



UNIVERSITÀ DI PISA

Corso di Laurea Triennale in Informatica (L-31)

TESI DI LAUREA

Protein Folding: dai metodi classici alla rivoluzione di AlphaFold

Relatore

Prof. Paolo Milazzo

Candidato

Ludovico Venturi

ANNO ACCADEMICO 2020/2021

Indice

1	Introduzione	2
2	Background	3
2.1	Background biologico	4
2.1.1	Organizzazione della vita: dagli atomi alle cellule	4
2.1.2	Concetti fondamentali in biologia	7
2.1.3	Dogma centrale della biologia	8
2.1.4	Proteine: le macromolecole più importanti della vita	13
2.2	Background informatico	16
2.2.1	Bioinformatica	16
2.2.2	Soft computing	17
2.2.3	Intelligenza Artificiale	18
2.2.4	Machine Learning	18
2.2.5	Reti neurali artificiali (ANN)	20
3	Protein Folding	22
4	Predizione della struttura di proteine	23
5	AlphaFold	24
6	Uso di AlphaFold e visualizzazione	25
7	Scenari aperti e conclusioni	26
	Bibliografia	27

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:

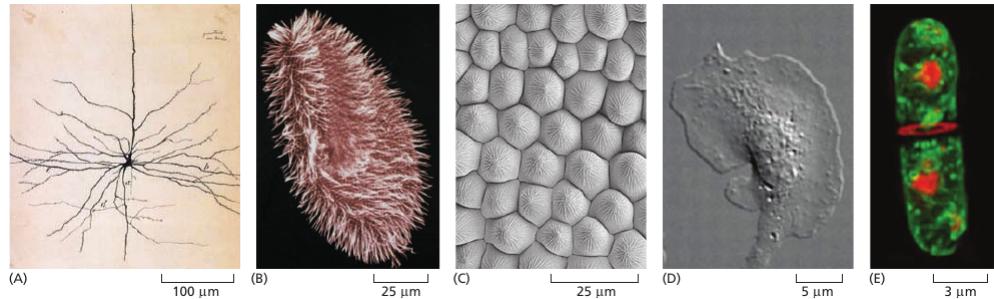


Figura 2.1: (A) disegno di un neurone. (B) Paramecium. (C) superficie di un petalo di fiore di bocca di leone. (D) Macrofago. (E) Un lievito di fissione viene catturato nell'atto di divisione cellulare. Fonte: [4]

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 10-100 μm) 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:

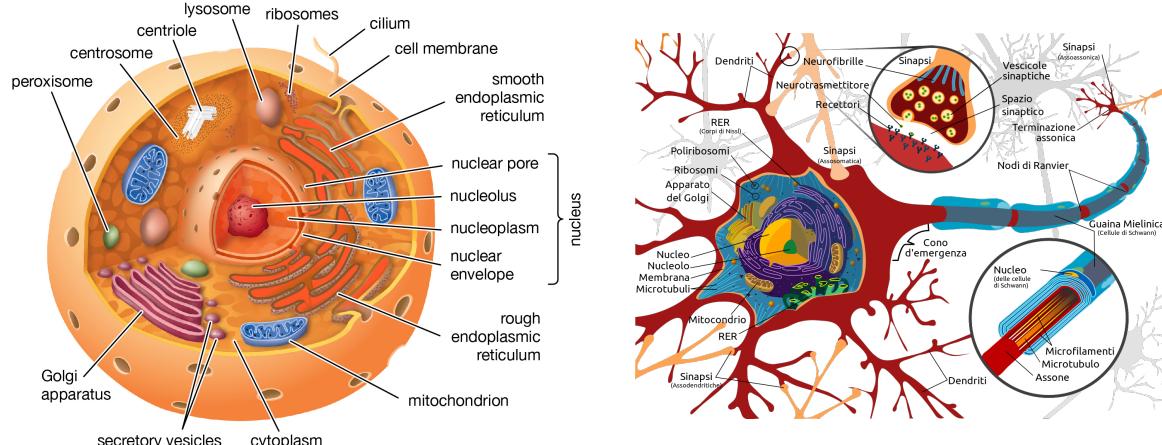


Figura 2.3: Neurone. Fonte [7]

