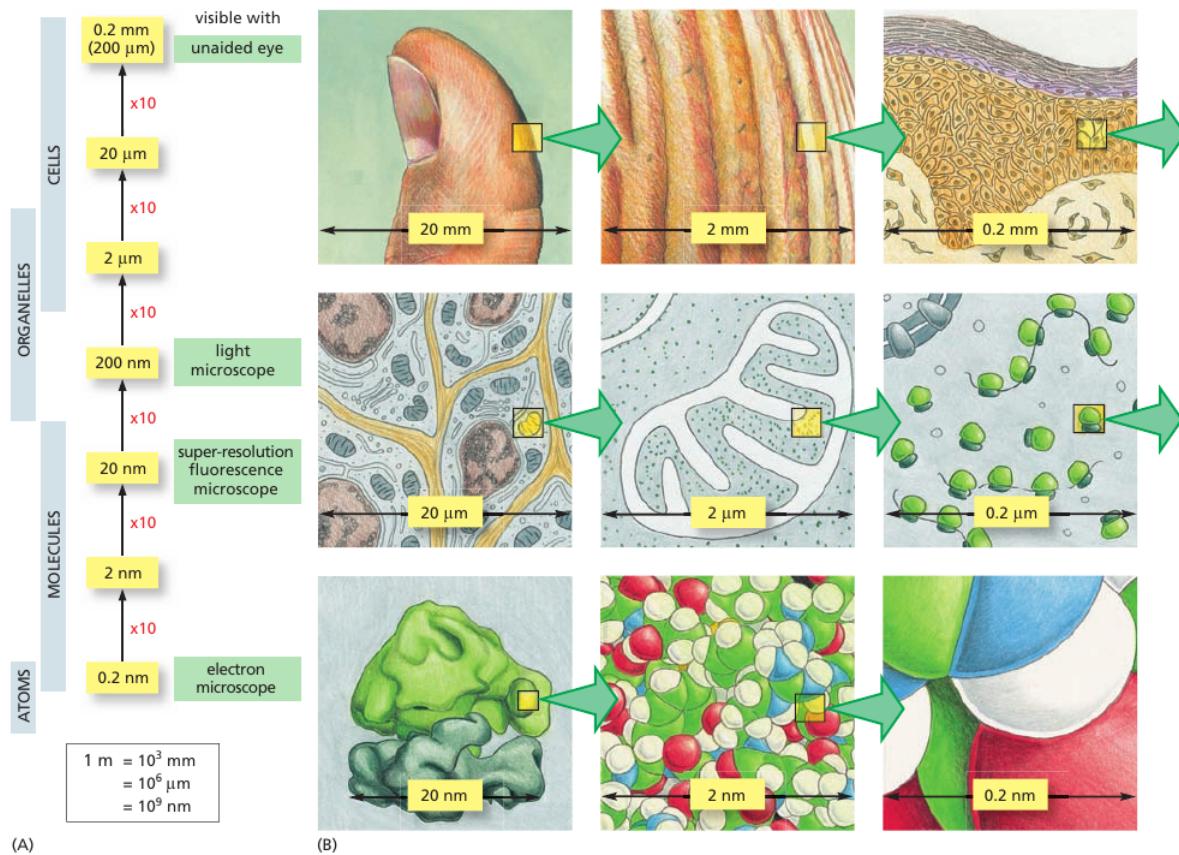


Figura 2.4: (A) Il grafico elenca le dimensioni dei livelli strutturali biologici, le unità di misura relative e gli strumenti necessari per visualizzarli. (B) Uno stesso dettaglio a varie scale di grandezza: pollice, pelle, cellule, mitocondrio, ribosomi, insieme di atomi che formano parte di una proteina. I dettagli molecolari sono oltre la potenza del microscopio elettronico. Fonte: [4]

2.1.2 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 striature non potrebbero svolgere la respirazione cellulare, se i neuroni non avessero

lunghi assoni non riuscirebbero a comunicare oppure si pensi al *paramecium* che si muove come un sommersibile grazie alle sue ciglia (vedi figura 2.1B).

- *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 codificano 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^[4]: una cellula di lievito ha quindi tutto il macchinario molecolare necessario per leggere e interpretare il nostro codice genetico e utilizzarlo 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).

2.1.3 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 sottoforma di RNA e se codificante questo è poi tradotto in proteine, concepite come la forma operativa e terminale delle informazioni contenute nel genoma^[9].

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 zucchero pentoso (*ribosio* nell’RNA e *desossiribosio*

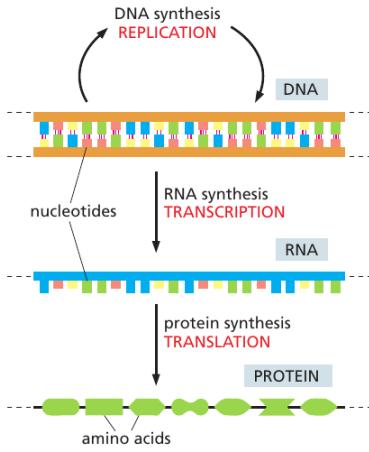


Figura 2.5: Dogma centrale in biologia. Fonte [4]

nel DNA) e un gruppo fosfato (vedi figura 2.7). 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):

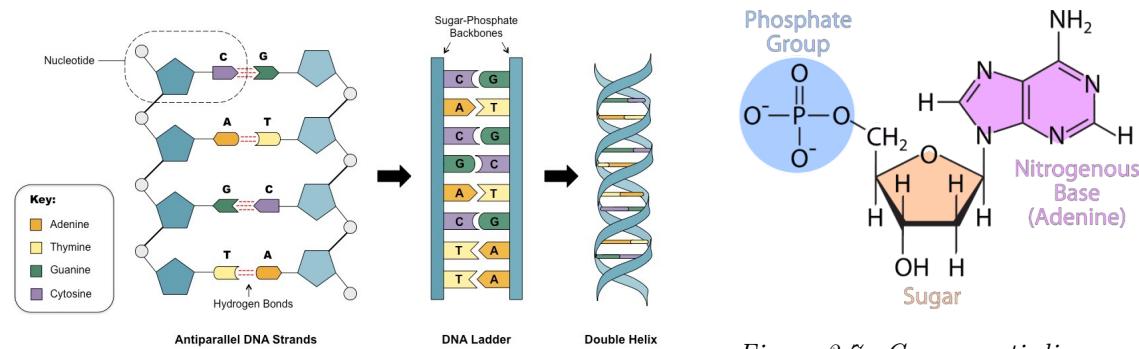


Figura 2.7: Componenti di un nucleotide con Adenina per base azotata. Fonte [11]

Il *genoma* indica il patrimonio complessivo del DNA di una cellula. Lo stesso gene nella stessa specie può esistere in varie forme, con leggere differenze nella sequenza nucleotidica: si sta parlando dei differenti *alleli* del gene. Gli alleli di tutti i geni di un individuo determinano il suo *genotipo*. 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 afferenti a organi o tessuti diversi esprimono geni differenti (*espressione genica*).

L'RNA (*acido ribonucleico*) esiste in varie forme. Le differenze con il DNA sono mostrate nella figura 2.8, 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
- RNA catalitico o ribozima, enzima ad RNA, è una molecola di RNA in grado di catalizzare una reazione chimica similmente agli enzimi
- snRNA, hnRNA

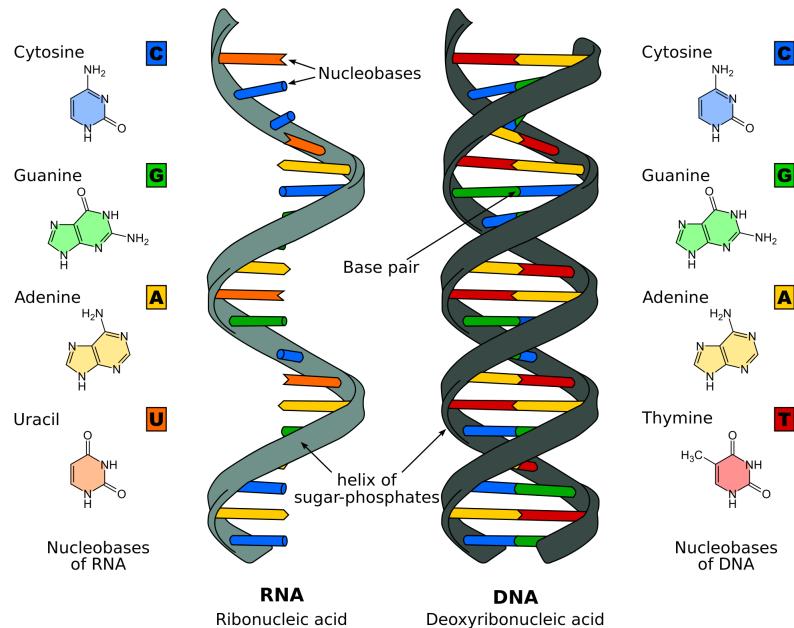


Figura 2.8: Differenze fra RNA e DNA Fonte: [12]

Il DNA dell'uomo contiene 3^9 coppie di nucleotidi, contenente circa 21000 geni codificanti: se il genoma umano venisse esteso in lunghezza sarebbe lungo 2,2 metri. Il batterio più semplice contiene 500 geni codificanti mentre il genoma di un'ameba è 100 volte più lungo di quello umano^[4], tanto per avere una visione quantitativa della diversità genetica tra gli organismi.

