



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

LAUREA TRIENNALE IN INGEGNERIA INFORMATICA

Reti neurali convoluzionali per lo studio di varianti non codificanti in sequenze genomiche

LAUREANDO

Alessandro Trigolo

Matricola 2043049

RELATORE

Prof.ssa Cinzia Pizzi

Università degli Studi di Padova

ANNO ACCADEMICO
2023/2024

Sommario

PROVA3 Questo elaborato mira ad approfondire il funzionamento delle reti neurali convoluzionali e di come questi modelli di deep learning siano in grado di estrarre significative informazioni da sequenze genomiche, analizzandone le zone non codificanti. In particolare, verranno comparati tre tool basati sulle CNN — DeepSEA, Basset e DeepSATA — e sarà fornita una revisione delle loro prestazioni.

Abstract

PROVA3 This thesis aims to deepen the understanding of the functioning of convolutional neural networks and how these deep learning models are able to extract significant information from genomic sequences, analyzing their non-coding regions. In particular, three CNN-based tools — DeepSEA, Basset and DeepSATA — will be compared and a review of their performance will be provided.

Indice

1	Introduzione	1
2	Background biologico	3
2.1	Dogma centrale	5
2.2	Varianti non codificanti	9
3	Reti neurali	11
3.1	Principi di base ed evoluzione	11
3.2	Reti neurali convoluzionali	11
4	Reti convoluzionali e varianti non codificanti	15
4.1	DeepSEA	15
4.2	Basset	15
4.3	DeepSATA	15
5	Discussione	17
6	Conclusioni	19
	Bibliografia	21

Indice delle Figure

2.1	Rappresentazione schematica della cellula eucariote.	3
2.2	Rappresentazione schematica del DNA.	4
2.3	Il processo di impacchettamento del DNA.	5
2.4	Il processo di trascrizione del DNA in RNA.	6
2.5	Il processo di traduzione da mRNA a polipeptide.	7
2.6	La mitosi cellulare.	8
2.7	Il processo di replicazione del DNA.	9

Indice delle Tabelle

Lista degli Acronimi

AI Intelligenza artificiale — *Artificial Intelligence*

ATP Adenosintrifosfato — *Adenosine TriPhosphate*

CNN Rete neurale convoluzionale — *Convolutional Neural Network*

DNA Acido desossiribonucleico — *DeoxyriboNucleic Acid*

RNA Acido ribonucleico — *RiboNucleic Acid*

1

Introduzione

Ad oggi l'avanzamento della genomica — ramo della biologia molecolare che si occupa di studiare il genoma degli esseri viventi — si è rivelato notevolmente significativo al fine di approfondire e comprendere malattie legate alle mutazioni del genoma degli individui. Si stima che solamente una percentuale tra l'1% e il 2% del DNA contiene i *geni*, ovvero particolari regioni che contengono tutte le informazioni necessarie per la sintesi degli aminoacidi che poi comporranno le proteine [1], [2]. Ciò nonostante, la quasi totalità dei disturbi genomici è dovuta alle mutazioni nelle regioni non codificanti [3] — dette *varianti non codificanti*. Le mutazioni in queste zone del genoma, che apparentemente svolgono funzioni marginali, sono responsabili dello sviluppo di disturbi importanti, come le *malattie mendeliane*¹, l'epilessia, malattie cardiovascolari e soprattutto tumori — tra cui il cancro del colon-retto e il tumore al seno [3]–[11]. Risulta quindi vitale continuare a studiare gli effetti che le varianti non codificanti in sequenze genomiche hanno sugli individui.

Negli ultimi decenni, il progredire delle tecniche di *sequenziamento* [12] ha dato uno slancio rilevante allo sviluppo della *bioinformatica* — disciplina che unisce informatica e biologia. La bioinformatica si interessa a organizzare dati biologici in modo tale da facilitare l'accesso e l'inserimento di nuove informazioni (come il *PDB* [13]), sviluppare *tool* che permettono l'analisi dei dati e infine fornire una interpretazione significativa dei risultati ottenuti [14]. Più recentemente, l'accrescimento dei dati biologici e il costante avanzamento della potenza di calcolo hanno reso possibile l'applicazione di tecniche di *deep learning* anche nel campo della bioinformatica. Questo notevole progresso consente di scoprire e perfezionare soluzioni informatiche che permettano di delineare con sempre maggior precisione il ruolo che hanno le mutazioni nelle regioni non codificanti del DNA. Grazie a queste nuove tecnologie, la *genomica funzionale* — area della genomica che si interessa a descrivere le relazioni che ci sono tra i componenti di un sistema biologico, come geni e proteine [15] — ha avuto un forte impulso nell'approfondire le varianti non codificanti, tuttavia rimangono ancora significative lacune

¹Le malattie mendeliane, causate dalla mutazione di un singolo gene, includono la fibrosi cistica e il morbo di Huntington.

nella comprensione della relazione tra mutazioni genetiche ed espressione genica. L'utilizzo di tecniche di deep learning risulta quindi cruciale per continuare la ricerca in questo ambito. L'obiettivo di questo elaborato è di discutere e confrontare tre tool che utilizzano le *reti neurali convoluzionali* per predire l'effetto delle varianti non codificanti su sequenze genomiche: DeepSEA [16], Basset [17] e DeepSATA [18].

