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

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'inconorazione 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 strutturali [1], 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: $1Da = 1.7 \times 10^{-24}g$. 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.

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

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

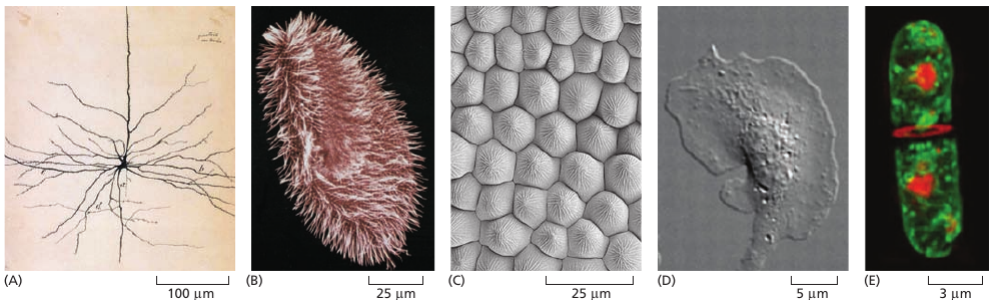

Figura 2.1: (A) disegno di un neurone. (B) Paramecium. (C) superficie di un petalo di fiore di bocca di leone. (D) Macrofago. (E) Un lievito di fissione viene catturato nell'atto di divisione cellulare. Fonte: [4]

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 $1\mu m$. 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 struttura tipica di una cellula animale è mostrata nella seguente figura:

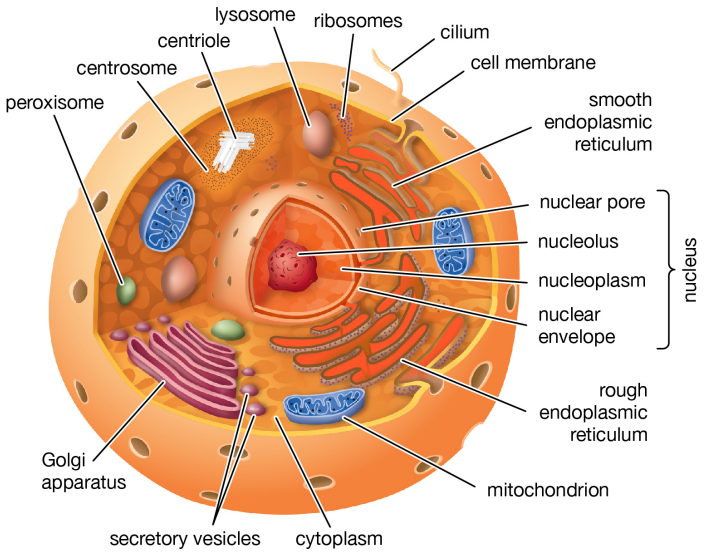

Figura 2.2: Cellula animale. Fonte: [6]

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 permettono lo scambio di sostanze con l'esterno. Tutto ciò che si trova all'interno della cellula è immerso nel citoplasma, gel acquoso contenente grandi e piccole molecole. Il citosol è la parte del citoplasma non contenuta all'interno delle membrane intracelullari. 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[7].

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):