2.1.4 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 tripletta: si parla di *codice a triplette*.

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 triplette 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 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 questa fase 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, da un'estremità vi è l'anticodone (interfaccia con il linguaggio genetico) e dall'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.

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 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.

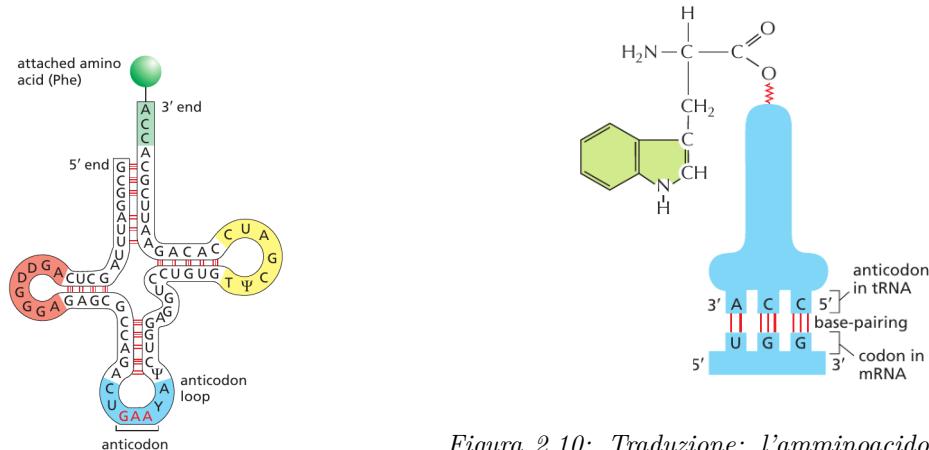


Figura 2.9: tRNA. Fonte [4]

Figura 2.10: Traduzione: l'amminoacido triptofano (*Trp*) è codificato dal codone UGG nell'mRNA e si lega al tRNA tramite un legame energetico forte. Fonte: [4]

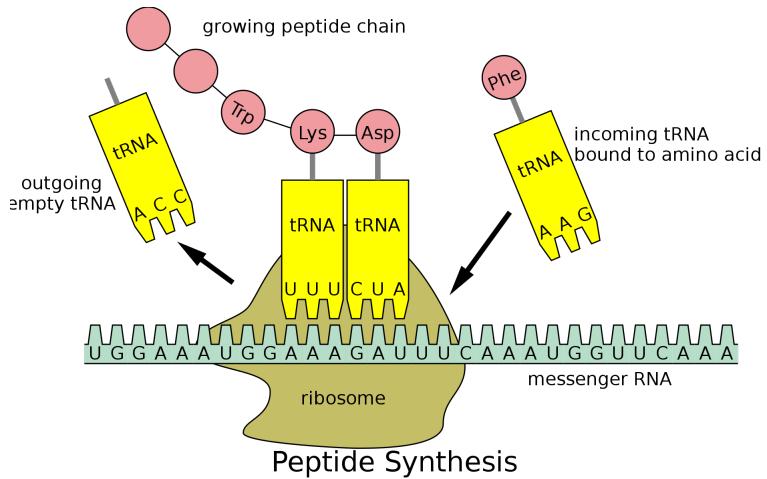


Figura 2.11: Traduzione, sintesi peptidica. Fonte: [13]

La traduzione comincia con il primo codone (AUG, che oltre a segnalare l'inizio codifica anche la metionina, vedi figura 2.12) 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.

	AGA		UUA		AGC		GUA		
	AGG		UUG		AGU		GUC		UAA
codons	GCA	CGA	GGA	CUA	CCA	UCA	ACA		
	GCC	CGC	GGC	AUC	CCC	UCC	ACC	GUC	UAG
	GCG	CGG	GAC	CAC	CCG	UCG	ACG	GUG	UAG
	GCU	CGU	AAC	GGG	CCU	UCU	ACU	GUU	UGA
		UGC	GAA	CAU	AUC	AAA	UUU		
		GAU	GAG	GGU	CUG	AAG	AUG		
		AAU	CAG	CAU	AAA				
amino acids	Ala	Arg	Asp	Asn	Ile	Leu	Lys	Met	Val
	A	R	D	N	I	L	K	M	V
					F	P	S	T	stop
						W	Y		

Figura 2.12: Codici a tripletta degli amminoacidi. Fonte: [4]

2.1.5 Proteine: le macromolecole più importanti della vita

Le proteine sono formate dall'unione di strutture più semplici: gli amminoacidi. Un polimero amminoacidico composto da meno di 50 amminoacidi è chiamato *peptide*, se supera tale soglia *polipeptide*. Una proteina può essere quindi sia un semplice peptide² 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_α circondato da 4 componenti (vedi fig. 2.14):

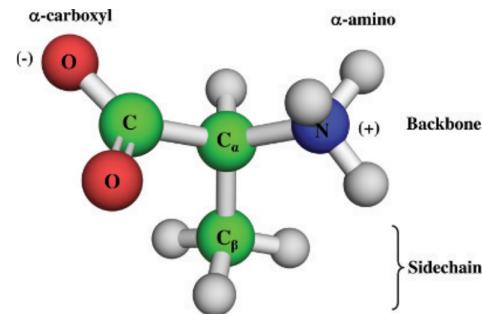
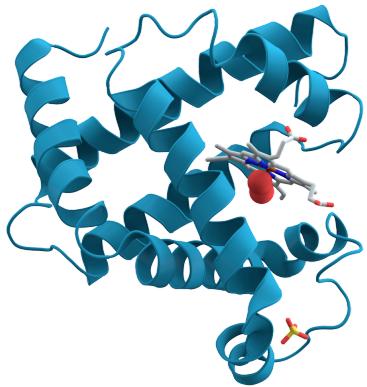
1. un atomo di idrogeno
2. un gruppo amminico ($\alpha - amino$), (-NH₂) in condizioni fisiologiche carico positivamente (-NH₃⁺)
3. un gruppo carbossilico ($\alpha - carboxyl$), (-COOH) carico negativamente (-COO⁻)
4. un gruppo R, gruppo laterale chiamato anche *residuo* che per sineddoche indica l'intero 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 2.12 o 2.17). 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 2.17) 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.

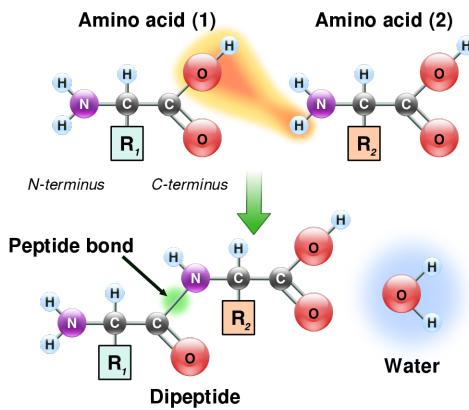
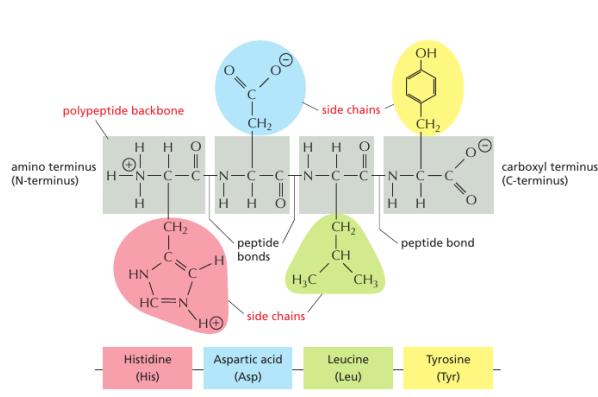
²Esempi di "semplici" peptidi che svolgono funzioni biologiche sono i *neuropeptidi* che agiscono da neurotrasmettore (ad es. endorfine) e *ormoni* quali l'insulina e il glucagone

bili. 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^[4]. Il legame peptidico comporta l'eliminazione della carica degli ex gruppi *amminico* e *carbossilico*.