Più precisamente, il Capitolo 2 introdurrà le basi della biologia molecolare, necessarie per comprendere interamente l'importanza delle varianti non codificanti. Successivamente, nel Capitolo 3 saranno approfonditi i principi fondamentali delle reti neurali e di come le reti convoluzionali possono essere utilizzate come ottimo strumento per predire l'effetto di sequenze genomiche. Il Capitolo 4 invece esaminerà i dettagli implementativi di ciascuno dei tre tool, indagando principalmente sugli aspetti legati alla codifica delle sequenze, alla struttura della rete e al *dataset* utilizzato per allenare il modello. Infine nel Capitolo 5 si riassumono le differenze analizzate nel capitolo precedente, offrendo una visione complessiva del confronto tra i tre tool.

Background biologico

La cellula è l'unità fondamentale della vita. La cellula è una piccola miscela acquosa con componenti chimici, racchiusi in una membrana, e possiede l'eccezionale capacità di replicarsi. Il primo elemento che permette di distinguere le cellule è la presenza di un nucleo. Vengono definite *procarioti* le cellule senza nucleo — che sono le più diffuse e compongono organismi unicellulari come i batteri e gli archei — mentre sono chiamate *eucarioti* le cellule che contengono un nucleo — le quali sono in genere più grandi e più complesse e costituiscono forme di vita multicellulari come animali, piante e funghi [19].

All'interno della cellula eucariote (Figura 2.1), immersi nel *citoplasma*, sono presenti diversi *organuli*, i quali svolgono una particolare funzione ciascuno. I *mitocondri* sono gli organuli più diffusi. Il loro compito è quello di generare energia chimica per la cellula: attraverso il processo di ossidazione di zuccheri e grassi, viene creata una sostanza che viene utilizzata nella



Figura 2.1: Rappresentazione schematica della cellula eucariote; si possono notare i principali organuli tra cui i mitocondri, lisosomi e perossisomi, il reticolo endoplasmatico, e il nucleo [2].

maggior parte delle attività cellulari¹; questo processo è anche chiamato *respirazione cellulare* perché consumando l'ossigeno viene rilasciata anidride carbonica. Oltre ad essere la fonte energetica primaria della cellula, i mitocondri hanno anche importanti ruoli nella regolazione del metabolismo, del ciclo cellulare, delle risposte antivirali e anche della morte della cellula [19]–[21].

Il *reticolo endoplasmatico* è invece un organulo molto esteso e svolge molteplici funzioni. Tra questi compiti rientrano quelli di traslocazione di proteine e il ripiegamento delle proteine (*protein folding*) [19], [22]. I *lisosomi* si occupano di degradare e riciclare gli scarti cellulari e giocano un ruolo fondamentale per l'omeostasi della cellula², il suo sviluppo e il suo invecchiamento [23]–[25]. Infine, i *perossisomi* sono delle piccole vescicole che forniscono un ambiente protetto per gestire molecole tossiche come gli acidi grassi i quali sono smaltiti tramite la β -ossidazione [19], [26].

L'organulo più importante della cellula rimane il *nucleo*. Racchiuso nell'*involucro nucleare*, all'interno di questo organulo sono presenti tutte le informazioni genetiche, racchiuse in una lunga molecola di acido desossiribonucleico (comunemente noto come DNA), che, una volta impacchettato forma il *cromosoma* [2], [19]. La molecola di DNA è una struttura a doppia elica formata da *nucleotidi*. Osservando la Figura 2.2, i nucleotidi sono composti a loro volta da tre elementi fondamentali: una *base azotata*, uno *zucchero* e un *gruppo fosfato*³. Le basi azotate

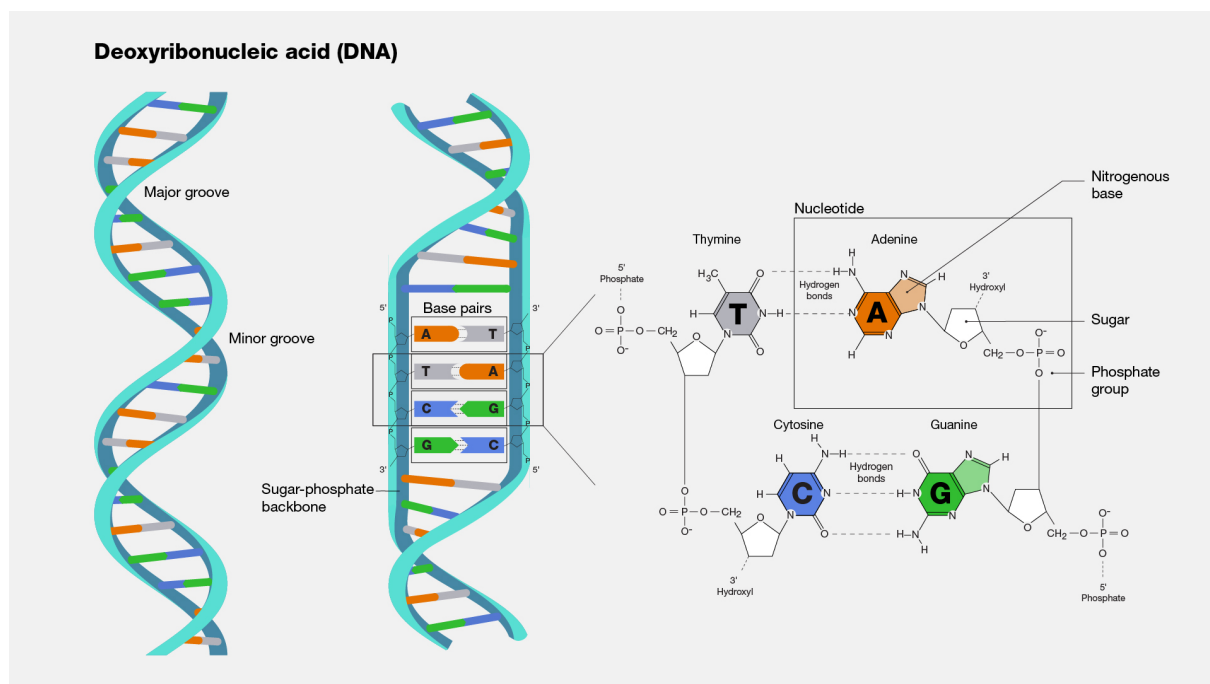


Figura 2.2: Rappresentazione schematica del DNA in cui si possono osservare le coppie di basi azotate, legate tra loro attraverso gli zuccheri e i gruppi fosfati [27].