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):

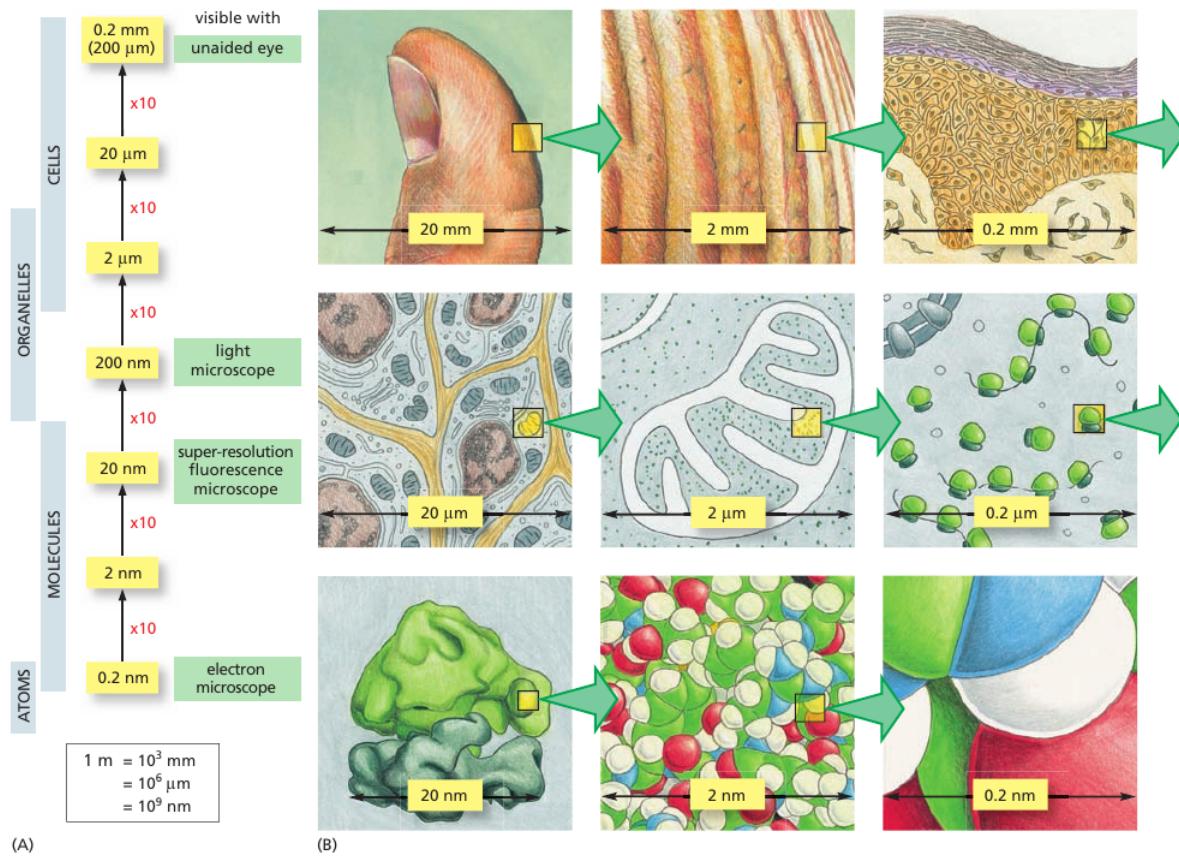


Figura 2.4: (A) Il grafico elenca le dimensioni dei livelli strutturali biologici, le unità di misura relative e gli strumenti necessari per visualizzarli. (B) Uno stesso dettaglio a varie scale di grandezza: pollice, pelle, cellule, mitocondrio, ribosomi, insieme di atomi che formano parte di una proteina. 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*