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^[4]. 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. 2.15). È presente però un fenomeno che permette ai residui di interagire elettrostaticamente: la *risonanza elettronica*. Gli elettroni 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.

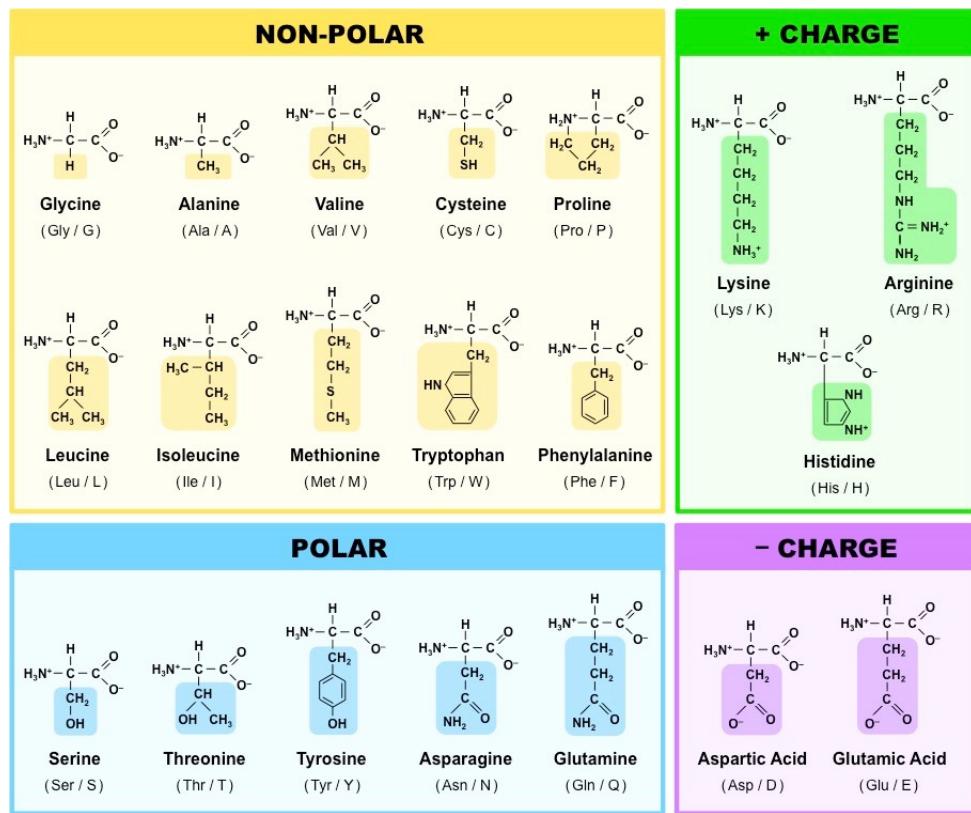


Figura 2.17: I 20 amminoacidi universali. Fonte: [16]

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 *dopamina*, *melanina* e *adrenalina*, il *triptofano* di *serotonin* 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 aminoacidi), 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 agli organi e ai tessuti del corpo così come a riportare CO₂ 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.

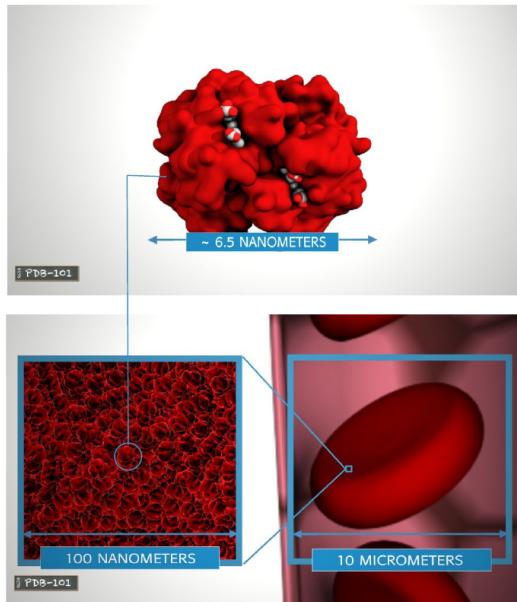


Figura 2.18: Emoglobina in diverse scale. Rapresentazione a superficie. Un globulo rosso contiene circa 280 milioni di molecole di emoglobina, per cui può portare più di 1 miliardo di molecole di ossigeno per volta. Fonte: [17]

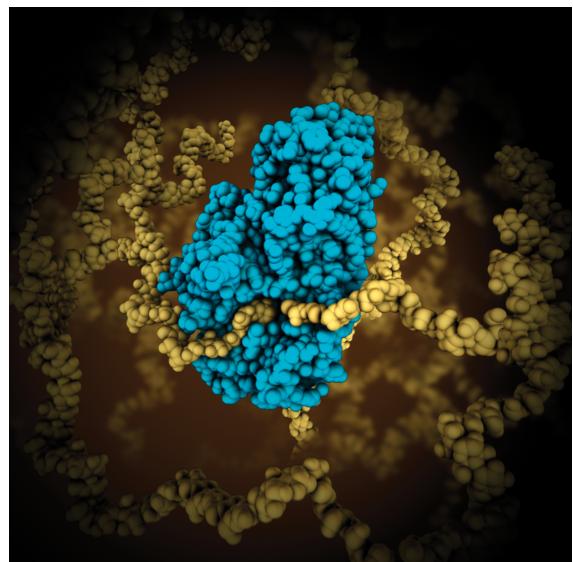


Figura 2.19: Enzima alpha Amilasi in turchese, rappresentazione di tipo space-filling. Si lega a catene di carboidrati (gialle) e le rompe in pezzi più piccoli di glucosio. Fonte [17]

Nelle cellule le proteine svolgono, fra le altre, funzioni di supporto strutturale (*collagene*), mobilità (*actina*, *miosina*), protezione (*anticorpi*), regolazione, ormoni (*insulina*), trasporto, catalisi, magazzino. Nel nostro corpo abbiamo un numero grandissimo di proteine: 10²⁷. Per usare una metafora di Ken Dill^[18] 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^[19] al contrastare l'inquinamento ambientale^[20].

2.2 Background informatico

2.2.1 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 provenienti 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^[21]. 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^[22].

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 l’idea di informazione e computazione possa essere la chiave per comprendere la biologia)^[23].

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 carat-

teristiche astratte, proprio come nel caso di AlphaFold per il problema della predizione della struttura di proteine.

2.2.2 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^[24] e vi entra pienamente: la stima di performance in ML è infatti l'*accuratezza predittiva*, stimata dall'errore calcolato sul test set.

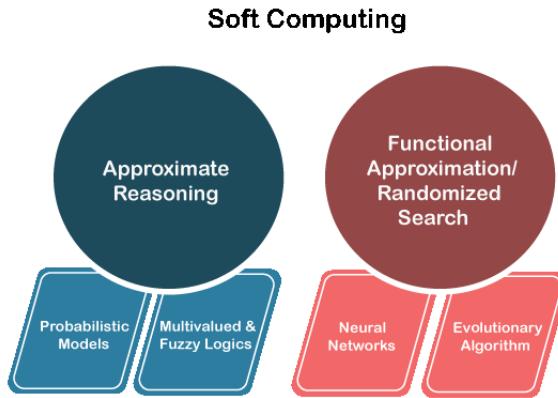


Figura 2.20: Branche del soft computing. Fonte: [25]

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 *crossover*.

2.2.3 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»*³

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

«*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»*⁴

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.