¹Questa sostanza è detta *adenosintrifosfato* o ATP ed ha una struttura simile ad un nucleotide: è infatti composta dall'Adenina, da uno zucchero e da tre gruppi fosfati.

²Con omeostasi cellulare si intende l'insieme di meccanismi necessari per mantenere ad un livello ottimale le funzioni della cellula.

³I gruppi fosfati hanno una carica negativa e forniscono alla molecola le proprietà di un acido.

sono quattro — Adenina (A), Citosina (C), Guanina (G) e Timina (T) — e si uniscono tra loro mediante dei legami ad idrogeno e secondo un preciso criterio: l'Adenina si lega solamente con la Timina (formando il legame *AT*) mentre la Citosina si unisce solo con la Guanina (creando la coppia *CG*) [1], [28]. Si osserva infine che il nucleotide di una coppia e quello successivo si legano mediante zucchero e gruppo fosfato sempre allo stesso modo: il gruppo fosfato di un nucleotide si lega sempre allo zucchero dell'altro. Di conseguenza, preso un filamento della doppia elica, le due estremità non sono uguali in quanto una termina con un gruppo fosfato (terminazione 5') e l'altra con uno zucchero (terminazione 3').

Attraverso una serie di ripiegamenti, una molecola di DNA lunga circa due metri riesce a raggomitolarsi in un cromosoma di grandezza inferiore a 2 micron (Figura 2.3). Il processo di *DNA-packaging* inizia avvolgendo la doppia elica di DNA attorno a delle proteine dette *istoni* e formando dei *nucleosomi*. In secondo luogo i nucleosomi si ammassano vicini tra loro formando una fibra, chiamata *cromatina* che, a sua volta si impacchetta su se stessa creando il cromosoma [29], [30].

2.1 DOGMA CENTRALE

La rilevanza del DNA è data dalle informazioni essenziali che questa molecola contiene. Tali informazioni risiedono nei geni, che sono delle sequenze genomiche che codificano uno o più prodotti biologici operativi [32]. L'*espressione genica* è il processo che permette di utilizzare i dati contenuti nel gene per la creazione di macromolecole, come le proteine. Per esempio, le cellule della pelle a contatto con luce solare intensa possono esprimere geni che regolano la pigmentazione della pelle [33]. L'espressione genica è divisa in due fasi principali: la *trascrizione*



Figura 2.3: Il processo di impacchettamento del DNA che permette di compattare la struttura a doppia elica nel cromosoma [31].

zione — che si occupa di produrre delle molecole di RNA che rispecchino il gene da esprimere — e la *traduzione* — la quale traduce le informazioni dell'RNA sintetizzando la proteina.

Nella prima fase dell'espressione genica, è necessario trascrivere il DNA in una molecola molto simile ovvero l'RNA — chiamato anche acido ribonucleico. Questa molecola differisce dall'acido desossiribonucleico per una base azotata — anziché la Timina è presente l'Uracile (U) — e per lo zucchero — da desossiribosio a ribosio [34]. La trascrizione del DNA in RNA inizia quando delle proteine, chiamate *fattori di trascrizione*, attratte dagli *enhancer* del DNA, riconoscono la regione che delimita l'inizio della molecola del gene da esprimere, detta *zona promotrice*. Dopo aver riconosciuto l'inizio della sequenza, queste proteine permettono ad un enzima chiamato *RNA polimerasi* di attaccarsi ed aprire la doppia elica del DNA [35]. Una volta aperta la doppia elica, inizia la vera e propria trascrizione in RNA: il filamento del DNA viene preso come modello per la creazione dell'RNA; in particolare il nucleotide dell'RNA sarà il complementare rispetto a quello del DNA (di conseguenza $A \rightarrow U$, $C \rightarrow G$, $G \rightarrow C$ e $T \rightarrow A$). Così facendo l'acido ribonucleico viene creato un nucleotide alla volta, analizzando quello del DNA [34]. La trascrizione termina nel momento in cui gli enzimi e le proteine incontrano la regione terminatrice del gene che determina la separazione dal filamento e la terminazione dell'RNA *messaggero* (*mRNA*) che contiene le informazioni presenti nel gene da esprimere. L'intero processo di trascrizione è illustrato nella Figura 2.4.

Prima di uscire dal nucleo l'RNA messaggero subisce una serie di elaborazioni necessarie per rendere le informazioni immagazzinate sicure: diverse sono le malattie che emergono per mutazioni presenti nell'mRNA tra cui la distrofia miotonica⁴ [37]. La prima elaborazione viene



Figura 2.4: Il processo di trascrizione del DNA del gene in RNA mediante la RNA polimerasi [36].