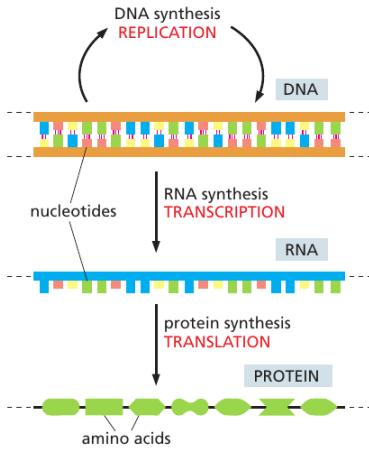


Figura 2.5: Dogma centrale in biologia. Fonte [4]

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):

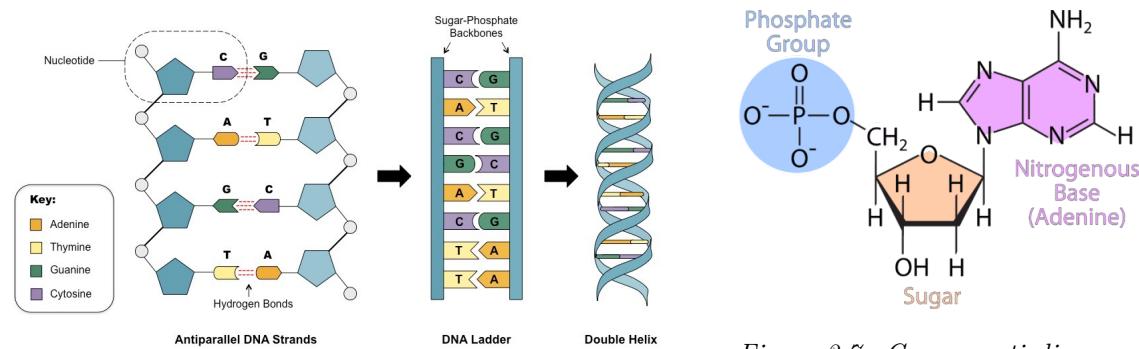


Figura 2.6: struttura del DNA. Fonte: [10]

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

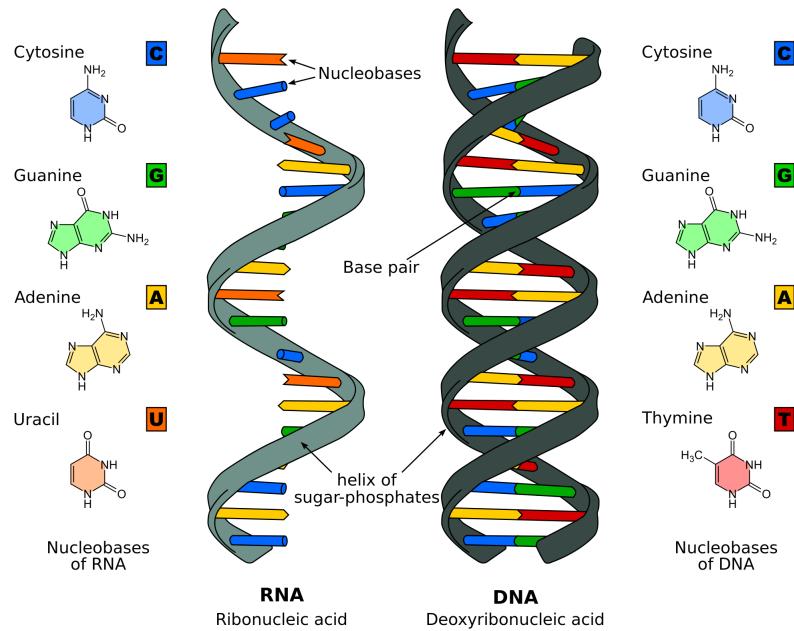


Figura 2.8: Differenze fra RNA e DNA Fonte: [12]

Il DNA dell'uomo contiene 3^9 coppie di nucleotidi, contenente quasi 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.