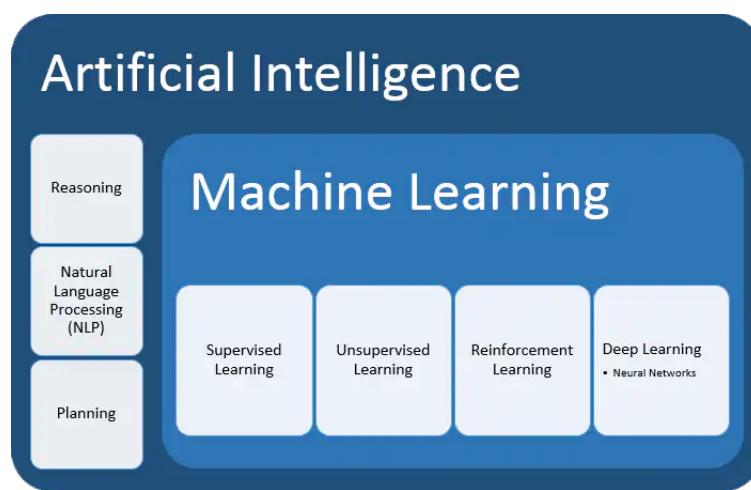


Figura 2.21: Schema riassuntivo dei campi dell'IA. Fonte: [28]

³R. Kurzweil, R. Richter, R. Kurzweil et al., *The age of intelligent machines*, 1990

⁴H. H. the XIV Dalai Lama, *The heart of the Buddha's path*, 2011

2.2.4 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?".

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.»⁵

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.

⁵T. Mitchell, *Machine learning*. McGraw hill New York, 1997

2.2.5 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 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.

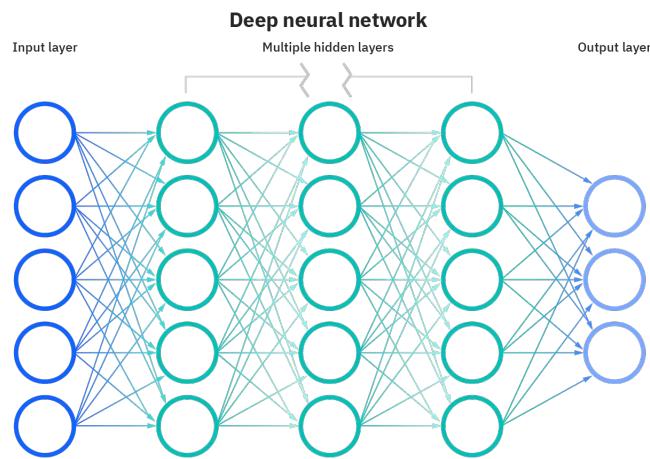


Figura 2.22: Rete neurale artificiale. Fonte: [30]

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 più di 3 strati (inclusi quello di input e output) può essere considerata un algoritmo di *deep learning*^[30]. Una rete neurale con 2 o 3 strati è una rete neurale semplice.

Capitolo 3

Protein Folding

«la forma è l'immagine plastica della funzione»¹

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.

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.

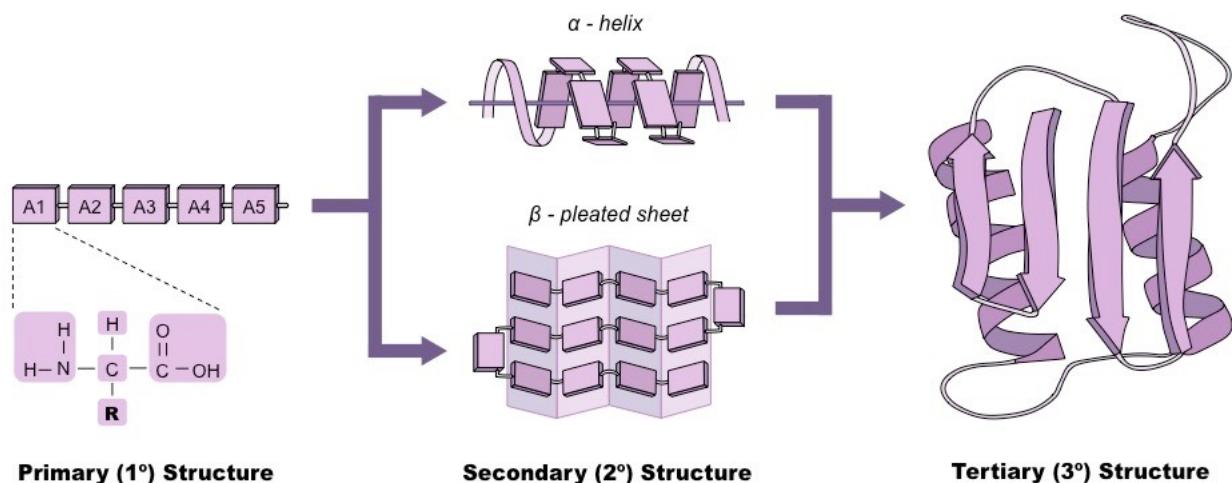


Figura 3.1: Protein folding: dagli amminoacidi alla struttura tridimensionale. Fonte: [32]

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

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 finale².

La prima teoria del ripiegamento proteico è stata proposta negli anni venti del 20° secolo da Hsien Wu^[33], in relazione al processo di denaturazione (vedi sezione 3.1.2). È però Anfinsen, premio Nobel per la chimica, negli anni '60 a compiere un fondamentale passo nella comprensione del processo del ripiegamento proteico^[34].

3.1 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*.

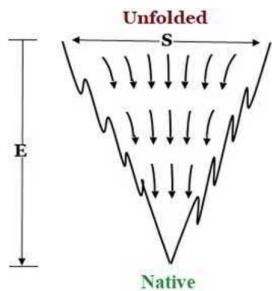


Figura 3.2: Un "panorama" idealizzato dell'energia libera a forma di imbuto. E=energia, S=entropia.
Fonte: [35]

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 di minimo (visivamente simile ad un imbuto, vedi fig. 3.2); la superficie di energia libera nelle vicinanze dello stato nativo deve essere abbastanza ripida ed elevata

²ciò non è vero nel 100% dei casi, alcune proteine possono avere più di una conformazione stabile per adempire funzioni diverse (vedi la sezione 3.4.2) e alcune proteine possono andare incontro a misfolding (vedi la sezione 3.3)

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

3.1.1 Esperimento di Anfinsen

L'esperimento, compiuto nel 1957^[36], 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 ponti disolfuro ($-CH_2 - S-S - CH_2 -$, vedi sez. 3.2.1). È stato usato un agente riducente per scindere questi ponti 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 ponti 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.

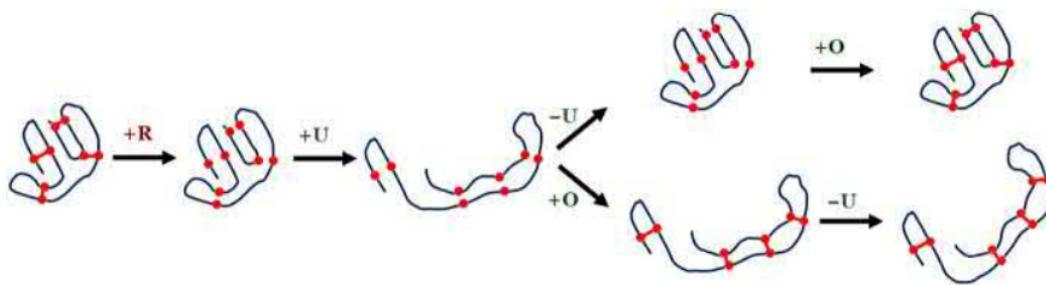


Figura 3.3: Rappresentazione schematica dell'esperimento di Anfinsen. R=reducing agent, U=Urea, O=oxidizing agent, punti rossi=cisteina, linee rosse=ponti disolfuro. Fonte: [35]

I ponti 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.

3.1.2 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.

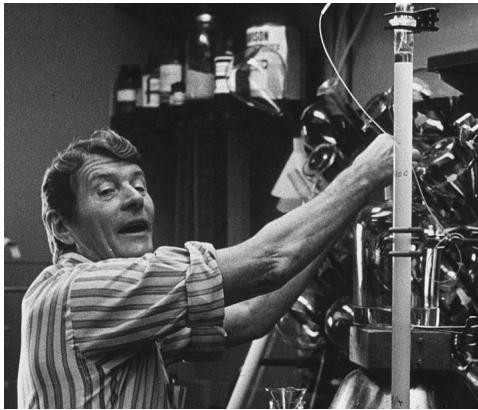


Figura 3.4: C.B. Anfinsen nel suo laboratorio.
Fonte: [37]

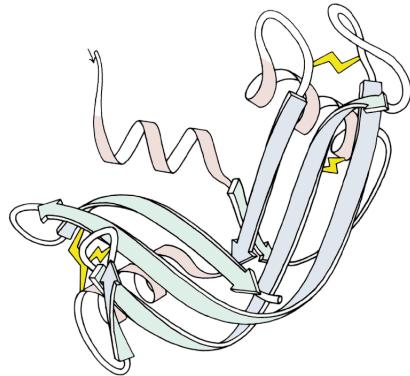


Figura 3.5: Ribonucleasi A, rappresentazione a nastro. In giallo i ponti disolfuro, rosa le α -eliche, verde e azzurro i β -foglietti. Fonte [38]

Può avvenire che una volta rimosso l'agente denaturante la proteina ritorni allo stato di partenza (*rinaturazione*) ma spesso il processo è irreversibile.