⁴Le distrofie miotoniche sono patologie che colpiscono principalmente l'apparato muscolo-scheletrico.

chiamata *5'-end capping* e si occupa di aggiungere alla terminazione 5' dell'mRNA una Guanine attraverso un collegamento inusuale che garantisce maggiore stabilità alla molecola. In secondo luogo avviene lo *splicing* che si occupa di rimuovere le zone non codificanti — dette *intron*i — dal gene trascritto mantenendo solo quelle che verranno utilizzate per essere sintetizzate in proteine — gli *esoni* — e quindi facilitando il processo di traduzione. Infine con il *3'-end processing* viene aggiunta alla terminazione 3' dell'mRNA una coda di Adenine — detta anche *polyA tail* — che, in maniera molto simile al *5'-end capping* garantisce una stabilità del filamento di acido ribonucleico [38], [39].

Dopo essere uscito dal nucleo attraverso i *pori*, l'RNA messaggero raggiunge il citoplasma ed è pronto per iniziare la seconda fase dell'espressione genica, la traduzione. La traduzione non è altro che la traduzione dell'mRNA in un *polipeptide*, ovvero una sequenza di aminoacidi che compongono la proteina. Gli aminoacidi sono più di 20, di conseguenza anziché codificare un solo nucleotide dell'RNA messaggero, vengono codificati tre nucleotidi alla volta: questa tripletta viene chiamata *codone*. Durante la fase della traduzione, giocano un ruolo fondamentale i *ribosomi* i quali sono degli organuli nei quali avviene la traduzione. I ribosomi sono composti da due sotto unità, ciascuna delle quali ha tre siti per l'RNA di *trasporto* (*tRNA*). Delle due sotto unità del ribosoma, quella dimensionalmente minore si lega all'mRNA e agli *anticodoni* (sequenze specifiche di tre basi nel tRNA) e controlla che la traduzione avvenga con successo. La sotto unità più voluminosa invece si prende carico di catalizzare il legame peptidico tra l'aminoacido trasportato dal tRNA e la catena di aminoacidi in crescita [39]–[41]. In questo modo i ribosomi, analizzando codone dopo codone riescono a creare la catena polipeptidica mediante l'RNA di trasporto, come mostrato nella Figura 2.5.

Una volta creata la sequenza polipeptidica, inizia il processo di ripiegamento della proteina.

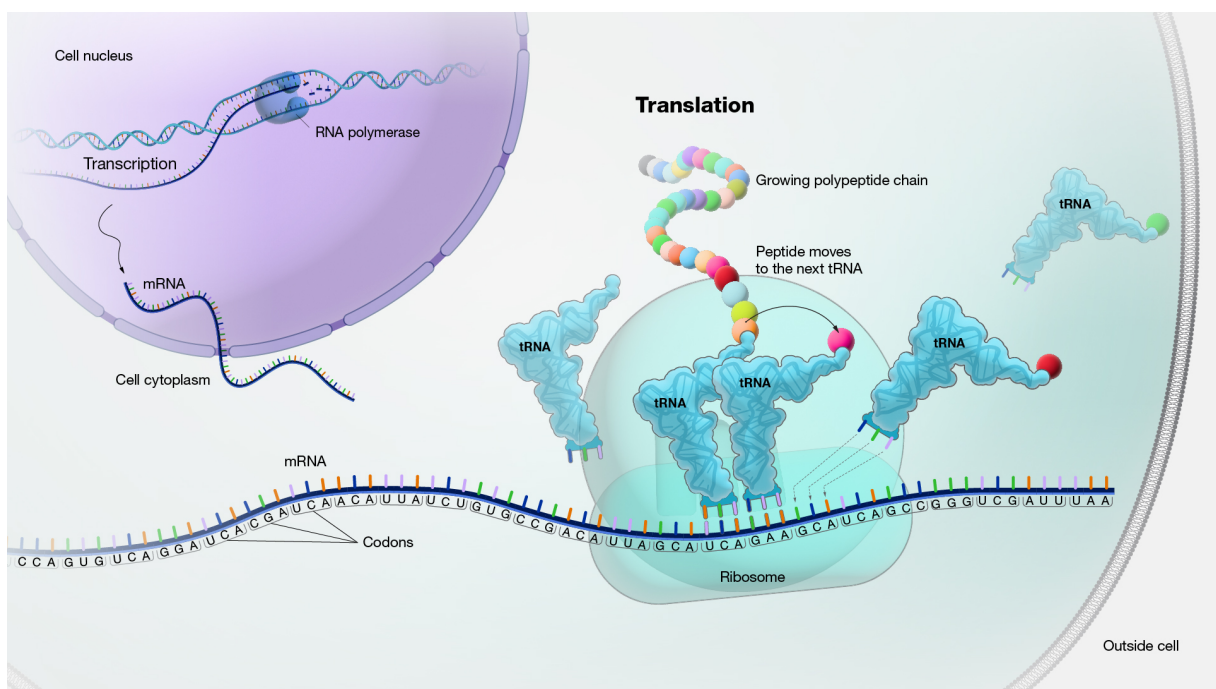


Figura 2.5: Il processo di traduzione da RNA messaggero a polipeptide attraverso il tRNA e i ribosomi [42].

In maniera molto simile a quanto visto per l'impacchettamento del DNA nel cromosoma, la sequenza di polipeptidi inizialmente si arrotola creando delle bobine che sono comunemente chiamate α -*helix*. Queste ultime poi si ripiegano nuovamente arrivando alla struttura terziaria della proteina, ovvero la proteina tridimensionale effettiva [43]. Una volta creata la proteina il gene è stato espresso definitivamente. Questo passaggio di informazioni dal DNA alla creazione della proteina è gergo definito come il *dogma della biologia molecolare*.