2.1.3.1 Dai geni alle proteine

Il codice genetico lavora a sequenze di codici di 3 lettere (es. "GAA" = Glutammato), questo perché si hanno a disposizione 4 lettere (le basi azotate) e si devono codificare i 20

diversi amminoacidi. Con 2 lettere avrei 4^2 possibilità che non sono sufficienti a descrivere 20 informazioni diverse, si utilizzano pertanto 3 lettere anche se ciò causa ridondanza nei codici. Un amminoacido è quindi codificato da una tripla: si parla di *codice a triplett*.

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

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

- *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 codificant, gli *introni*. In questa fase vengono rimossi e gli *esoni* (sequenze codificant) 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 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.

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.

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

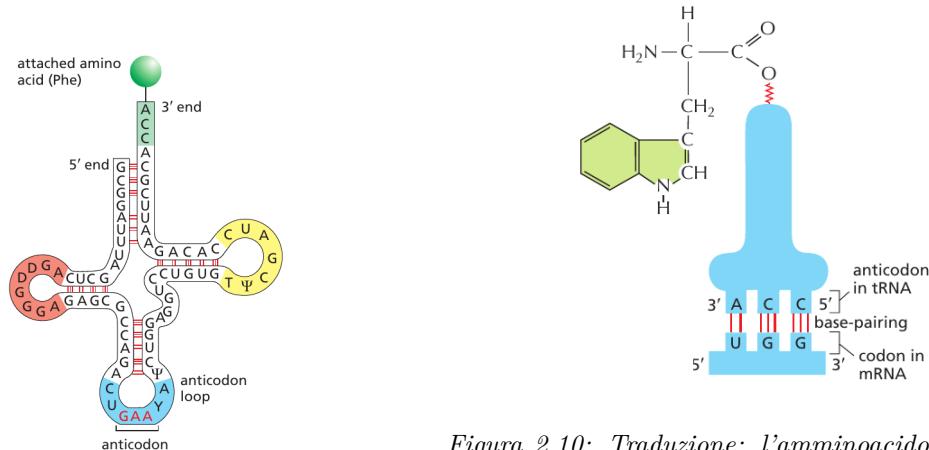


Figura 2.9: tRNA. Fonte [4]

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]

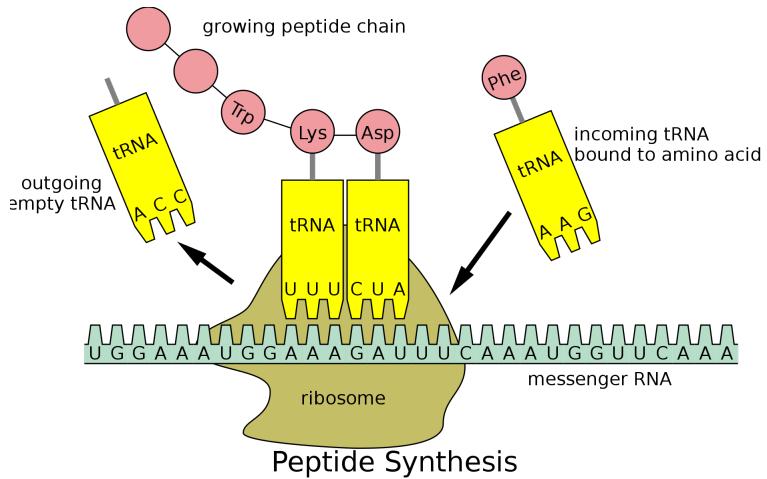


Figura 2.11: Traduzione, sintesi peptidica. Fonte: [13]

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																									
	AGG																									
codons	GCA	CGA																								
	GCC	CGC																								
	GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC	CUG	AAA	AAG	AUG	UUC	CCG	UCG	ACG	CCC	UCA	ACA	GUA	GUC	UAA		
	GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAA	UUG	UUU	CCU	UCU	ACU	UGG	UAC	UCC	ACC	GUC	GUU	UAG		
amino acids	Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop					
	A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V						

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

2.1.4 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 3.1). 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 3.1) 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 so-

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

vrapponibili. Nella grande maggioranza degli organismi viventi le proteine sono composte solo da amminoacidi della serie L.