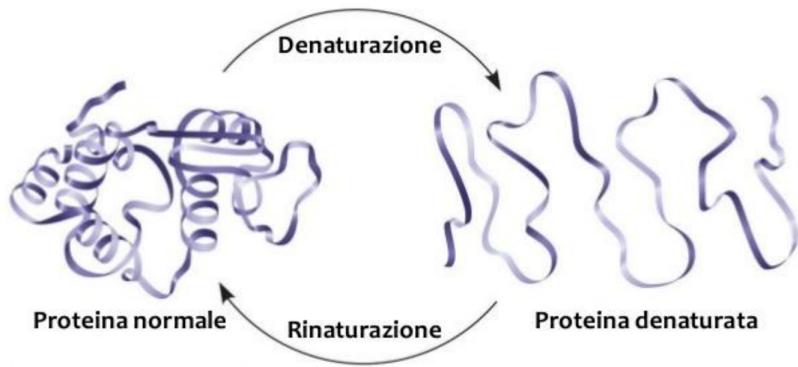


Figura 3.6: Denaturazione e rinaturazione. Fonte: [39]

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 termodynamicamente 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' α -cheratina, rompendo e riformando ponti disolfuro).

3.2 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:

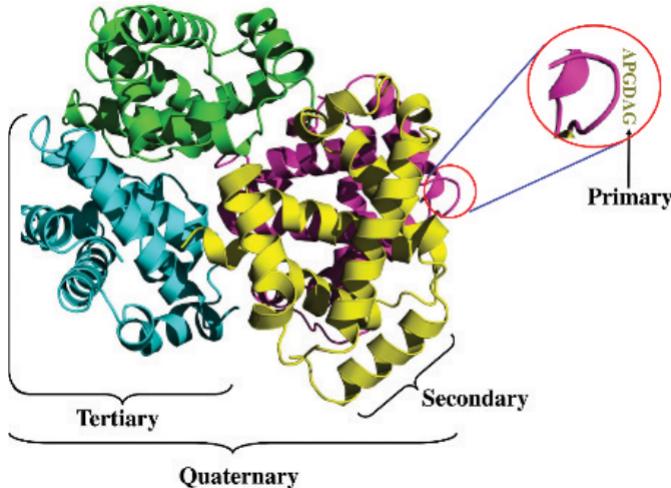


Figura 3.7: Livelli strutturali di una proteina. Fonte: [3]

- *struttura primaria*: la sequenza ordinata degli aminoacidi
- *struttura secondaria*: regioni ripetitive locali stabilizzate da legami idrogeno tra atomi della backbone (α -eliche e β -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.

3.2.1 Legami e interazioni molecolari

La chimica della vita è di un tipo speciale e niente disobeisce alle leggi della chimica e della fisica³. È una chimica organica formata da composti carboniosi, in un ambiente acquoso, con temperature "terrestri" e complicata, basata su grandi polimeri. Gli elementi

³Assumendo un paradigma meccanicistico, come accennato all'inizio di questa tesi

non si possono modificare chimicamente ma gli atomi, particelle più piccole associate ad un elemento, 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 spostano⁴ per ricercare stabilità, ovvero per 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).

- *legame covalente*: prevede la compartecipazione 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 δ^-) 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 architettonicale fondamentale nella struttura delle proteine. La cisteina presenta un gruppo -SH nella catena laterale e può quindi formare ponti disolfuro.
- *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.

⁴Protoni 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 3.8: 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]

– *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 eletropositivo è parzialmente condiviso da due atomi elettonegativi; 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).

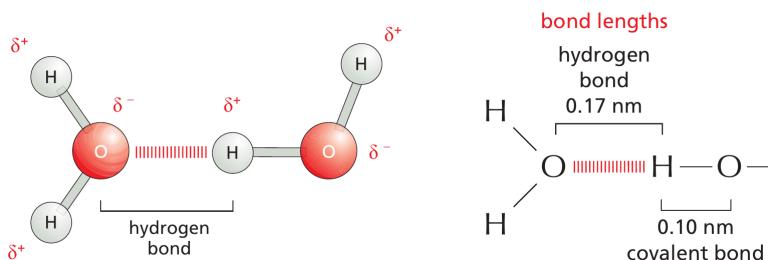


Figura 3.9: Legame idrogeno tra due molecole d'acqua. Fonte: [4]

– *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)

- *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.

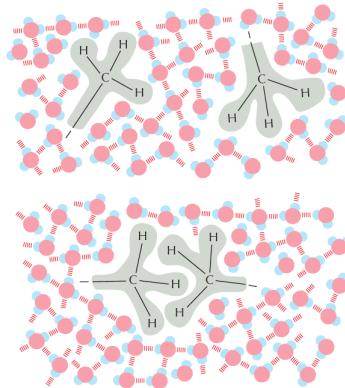


Figura 3.10: Forze idrofobiche. Fonte: [4]

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).

Struttura primaria

La struttura primaria delle proteine è la sequenza ordinata degli amminoacidi che ne determina la conformazione nativa. 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: α -eliche e β -foglietti.

Questo livello di organizzazione è una conseguenza dei legami a idrogeno tra gli amminoacidi appartenenti a una stessa catena, o tra gli amminoacidi di catene diverse. All'interno della backbone del polipeptide gli atomi di ossigeno hanno una parziale carica

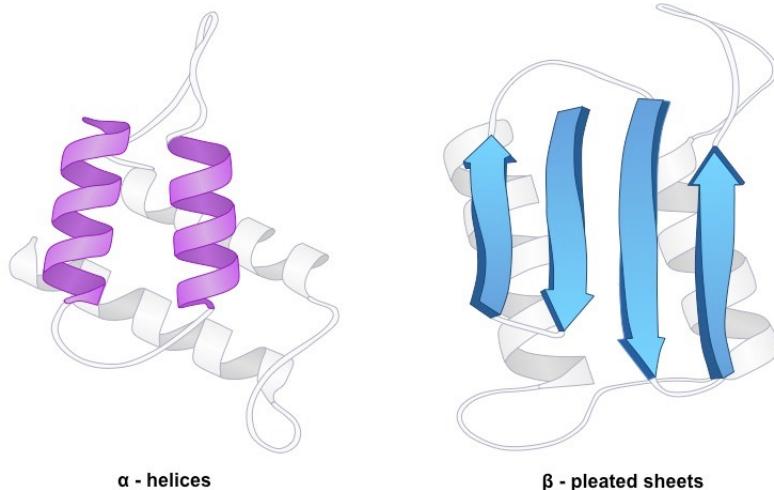


Figura 3.11: Struttura secondaria delle proteine, α -eliche e β -foglietti. Fonte: [32]

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 α -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 appunto a distanza di quattro 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' α -elica non può formarsi, a causa delle forze di repulsione che si generano tra i residui. Alcune proteine fibrose, come l' α -cheratina, la proteina strutturale di capelli, lana e unghie hanno formazioni di α -eliche sulla maggior parte della loro lunghezza.