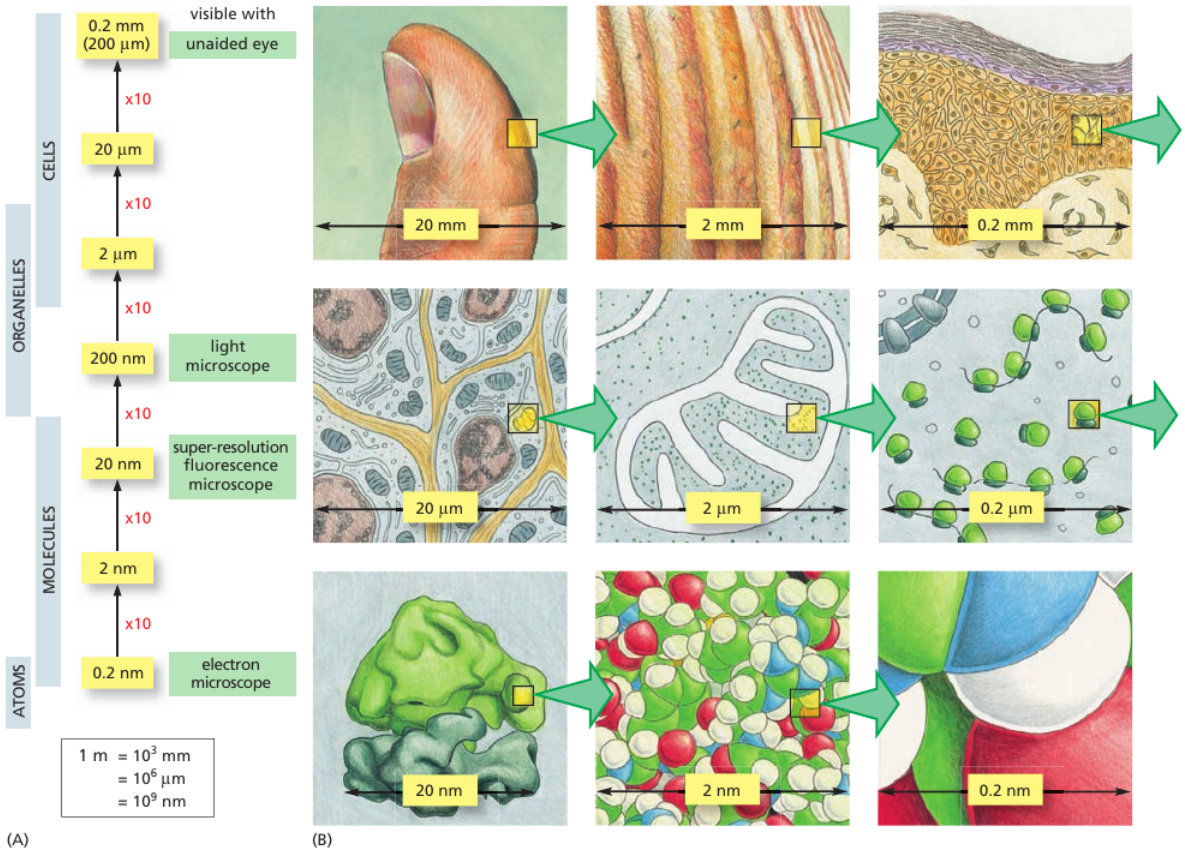

Figura 2.3: (A) Il grafico elenca le dimensioni dei livelli strutturali biologici, le unità di misura relative e gli strumenti necessari per visualizzarli. (B) Uno stesso dettaglio a varie scale di grandezza: pollice, pelle, cellule, mitocondrio, ribosomi, insieme di atomi che formano parte di una proteina. 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 sommergibile 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 triplette) sono universali, gli amminoacidi si condificano nello stesso modo per tutti gli organismi. Diversità e unità della vita sulla Terra sono due facce della stessa medaglia. Il sequenziamento dei genomi e il loro confronto, basato su approcci informatici, ha rivelato una conservazione evoluzionistica, un'eredità comune: è possibile infatti scambiare geni omologhi codificanti proteine del ciclo di divisione cellulare fra uomini e lievito[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[8].

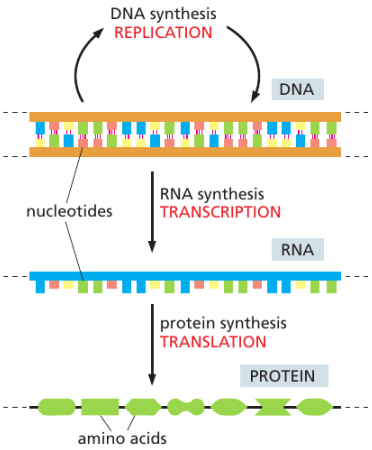

Figura 2.4: Dogma centrale in biologia. Fonte [4]

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* nel DNA) e un gruppo fosfato (vedi figura 2.6). 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):

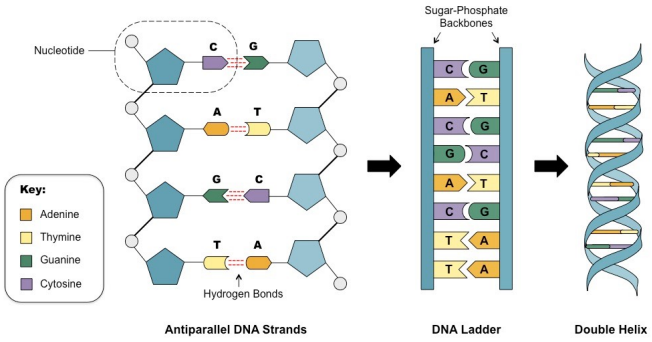

Figura 2.5: struttura del DNA. Fonte: [9]