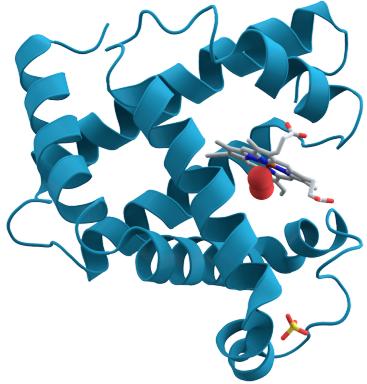


Figura 2.13: Rappresentazione a nastro della struttura tridimensionale della mioglobina. È presente un gruppo hemo al quale è legata una molecola di ossigeno (rossa). Fonte: [14]

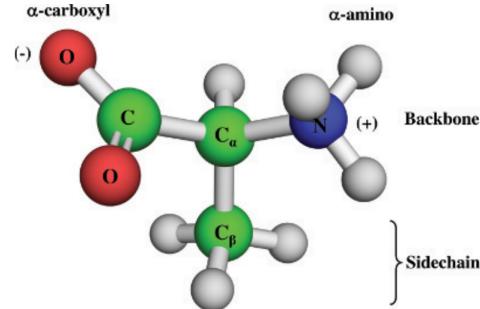


Figura 2.14: Struttura principale degli amminoacidi. Fonte [3]

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*.

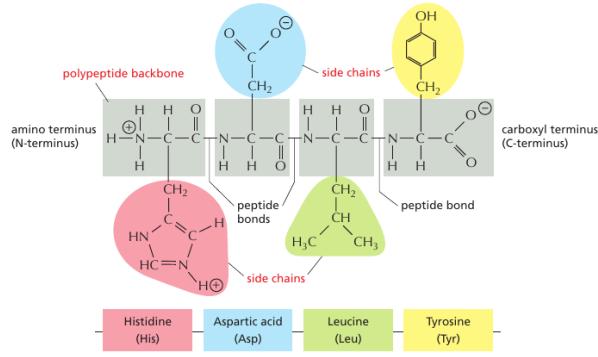


Figura 2.15: Backbone delle proteine. Fonte: [4]

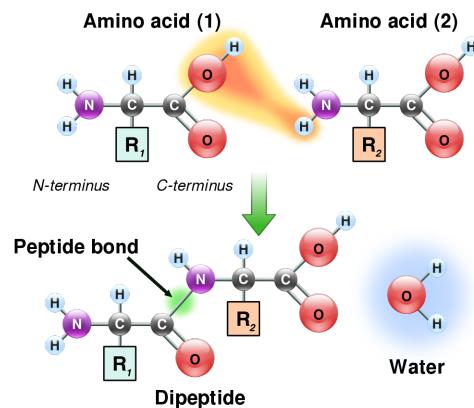


Figura 2.16: Legame peptidico. Fonte [15]

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.

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

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

Oltre agli enzimi ci sono tante altre proteine importanti. Uno degli esempi più noti è l'emoglobina, proteina animale adibita a trasportare ossigeno dai polmoni 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.

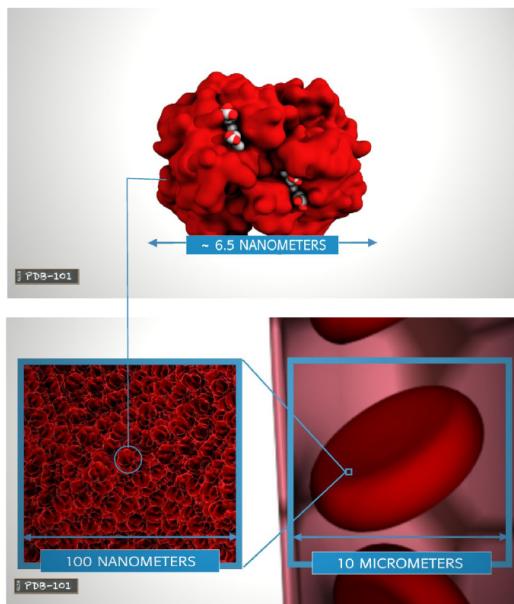


Figura 2.17: Emoglobina in diverse scale. Rapresentazione a superficie. Un globulo rosso contiene circa 280 milioni di molecole di emoglobina, per cui può portare più di 1 miliardo di molecole di ossigeno per volta. Fonte: [16]