Altre proteine fibrose sono invece dominate dai β -foglietti, come le proteine della seta (β -cheratina) e della tela prodotta dai ragni. In queste conformazioni due o più segmenti della catena polipeptidica giacenti lato su lato (chiamati β -filamenti) sono connessi da tre o più legami idrogeno. Si definisce β -filamento una sequenza peptidica di amminoacidi (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. In questo caso, i legami a idrogeno si formano tra gli amminoacidi di due catene adiacenti. I gruppi amminici di uno scheletro peptidico formano legame con quelli carbossilici del

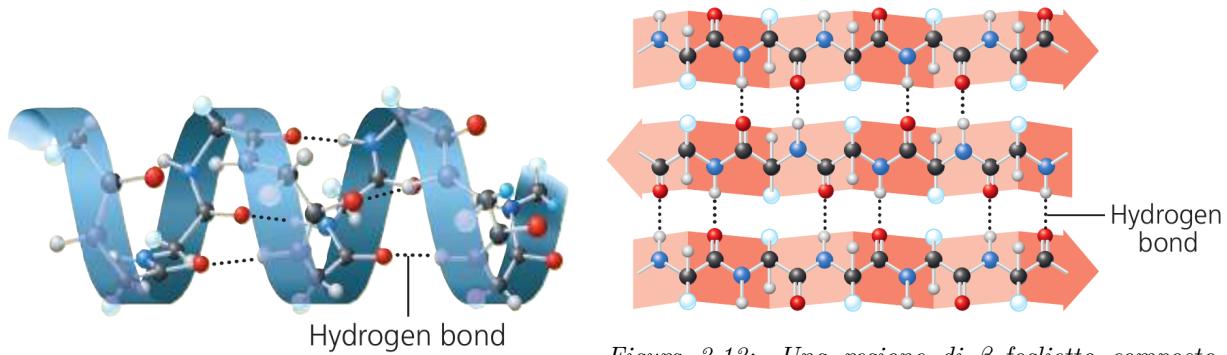


Figura 3.12: Regione di α -elica. Fonte: [39]

Figura 3.13: Una regione di β -foglietto composta da β -filamenti adiacenti, spesso mostrati come una freccia pieghettata o piatta puntata in direzione C-terminus. Fonte [39]

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 β -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. Può essere considerata come esempio di struttura supersecondaria la triplice elica allungata del collagene.

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

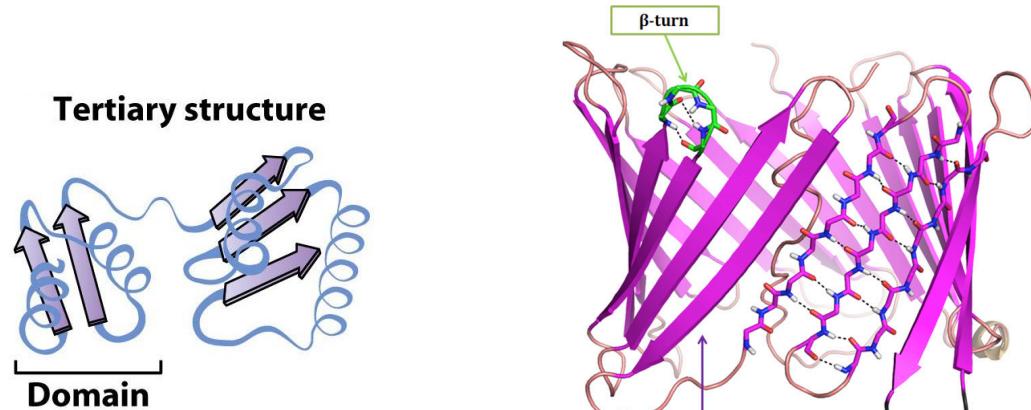


Figura 3.14: Dominio in una proteina. Fonte: [40]

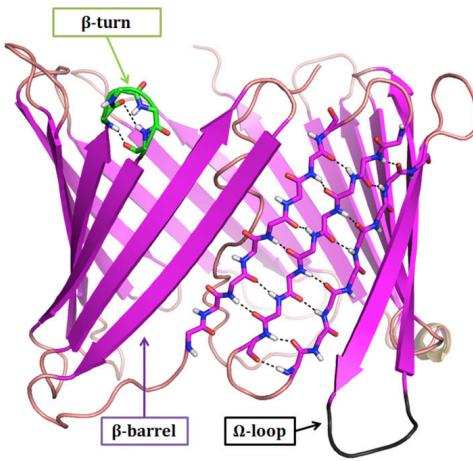


Figura 3.15: Struttura con giri, loop e motivo β -barile. Fonte: [41]

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.

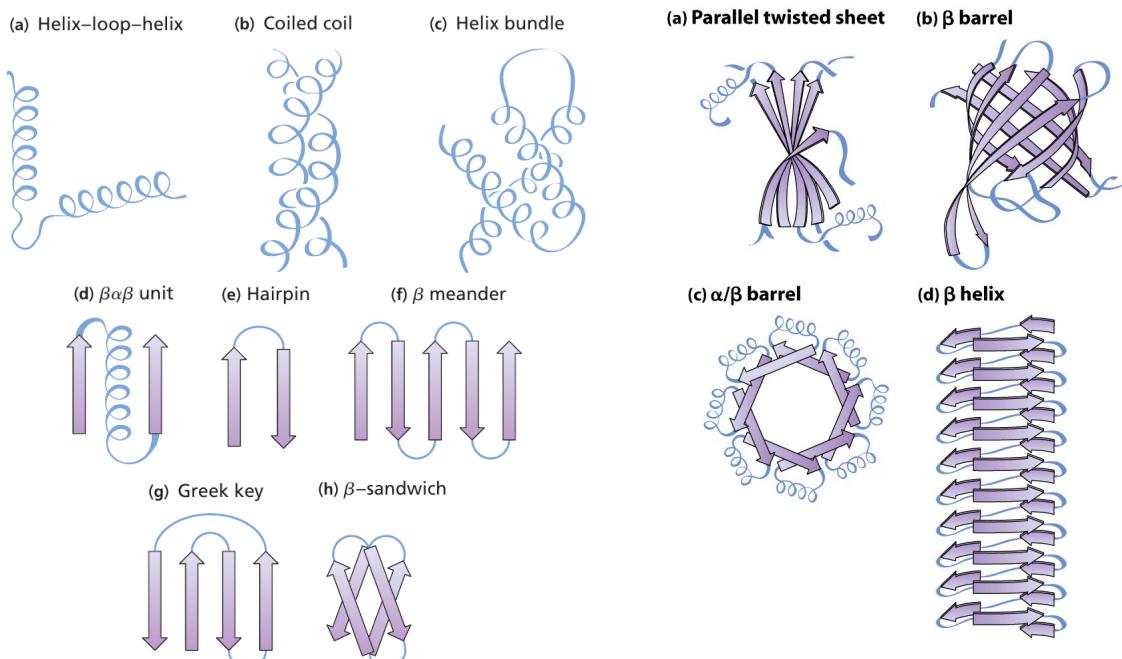


Figura 3.16: Motivi comuni. Fonte [40]

Figura 3.17: Domain folds (ripiegamenti di dominio). Fonte: [40]

In figura 3.16 sono illustrati alcuni motivi comuni nelle strutture proteiche. Il motivo

elica-loop-elica ad esempio consiste di due α -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 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 β -filamenti antiparalleli in un β -foglietto dove l'ordine dei foglietti lungo la catena polipeptidica è 4,1,2,3⁵. In figura 3.15 e 3.17 è illustrato il motivo β -barile composto da β -foglietti ripiegati circolarmente a formare una struttura somigliante ad un barile comune in molte proteine di membrana.

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 che vi sono un numero limitato di possibili ripiegamenti nell'universo delle strutture proteiche^[42]. 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^[43].

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.

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.

⁵I numeri indicano l'ordine dei filamenti ovvero la loro posizione nel β -foglietto da destra a sinistra

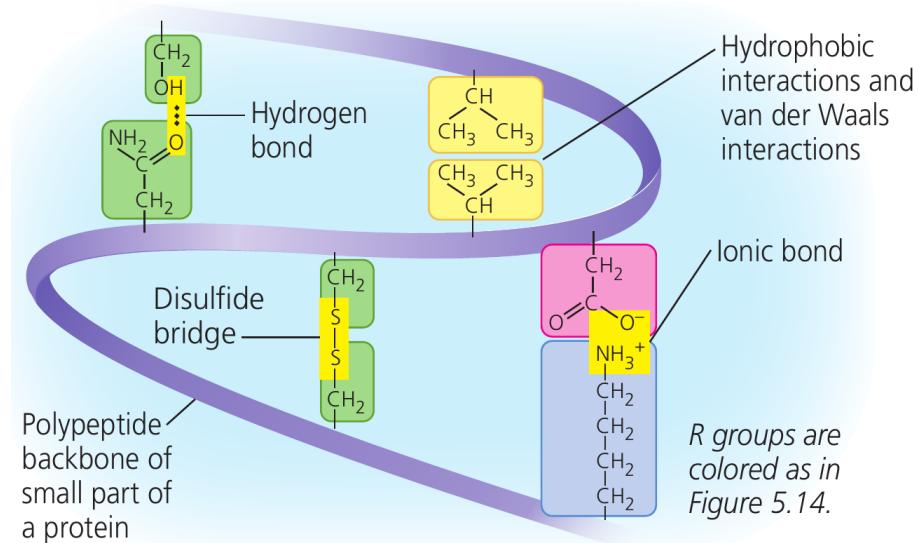


Figura 3.18: I diversi tipi di interazioni che possono contribuire alla struttura terziaria di una proteina.
Fonte: [39]

La glicina assume una speciale posizione tra gli amminoacidi dato che ha il gruppo R più piccolo, un solo atomo di idrogeno (vedi fig. 2.17): 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. 3.5) era visto come la somma di molte piccole interazioni (legami idrogeno, interazioni di van der Waals, attrazioni elettrostatiche ...) ma senza nessuna forza dominante^[44]. Negli anni '80, grazie alla modellazione basata sulla meccanica statistica, è emerso un nuovo paradigma: la componente dominante nel folding code è sono le forze idrofobiche, il folding code è distribuito sia localmente che non localmente nella sequenza e che le strutture secondarie di una proteina sono una conseguenza della struttura terziaria, tanto quanto una causa.