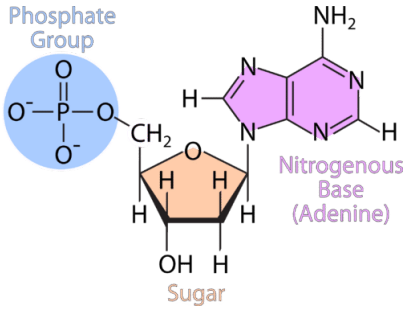

Figura 2.6: Componenti di un nucleotide con Adenina per base azotata. Fonte [10]

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. Ricapitolando: l'informazione contenuta in ogni allele è determinata dalle sequenze delle quattro possibili basi azotate.

L'RNA (*acido ribonucleico*) esiste in varie forme. Le differenze con il DNA sono mostrate nella figura 2.7, 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

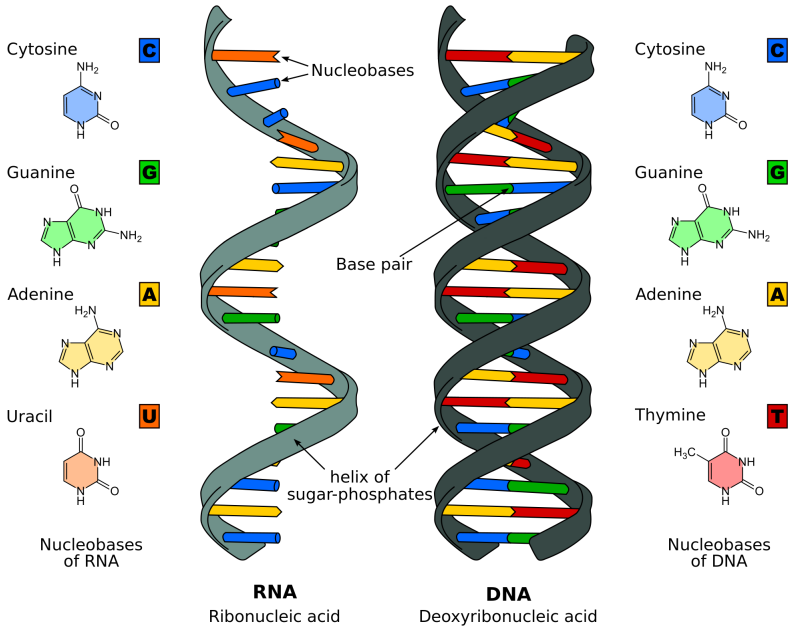

Figura 2.7: Differenze fra RNA e DNA Fonte: [11]

Il DNA dell'uomo contiene 3^9 coppie di nucleotidi: 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.

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 formato 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. Ci sono di conseguenza più tRNA (circa 45).

— terminare traduzione

2.1.4 Proteine: le macromolecole più importanti della vita

Oltre agli enzimi ci sono 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 CO_2 ai polmoni.

Importante funzione degli enzimi 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

Tutti gli enzimi sono proteine, ma non tutti i catalizzatori biologici sono enzimi, dal momento che esistono anche catalizzatori costituiti di RNA, chiamati ribozimi

In un organismo, nonostante tutte le cellule condividano gli stessi geni, cellule afferenti a organi o tessuti diversi esprimono geni differenti (*espressione genica*).

2.2 Background informatico

Background informatico • bioinformatica • database bioinformatici • machine learning • reti neurali, deep learning

Capitolo 3

Protein Folding

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*, 2ª ed. Chapman e Hall/CRC, 2018.

[4] B. Alberts, D. Bray, K. Hopkin et al., *Essential cell biology*, 5ª ed. W. W. Norton e Company, 2019.

[10] S. Bewick, R. Parsons, T. Forsythe, S. Robinson e J. Dupon, *Introductory Chemistry (CK-12)*. Libretexts, 1 giu. 2021. indirizzo: `https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Book%3A_Introductory_Chemistry_(CK-12)` (visitato il 22/01/2022).

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] "Saccharomyces cerevisiae - Wikipedia." (25 set. 2021), indirizzo: `https://it.wikipedia.org/wiki/Saccharomyces_cerevisiae` (visitato il 22/01/2022).

[8] "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).

[9] "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] "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).

Altre fonti

[5] *Appunti del corso Elementi di Biologia e Neuroscienze*, 2021.