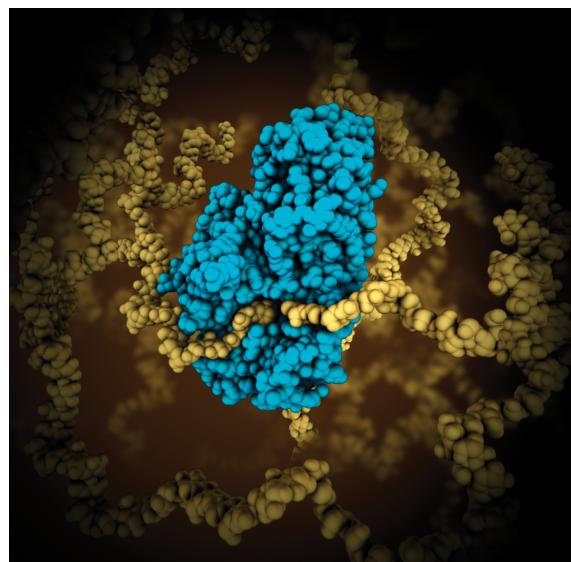


Figura 2.18: Enzima alpha Amilasi in turchese, rappresentazione di tipo space-filling. Si lega a catene di carboidrati (gialle) e le rompe in pezzi più piccoli di glucosio. Fonte [16]

Nelle cellule le proteine svolgono, fra le altre, funzioni di supporto strutturale, mobilità, protezione, regolazione, trasporto, catalisi, magazzino. Nel nostro corpo abbiamo un numero grandissimo di proteine: 10²⁷. Per usare una metafora di Ken Dill^[17] 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^[18] al contrastare l’inquinamento ambientale^[19].

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 proveniente da esperimenti biochimici
- organizzazione e archiviazione conoscenze su genomi e proteomi
- modellizzazione di sistemi e reti biologiche

Come si può notare da questa lista una parte importante della bioinformatica si occupa dell’utilizzo di strumenti informatici finalizzati a manipolare, archiviare e confrontare stringhe e sequenze di caratteri. Tuttavia questa disciplina non si ferma all’analisi delle sequenze. Tra le più interessanti applicazioni bioinformatiche odiere vi sono quelle incentrate sull’analisi strutturale^[20]. 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^[21].

Non va confusa la bioinformatica (o biologia computazionale) con la *computazione bio-inspirata* (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)^[22].

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

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^[23] e vi entra pienamente: la stima di performance in ML è infatti l'accuratezza predittiva stimata dall'errore calcolato sul test set.

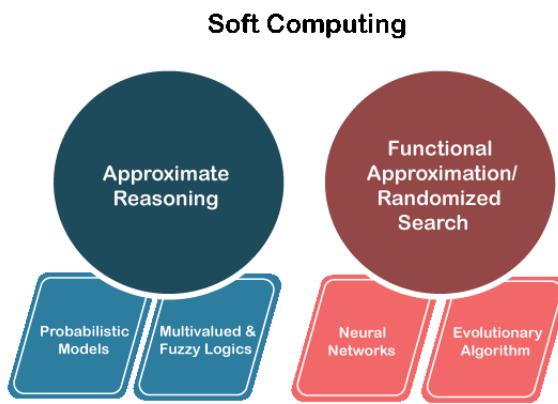


Figura 2.19: Branche del soft computing. Fonte: [24]

2.2.2.1 Algoritmi genetici

Gli algoritmi genetici fanno parte del paradigma relativo alle tecniche informatiche *bio-inspirate*, 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.