Struttura quaternaria

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. Un altro esempio è l'emoglobina, proteina globulare formata da 4 subunità polipeptidiche. Le strutture terziarie delle subunità non vengono alterate.

La classificazione delle proteine può essere basata su somiglianze strutturali e/o di sequenza.

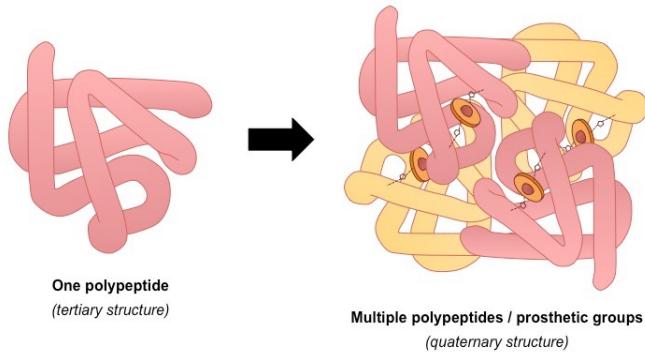


Figura 3.19: Rappresentazione di una struttura quaternaria composta da più polipeptidi e alcuni gruppi prostetici. Fonte [32]

3.2.2 Geometria dei legami

Le tante conformazioni possibili della catena polipeptidica sono possibili grazie alla rotazione di essa attorno all'atomo C_α .

- Limiti al ripiegamento: angoli di tensione e piano di Ramachandran
- Domini, Residui, Motivi, Giri, Loops, Turns

3.2.3 Energetica del ripiegamento

- processo spontaneo: energia di Gibbs, entalpia, entropia

3.3 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.

Più in dettaglio i chaperoni molecolari svolgono le seguenti funzioni:

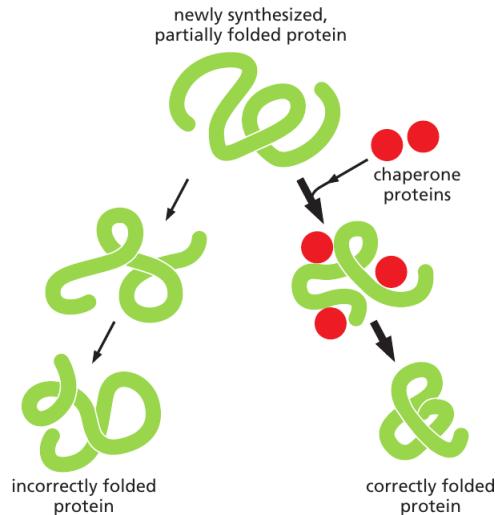


Figura 3.20: Schema della funzione dei chaperoni molecolari. Fonte: [4]

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 conservate, 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^[45], famosa è la chaperonina procariotica GroEL (vedi fig. 3.21), che può essere assunta come modello di riferimento delle chaperonine.

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^[4].

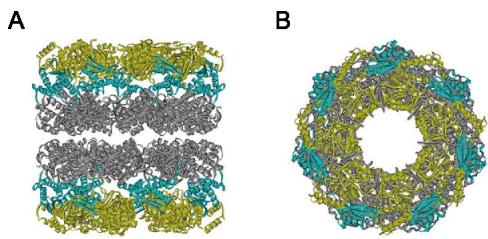


Figura 3.21: Strutture dei complessi GroEL e GroEL-GroES. (B) si può osservare la tipica forma ad anello.
Fonte: [46]

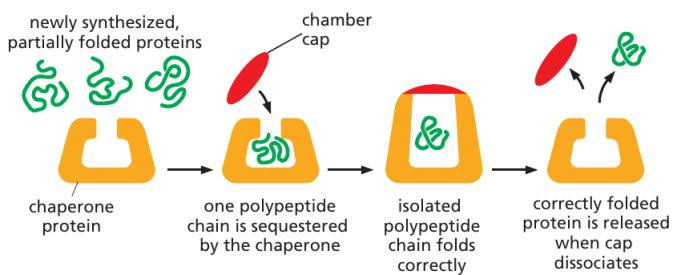


Figura 3.22: Rappresentazione schematica della funzione della camera di isolamento nelle chaperonine. Fonte [4]

3.3.1 Misfolding e malattie

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 β -foglietti organizzati in una struttura denominata cross- β , disposizione molto stabile e insolubile, generalmente 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. 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^[47].

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.⁶ 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 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^c*(cellulare, forma normale) è codificata da un solo gene, PRNP. La *PrP^{sc}*(scrapie, forma patologica) differisce dalla proteina naturale *PrP^c* per la conformazione tridimensionale: la *PrP^c* ha una struttura più aperta contenente 3 segmenti ad α -eliche e pochi β -foglietti; la *PrP^{sc}* invece ha una struttura più compatta e stabile e presenta un aumento di β -foglietti.

3.3.2 Controllo qualità e proteasomi

L'uscita delle proteine dal reticolo endoplasmatico (RE) è controllata per assicurare la qualità delle proteine. Sebbene alcune proteine 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

⁶I prioni sono stati studiati e denominati in questo modo dal premio Nobel per la medicina nel 1997 Stanley Prusiner^[48]

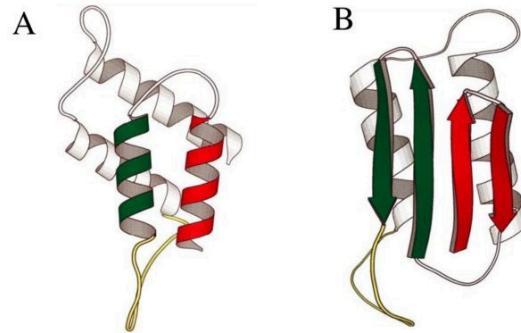


Figura 3.23: (A) Struttura della *PrP^c*. (B) Struttura della *PrP^{sc}*. Fonte: [49]

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 conservate 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'ubiquitina⁷.

Ad esempio molecole di anticorpi sono composte 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^[4] (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.

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

⁷Per "la scoperta della degradazione delle proteine mediata da ubiquitina" è stato assegnato il Premio Nobel per la chimica del 2004 ad Aaron Ciechanover, Avram Hershko ed Irwin Rose.

la funzione di questi complessi è altamente conservata.

A causa del ruolo dei proteasomi nella regolazione del ciclo cellulare e dell'apoptosi⁸, 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.

3.3.3 Unfolded protein response

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 innesca 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 secerne insulina sono eliminate e la richiesta per quelle sopravvissute aumenta rendendole sempre più vulnerabili a questo meccanismo, esacerbando ulteriormente la malattia^[4].

3.4 Eccezioni al postulato di Anfinsen

– IDP [soft computing articolo]— The structure of some proteins is difficult to determine for a simple reason: A growing body of biochemical research has revealed that a significant number of proteins, or regions of proteins, do not have a distinct 3-D structure until they interact with a target protein or other molecule. Their flexibility and indefinite structure

⁸Il 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

are important for their function, which may require binding with different targets at different times. These proteins, which may account for 20–30mammalian proteins, are called intrinsically disordered proteins and are the focus of current research.