Come accennato all'inizio del capitolo, la cellula possiede la notevole capacità di replicarsi. In genere una cellula si duplica durante la crescita e lo sviluppo dell'organismo, quando deve essere rimpiazzata o rigenerata oppure nella riproduzione asessuata di alcuni micro organismi [44]. Il processo replicazione cellulare, chiamato *mitosi*, è preceduto dall'*interfase*, processo fondamentale in cui la cellula cresce di dimensioni e il DNA nei cromosomi si duplica, favorendo la replicazione cellulare. La mitosi può essere suddivisa in quattro fasi principali [44]–[47] le quali sono riassunte anche nella Figura 2.6:

1. Nella *profase* i cromosomi duplicati si condensano nel nucleo ed iniziano ad avvicinarsi dei microtuboli al nucleo, chiamati *centrosomi*; allo stesso tempo la membrana nucleare inizia a svanire;
2. Dopo che i microtuboli si sono attaccati ai cromosomi (fase intermedia detta *prometafase*) si giunge alla *metafase*, situazione in cui tutti i cromosomi sono allineati lungo la linea equatoriale della cellula;
3. Durante l'*anafase*, ciascuna coppia di cromosomi si divide e raggiunge i poli della cellula;
4. La fase finale della mitosi è la *telofase* nella quale le due cellule si dividono; le membrane nucleari delle due cellule si riformano attorno ai cromosomi divisi.

Affinché la mitosi abbia successo, è prima necessario duplicare il DNA all'interno della cellula. Il processo di replicazione di DNA che precede la mitosi è anche definito come *fase*

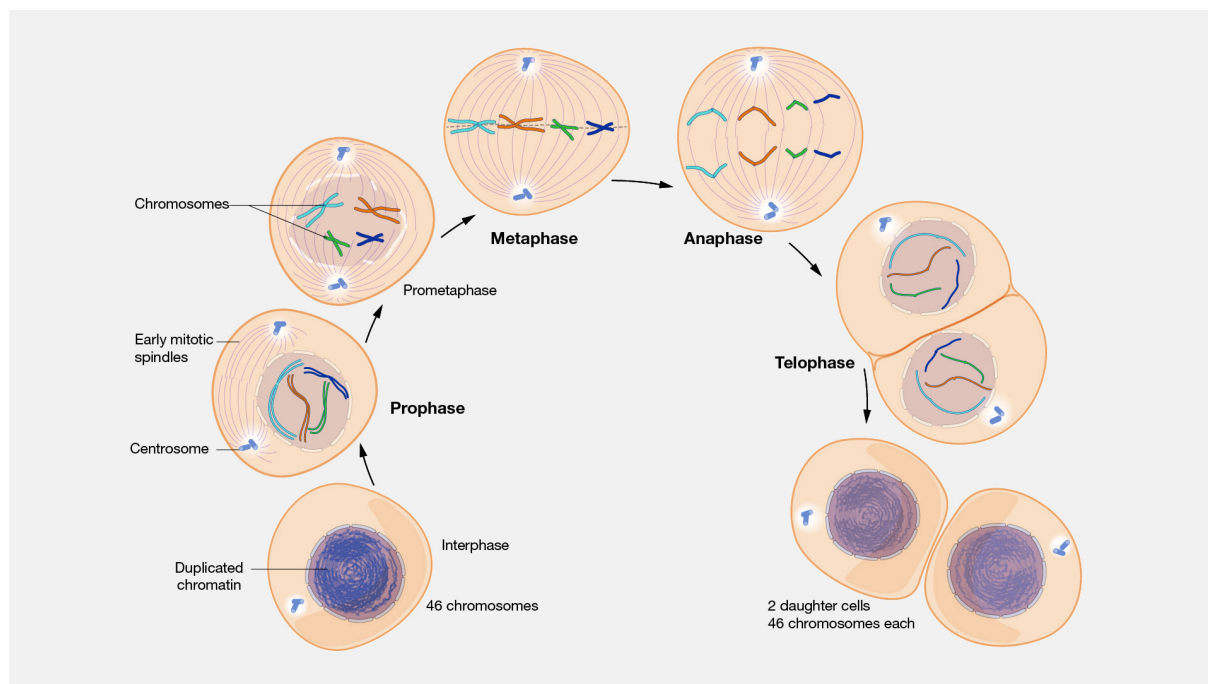


Figura 2.6: Rappresentazione delle quattro fasi che comprendono la mitosi cellulare [48].

di sintesi — *S-phase*. La duplicazione del DNA inizia con l'identificazione dell'*origine della replicazione*, ovvero una sequenza del DNA che specifica da quale punto della sequenza il DNA deve essere replicato (ci sono più di cento mila siti che segnalano un punto di origine nel DNA di una cellula). Una *proteina iniziatrice* è legata al punto di origine promuovendo l'attaccamento al DNA del *replisoma* che è composto da un enzima chiamato *elicasi* che si occupa di dividere i due filamenti di DNA procedendo nella direzione $5' \rightarrow 3'$. A questo punto il *RNA prime* inizia la sintesi del DNA favorendo l'attaccamento della *DNA polimerasi* entrambi i filamenti per duplicare il DNA. Essendo che il genoma è complementare, un filamento avrà un verso $5' \rightarrow 3'$ (*leading strand*) mentre l'altro filamento avrà verso opposto, $3' \rightarrow 5'$ (*lagging strand*). Di conseguenza, nel filamento concorde al replisoma, la polimerasi non incontrerà problemi nella duplicazione, invece nel filamento $3' \rightarrow 5'$ il DNA dovrà essere duplicato a segmenti, detti *frammenti di Okazaki* che verranno collegati tramite la *DNA ligasi* [49]–[52]. La Figura 2.7 racchiude quanto descritto sulla fase di sintesi.