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

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

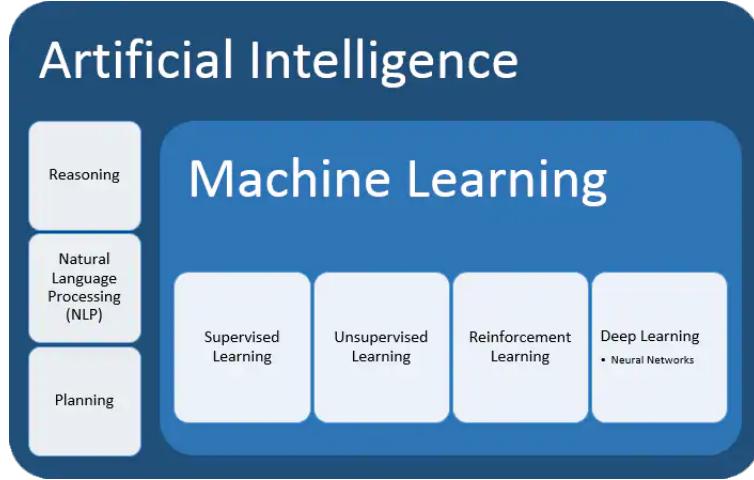


Figura 2.20: Schema riassuntivo dei campi dell'IA. Fonte: [27]

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 gio-

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

care 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.

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 ma tentare di risolvere problemi ingegneristici, dato che i modelli utilizzati sono eccessivamente semplici. Nonostante ciò le ANN riflettono tratti di comportamento del cervello umano e consentono di riconoscere pattern e risolvere problemi difficili.

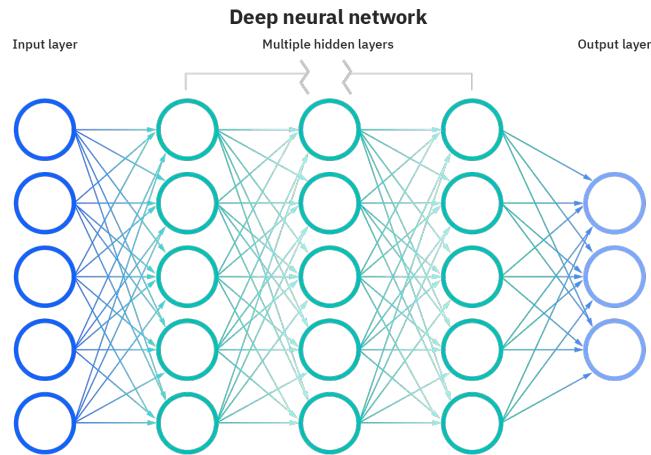


Figura 2.21: Rete neurale artificiale. Fonte: [29]

Le ANN sono composte da almeno 3 strati di nodi: uno strato di input, uno o più nascosti e uno di output. Ogni nodo è un neurone artificiale, si connette agli altri nodi dello strato successivo e ha associato un peso e una soglia. Se l'output di un nodo è sopra la soglia allora il nodo 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 che consiste di più di 3 strati (inclusi quello di input e output) può essere considerata un algoritmo di *deep learning*^[29]. Una rete neurale con 2 o 3 strati è una rete neurale semplice.