3.4.1 Intrinsically disordered proteins

3.4.2 Fold switching proteins

Some proteins have multiple native structures, and change their fold based on some external factors. For example, the KaiB protein complex switches fold throughout the day, acting as a clock for cyanobacteria. It has been estimated that around 0.5–4% of PDB proteins switch folds.[7] The switching between alternative structures is driven by interactions of the protein with small ligands or other proteins, by chemical modifications (such as phosphorylation) or by changed environmental conditions, such as temperature, pH or membrane potential. Each alternative structure may either correspond to the global minimum of free energy of the protein at the given conditions or be kinetically trapped in a higher local minimum of free energy.[8]

— porter, youtube — “I study proteins and proteins have been thought to have one structure that has one function or fold. I’m studying this group of proteins called fold switching proteins. So, they can actually change their structures and their functions in response to changes in the cell. So, you can kind of imagine fold switching proteins are like a transformer where in one case the protein is like a robot that does one thing and then in another case, in response to changes in our bodies, it becomes a car and can do something else. And an advantage to this is it can respond really quickly to changes in our bodies

3.4.3 box: Filosofia della scienza

3.5 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 stati strutture intermedie sulla loro 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^[44]:

- *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 cinetica di quali percorsi alcune proteine usano 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.

- le domande di principio del protein folding
- paradosso di Levinthal

Despite their large conformational spaces, proteins often fold in the microsecond or millisecond time range.

Come accennato nella sezione 3.3, 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.

Capitolo 4

Predizione della struttura di proteine

I biochimici conoscono oggi la sequenza amminoacidica per più di 225 milioni di proteine^[50] (UniProt), con circa 4.5-5 milioni aggiunte ogni mese mentre la struttura tridimensionale è conosciuta per quasi 200.000 proteine^[50] (PDB)¹ con più di 10.000 strutture aggiunte ogni anno.

Anche quando gli scienziati hanno una proteina correttamente ripiegata fra le mani non è così semplice determinare la sua esatta conformazione tridimensionale, considerando che si parla di migliaia di atomi.

¹al 25 Gennaio 2022 sono presenti 197.514 strutture di macromolecole quindi non solo proteine ma anche acidi nucleici e altre strutture complesse^[51]

Bibliografia

Libri

- [3] A. Kessel e N. Ben-Tal, *Introduction to proteins: Structure, function and motion*, 2^a ed. Chapman e Hall/CRC, 2018.
- [4] B. Alberts, D. Bray, K. Hopkin et al., *Essential cell biology*, 5^a ed. W. W. Norton e Company, 2019.
- [21] A. D. Baxevanis, G. D. Bader e D. S. Wishart, *Bioinformatics*. John Wiley & Sons, 2020.
- [29] T. Mitchell, *Machine learning*. McGraw hill New York, 1997.
- [35] S. Pal, *Fundamentals of Molecular Structural Biology*. Academic Press, 2019.
- [39] L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, R. B. Orr e N. A. Campbell, *Campbell Biology*. Pearson, 2021.
- [40] L. A. Moran, H. R. Horton, K. G. Scrimgeour, M. D. Perry e D. Rawn, *Principles of biochemistry*. Pearson London, 2012.

Articoli

- [19] 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.
- [20] 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.
- [22] 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.
- [23] M. Mitchell, “Biological Computation,” *PDXScholar*, 2010. indirizzo: https://pdxscholar.library.pdx.edu/compsci_fac/2.

- [33] 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.
- [34] C. B. Anfinsen, “The formation and stabilization of protein structure.,” *Biochemical Journal*, vol. 128, n. 4, p. 737, 1972.
- [36] 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.
- [41] 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>.
- [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.
- [45] N. A. Ranson, H. E. White e H. R. Saibil, “Chaperonins,” *Biochemical Journal*, vol. 333, n. 2, pp. 233–242, 1998.
- [46] 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.
- [48] 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.
- [49] B. Ruttkay-Nedecky, E. Sedlackova, D. Chudobova et al., “Prion protein and its interactions with metal ions (Cu^{2+} , Zn^{2+} , and Cd^{2+}) and metallothionein 3,” *ADMET and DMPK*, vol. 3, n. 3, pp. 287–295, 2015.

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] “eukaryote. Definition, Structure, Facts.” (19 set. 2019), indirizzo: <https://www.britannica.com/science/eukaryote> (visitato il 22/01/2022).
- [7] “Neurone - Wikipedia.” (27 dic. 2021), indirizzo: <https://it.wikipedia.org/wiki/Neurone> (visitato il 23/01/2022).

- [8] “Saccharomyces cerevisiae - Wikipedia.” (25 set. 2021), indirizzo: https://it.wikipedia.org/wiki/Saccharomyces_cerevisiae (visitato il 22/01/2022).
- [9] “Dogma centrale della biologia molecolare - Wikipedia.” (16 set. 2021), indirizzo: https://it.wikipedia.org/wiki/Dogma_centrale_della_biologia_molecolare (visitato il 22/01/2022).
- [10] “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).
- [11] 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%3AIntroductory_Chemistry_\(CK-12\)](https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Book%3AIntroductory_Chemistry_(CK-12)) (visitato il 22/01/2022).
- [12] “File: Difference DNA RNA-EN.svg - Wikimedia Commons.” (23 mar. 2010), indirizzo: https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-EN.svg (visitato il 22/01/2022).
- [13] “Transfer RNA - Wikipedia.” (23 gen. 2022), indirizzo: https://en.wikipedia.org/wiki/Transfer_RNA (visitato il 23/01/2022).
- [14] “Protein - Wikipedia.” (21 dic. 2021), indirizzo: <https://en.wikipedia.org/wiki/Protein> (visitato il 23/01/2022).
- [15] “Peptide bond - Wikipedia.” (4 nov. 2021), indirizzo: https://en.wikipedia.org/wiki/Peptide_bond (visitato il 23/01/2022).
- [16] “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).
- [17] “PDB101: Learn: Videos: What is a Protein?” (20 Nov. 2017), indirizzo: <https://pdb101.rcsb.org/learn/videos/what-is-a-protein-video> (visitato il 23/01/2022).
- [18] 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=zmv3kovWpNQ> (visitato il 06/01/2022).
- [24] “Apprendimento automatico - Wikipedia.” (1 dic. 2021), indirizzo: <https://it.wikipedia.org/wiki/ApprendimentoAutomatico> (visitato il 23/01/2022).
- [25] “What is soft computing - Javatpoint.” (3 lug. 2021), indirizzo: <https://www.javatpoint.com/what-is-soft-computing> (visitato il 24/01/2022).

- [28] “Machine Learning - IBM.” (29 ago. 2020), indirizzo: <https://www.ibm.com/it-it/analytics/machine-learning> (visitato il 23/01/2022).
- [30] “What are Neural Networks?” (1 Giu. 2021), indirizzo: <https://www.ibm.com/cloud/learn/neural-networks> (visitato il 24/01/2022).
- [32] “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).
- [37] “Christian B. Anfinsen - Digital Collections - National Library of Medicine.” (1 gen. 2022), indirizzo: <http://resource.nlm.nih.gov/101408166> (visitato il 25/01/2022).
- [38] “File:RibonucleaseA SS paleRib.png - Wikimedia Commons.” (16 mar. 2012), indirizzo: https://commons.wikimedia.org/wiki/File:RibonucleaseA_SS_paleRib.png (visitato il 25/01/2022).
- [42] 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).
- [43] “Protein structure prediction - Wikipedia.” (30 dic. 2021), indirizzo: https://en.wikipedia.org/wiki/Protein_structure_prediction (visitato il 27/01/2022).
- [47] “Chaperonina - Wikipedia.” (21 ott. 2021), indirizzo: <https://it.wikipedia.org/wiki/Chaperonina> (visitato il 26/01/2022).
- [50] “Latest Release Information.” (27 gen. 2022), indirizzo: <https://www.ddbj.nig.ac.jp/latest-releases-e.html> (visitato il 27/01/2022).
- [51] “wwPDB: Deposition Statistics.” (25 gen. 2022), indirizzo: <https://www.wwpdb.org/stats/deposition> (visitato il 27/01/2022).

Altre fonti

- [5] *Appunti del corso Elementi di Biologia e Neuroscienze, prof. Mario Pirchio, Unipi CdL Filosofia, 2021.*
- [26] R. Kurzweil, R. Richter, R. Kurzweil e M. L. Schneider, *The age of intelligent machines*, 1990.
- [27] H. H. the XIV Dalai Lama, *The heart of the Buddha's path*, 2011.
- [31] A. Ruffini, *Fisiogenia, la biodinamica dello sviluppo ed i fondamentali problemi morfologici dell'embriologia generale*, 1925.