2.2 VARIANTI NON CODIFICANTI

Come descritto fino ad ora, il ruolo del DNA è fondamentale in quanto trasmesso da cellula a cellula durante la replicazione per poi essere utilizzato nell'espressione genica creando le proteine. Alle regioni del DNA che prendono parte al processo di espressione di un gene, si contrappongono le regioni che non vengono codificate, come le zone intensificatrici e promotrici del DNA e gli introni dei geni. Anche se queste sequenze non vengono effettivamente espresse, ricoprono un ruolo fondamentale nell'espressione genica (Figura 2.4 e Figura 2.5) e durante la duplicazione del DNA (Figura 2.7). Proprio per questo motivo le mutazioni in queste



Figura 2.7: Il processo di replicazione del DNA durante la fase di sintesi [53].

2.2. VARIANTI NON CODIFICANTI

zone possono indurre a disturbi genetici critici. Ciò nonostante non si è ancora in grado di comprendere a fondo gli effetti che queste particolari mutazioni hanno. Risulta quindi importante continuare a studiare le varianti non codificanti e gli effetti collaterali genetici che provocano.

spiega in maniera più approfondita la varianti non codificanti

3

Reti neurali

3.1 PRINCIPI DI BASE ED EVOLUZIONE

3.2 RETI NEURALI CONVOLUZIONALI

VIDEO ENKK

Nel 1950 Alan Turing pubblica "Computing machinery and intelligence" dove viene introdotto il concetto di macchina intelligente. Viene introdotto anche il test di Turing che è un test che determina se una macchina è più o meno capace di imitare l'uomo. Nel 1956, la conferenza di Dartmouth (c'è Shannon, Minsky che copione il temine intelligenza artificiale). Gli obiettivi erano quelli di machine translation. Rosenblatt negli anni 60 inventa il perceptrone, rete neurale semplicissima. Agli anni 60 seguono due risultati allo stop dell'IA, chiamati inverno della intelligenza artificiale. Minsky pubblica un libro, chiamato perceptrons dove viene dimostrato che il perceptrone è in grado di risolvere solo problemi semplici, letteralmente separabili. Esce un report, ALPAC, dove viene osservato che nemmeno la task di traduzione funziona.

Dagli anni 80, viene inventata la backpropagation dove le reti neurali diventano efficaci. Qui nascono anche i sistemi esperti, che sono dei sistemi che, per esempio, si basano su delle regole logiche: qualcosa di più dettagliato ed un esperto inietta dentro la sua conoscenza. Nel 1996/1997 Gary Kasparov sfida DeepBLUE, progettato da IBM. La prima sfida, Kasparov vince il match overall ma perde due partite. Nella seconda edizione del '97, DeepBLUE vince tutto. Per la prima volta, un esperto viene sconfitto da una macchina. Negli anni 2000 arriva il machine learning: apprendimento automatico. Una macchina che apprende dai dati. Reti neurali sono un sottoinsieme del ML, il deep learning sono un sottoinsieme di rete neurale (LLM sono un sottoinsieme di deep NN).

LIBRO

Il primo modello di intelligenza artificiale tento di creare il funzionamento di un neurone (ce la foto se vuoi metterla). I primi modelli più semplici furono delle semplici input-output function. IN seguito le funzioni divennero più complicate, più livelli furono aggiunti e feedback bidirezionale fino ad arrivare ai modelli odierni di DL. La prima pubblicazione sulla AI, fu del 43 ([54]) che descrive un modello di computer che impara attraverso un processo paragonabile a quello dei neuroni: vengono quindi introdotti i neuroni MCP che prendono il nome dai due creatori (McCulloch and Pitts). Questo neurone prendeva in input delle variabili booleane, processarle attraverso una funzione scelta a priori e poi, se il risultato superava una certa threshold, l'output avrà un valore. Quando la threshold è superata si dice che il neurone si attiva. Aveva molte limitazioni, generava un output binario e richiedeva un numero fisso di pesi.

Un metodo più sofisticato venne introdotto nel 1958 da Rosenblatt, chiamato perceptrone il quale processava input non booleani e pesi per bilanciare. Una funzione non lineare processa la somma dei prodotti degli input e dei pesi che rende il modello più flessibile: questo sarà la base per le reti neurali.

Uno dei primi articoli che introducono una metodologia per testare l'intelligenza di un modello fu il test di Turing nel 1950. Questo test, chiamato Test di Turing, si domandava se una macchina fosse in grado di imitare l'intelligenza umana ([55, Computing machinery and intelligence]). Il test prende in causa un interrogatore umano che fa la stessa domanda a due ascoltatori, una persona ed una macchina. Se l'interrogatore non riesce a distinguere l'uomo dalla macchina allora la macchina ha superato il test di Turing. Questo test è stato da sempre l'obiettivo dell'intelligenza artificiale anche se ad oggi alcuni dubbi stanno sorgendo. La conferenza del 1956 a Dartmouth (organizzata da Marvin Minsky, John McCarthy, Claude Shannon e Nathan Rochester) è considerata il momento in cui l'AI è stata globalmente riconosciuta.