Capitolo 3

Protein Folding

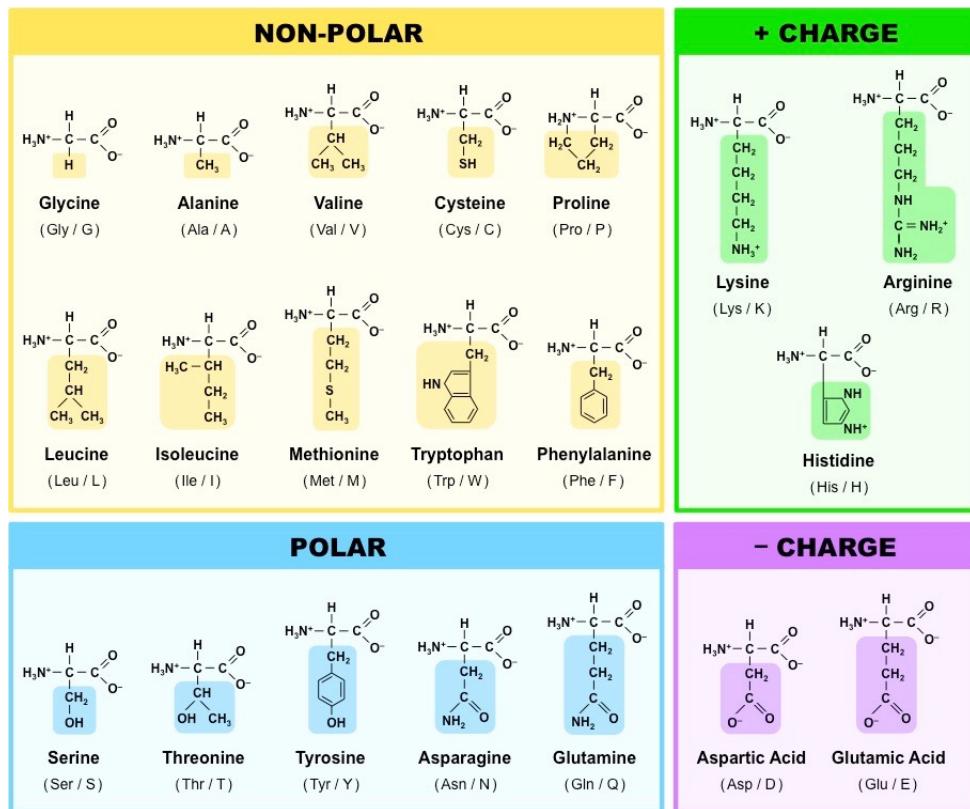


Figura 3.1: I 20 amminoacidi universali. Fonte: [30]

Capitolo 4

Predizione della struttura di proteine

Capitolo 5

AlphaFold

Capitolo 6

Uso di AlphaFold e visualizzazione

Capitolo 7

Scenari aperti e conclusioni

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.
- [20] A. D. Baxevanis, G. D. Bader e D. S. Wishart, *Bioinformatics*. John Wiley & Sons, 2020.
- [28] T. Mitchell, *Machine learning*. McGraw hill New York, 1997.

Articoli

- [18] 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.
- [19] 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.
- [21] 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.
- [22] M. Mitchell, “Biological Computation,” *PDXScholar*, 2010. indirizzo: https://pdxscholar.library.pdx.edu/compsci_fac/2.

Risorse Online

- [1] “enzima nell’Encyclopedia 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%3A_Introductory_Chemistry_\(CK-12\)](https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Book%3A_Introductory_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] “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).
- [17] 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=zm-3kovWpNQ> (visitato il 06/01/2022).

- [23] “Apprendimento automatico - Wikipedia.” (1 dic. 2021), indirizzo: https://it.wikipedia.org/wiki/Apprendimento_automatico (visitato il 23/01/2022).
- [24] “What is soft computing - Javatpoint.” (3 lug. 2021), indirizzo: <https://www.javatpoint.com/what-is-soft-computing> (visitato il 24/01/2022).
- [27] “Machine Learning - IBM.” (29 ago. 2020), indirizzo: <https://www.ibm.com/it-it/analytics/machine-learning> (visitato il 23/01/2022).
- [29] “What are Neural Networks?” (1 Giu. 2021), indirizzo: <https://www.ibm.com/cloud/learn/neural-networks> (visitato il 24/01/2022).
- [30] “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).

Altre fonti

- [5] *Appunti del corso Elementi di Biologia e Neuroscienze, prof. Mario Pirchio, Unipi CdL Filosofia*, 2021.
- [25] R. Kurzweil, R. Richter, R. Kurzweil e M. L. Schneider, *The age of intelligent machines*, 1990.
- [26] H. H. the XIV Dalai Lama, *The heart of the Buddha's path*, 2011.