Il perceptrone di Rosenblatt aveva fatto molto clamore ma le aspettative dell'opinione pubblica non furono incontrate. Le aspettative del pubblico non furono assolutamente incontrate, tanto che nel '68 un paper di Minsky e Papert dimostrò le forti limitazioni del perceptrone, dove di fatto veniva descritto che il perceptrone non sarebbe stato in grado di emulare una operazione di XOR esclusivo con due input (non erano linearmente separabili). Nel '73 una seconda pubblicazione di Lighthill sottolinea le forti aspettative del pubblico e gli scarsi risultati ottenuti dal progresso. Dal '74 all'80 ci fu il primo inverno dell'intelligenza artificiale dove essenzialmente nulla accadde e tutti persero l'interesse al riordinamento (controlla bene le fonti che qui me la sono un po' inventata ecco letsgoski).

Nel 1985 l'inverno dell'AI finì quando venne introdotto il Gradient Descent Optimization per minimizzare l'errore in una rete (Rumelhart, Hinton and Williams). Inoltre nel '86 Rumelhart e colleghi espansero il lavoro introducendo il concetto di back-propagation nelle reti neurali multilivello (molti livelli di neuroni attaccati tra loro). Questo algoritmo rivoluzionò le capacità di imparare delle reti. Il secondo inverno dell'AI iniziò negli anni 90, dove ci si rese conto che queste reti neurali non erano scalabili: c'era troppa poca potenza di calcolo.

Questo inverno venne causato dalle aspettative sulle capacità delle reti neurali che però non andavano di pari passo con lo sviluppo della potenza computazionale. Per questo motivo i ricercatori si spostarono su altri algoritmi, che richiedevano meno potenza di calcolo come le support vector machines (1963, Vapnik and Chervonenkis). Quando questi algoritmi furono implementati in kernel non lineari nel 92 (kernel trick) si fu in grado di risolvere iperpiano non lineari senza la necessità di requisiti computazionali elevati.

Quando alla metà degli anni 90 la potenza computazionale crebbe, l'attenzione per la AI aumentò di conseguenza. Nel '97, IBM sviluppò un super computer, chiamato Deep Blue, che sconfisse Kasparov a scacchi. Questo evento rese le reti neurali risolverse, con l'introduzione delle reti neurali convoluzionali.

Nel '98 venne pubblicata LeNet-5, che è una rete convoluzionale a sette livelli. Viene usata la convoluzione per fare subsampling per poi passare ai livelli fully connected per predire l'output. Anche in questo caso il modello era difficile da far scalare a causa di hardware e data constraints. Questo problema continuò fino all'ultimo decennio, dove due grandi avanzamenti furono cruciali: lo storage dei dati e la GPU (graphical processing unit). I dati furono più accessibili e con meno costi. Con un alto livello di dati disponibili, la performance degli algoritmi di machine learning aumentò.

Questo introdusse il deep learning, nel 2006 sbloccando una performance eccellente nella speech recognition

[56]

Articoli: [54]

quando fai CNN dici che ML fa cagare per questi

4

Reti convoluzionali e varianti non codificanti

4.1 DEEPSEA

4.2 BASSET

4.3 DEEPSATA



Discussione

Tabella che specifica e riassume per ogni tool encoding, dataset etc

riassume quanto analizzato prima, riporta i risultati del paper piu recente in modo da avere un momento in cui riassumo la situa

Sperimentalmente, per risorse a disposizione, per il confronto ci si basa sui risultati dell'ultimo paper



Conclusioni

Bibliografia

- [1] S. S. Sahu e G. Panda, «Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach,» *Genomics, Proteomics and Bioinformatics*, vol. 9, n. 1-2, pp. 45–55, 2011.
- [2] T. D. Pollard, W. C. Earnshaw, J. Lippincott-Schwartz e G. Johnson, *Cell Biology E-Book: Cell Biology E-Book*. Elsevier Health Sciences, 2022.
- [3] F. Zhang e J. R. Lupski, «Non-coding genetic variants in human disease,» *Human molecular genetics*, vol. 24, n. R1, R102–R110, 2015.
- [4] J. French e S. Edwards, «The role of noncoding variants in heritable disease,» *Trends in Genetics*, vol. 36, n. 11, pp. 880–891, 2020.
- [5] H. Chial, «Mendelian genetics: patterns of inheritance and single-gene disorders,» *Nature Education*, vol. 1, n. 1, p. 63, 2008.
- [6] S. Pagni, J. D. Mills, A. Frankish, J. M. Mudge e S. M. Sisodiya, «Non-coding regulatory elements: Potential roles in disease and the case of epilepsy,» *Neuropathology and Applied Neurobiology*, vol. 48, n. 3, e12775, 2022.
- [7] A. Kapoor et al., «An enhancer polymorphism at the cardiomyocyte intercalated disc protein NOS1AP locus is a major regulator of the QT interval,» *The American Journal of Human Genetics*, vol. 94, n. 6, pp. 854–869, 2014.
- [8] E. Khurana, Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin e M. Gerstein, «Role of non-coding sequence variants in cancer,» *Nature Reviews Genetics*, vol. 17, n. 2, pp. 93–108, 2016.
- [9] J. Tian et al., «Systematic functional interrogation of genes in GWAS loci identified ATF1 as a key driver in colorectal cancer modulated by a promoter-enhancer interaction,» *The American Journal of Human Genetics*, vol. 105, n. 1, pp. 29–47, 2019.
- [10] S. E. Bojesen et al., «Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer,» *Nature genetics*, vol. 45, n. 4, pp. 371–384, 2013.
- [11] K. Michailidou et al., «Association analysis identifies 65 new breast cancer risk loci,» *Nature*, vol. 551, n. 7678, pp. 92–94, 2017.
- [12] C. S. Pareek, R. Smoczynski e A. Tretyn, «Sequencing technologies and genome sequencing,» *Journal of applied genetics*, vol. 52, pp. 413–435, 2011.

- [13] S. K. Burley, H. M. Berman, G. J. Kleywegt, J. L. Markley, H. Nakamura e S. Velankar, «Protein Data Bank (PDB): the single global macromolecular structure archive,» *Protein crystallography: methods and protocols*, pp. 627–641, 2017.
- [14] N. M. Luscombe, D. Greenbaum e M. Gerstein, «What is bioinformatics? A proposed definition and overview of the field,» *Methods of information in medicine*, vol. 40, n. 04, pp. 346–358, 2001.
- [15] C. Caudai et al., «AI applications in functional genomics,» *Computational and Structural Biotechnology Journal*, vol. 19, pp. 5762–5790, 2021.
- [16] J. Zhou e O. G. Troyanskaya, «Predicting effects of noncoding variants with deep learning-based sequence model,» *Nature methods*, vol. 12, n. 10, pp. 931–934, 2015.
- [17] D. R. Kelley, J. Snoek e J. L. Rinn, «Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks,» *Genome research*, vol. 26, n. 7, pp. 990–999, 2016.
- [18] W. Ma et al., «DeepSATA: A Deep Learning-Based Sequence Analyzer Incorporating the Transcription Factor Binding Affinity to Dissect the Effects of Non-Coding Genetic Variants,» *International Journal of Molecular Sciences*, vol. 24, n. 15, p. 12 023, 2023.
- [19] B. Alberts et al., *Essential cell biology*. Garland Science, 2015.
- [20] P. F. Chinnery e E. A. Schon, «Mitochondria,» *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, n. 9, pp. 1188–1199, 2003.
- [21] H. M. McBride, M. Neuspiel e S. Wasiak, «Mitochondria: more than just a powerhouse,» *Current biology*, vol. 16, n. 14, R551–R560, 2006.
- [22] G. K. Voeltz, M. M. Rolls e T. A. Rapoport, «Structural organization of the endoplasmic reticulum,» *EMBO reports*, vol. 3, n. 10, pp. 944–950, 2002.
- [23] A. Ballabio, «The awesome lysosome,» *EMBO molecular medicine*, vol. 8, n. 2, pp. 73–76, 2016.
- [24] C. Yang e X. Wang, «Lysosome biogenesis: Regulation and functions,» *The Journal of cell biology*, vol. 220, n. 6, 2021.
- [25] E. C. Dell’Angelica, C. Mullins, S. Caplan e J. S. Bonifacino, «Lysosome-related organelles,» *The FASEB Journal*, vol. 14, n. 10, pp. 1265–1278, 2000.
- [26] M. Islinger, S. Grille, H. D. Fahimi e M. Schrader, «The peroxisome: an update on mysteries,» *Histochemistry and cell biology*, vol. 137, pp. 547–574, 2012.
- [27] National Human Genome Research Institute, *Deoxyribonucleic acid (DNA) Image*, <https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid>, 2024.
- [28] C. Fonseca Guerra, F. M. Bickelhaupt, J. G. Snijders e E. J. Baerends, «Hydrogen bonding in DNA base pairs: reconciliation of theory and experiment,» *Journal of the American Chemical Society*, vol. 122, n. 17, pp. 4117–4128, 2000.

- [29] A. Jansen e K. J. Verstrepen, «Nucleosome positioning in *Saccharomyces cerevisiae*,» *Microbiology and molecular biology reviews*, vol. 75, n. 2, pp. 301–320, 2011.
- [30] G. Zheng, *The packaging of DNA in chromatin*. Rutgers The State University of New Jersey, School of Graduate Studies, 2010.
- [31] National Human Genome Research Institute, *Chromosome Image*, <https://www.genome.gov/genetics-glossary/Chromosome>, 2024.
- [32] M. B. Gerstein et al., «What is a gene, post-ENCODE? History and updated definition,» *Genome research*, vol. 17, n. 6, pp. 669–681, 2007.
- [33] R. J. White, *Gene transcription: mechanisms and control*. John Wiley & Sons, 2009.
- [34] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts e P. Walter, «From DNA to RNA,» in *Molecular Biology of the Cell. 4th edition*, Garland Science, 2002.
- [35] P. Cramer, «Organization and regulation of gene transcription,» *Nature*, vol. 573, n. 7772, pp. 45–54, 2019.
- [36] National Human Genome Research Institute, *Transcription Image*, <https://www.genome.gov/genetics-glossary/Transcription>, 2024.
- [37] A. Philips e T. Cooper*, «RNA processing and human disease,» *Cellular and Molecular Life Sciences CMLS*, vol. 57, pp. 235–249, 2000.
- [38] S. Hocine, R. H. Singer e D. Grünwald, «RNA processing and export,» *Cold Spring Harbor perspectives in biology*, vol. 2, n. 12, a000752, 2010.
- [39] M. Livingstone, E. Atas, A. Meller e N. Sonenberg, «Mechanisms governing the control of mRNA translation,» *Physical biology*, vol. 7, n. 2, p. 021 001, 2010.
- [40] V. Ramakrishnan, «Ribosome structure and the mechanism of translation,» *Cell*, vol. 108, n. 4, pp. 557–572, 2002.
- [41] J. Lemonnier, N. Lemonnier, S. Pascolo e C. Pichon, «The Marathon of the Messenger,»
- [42] National Human Genome Research Institute, *Translation Image*, <https://www.genome.gov/genetics-glossary/Translation>, 2024.
- [43] G. E. Schulz e R. H. Schirmer, *Principles of protein structure*. Springer Science & Business Media, 2013.
- [44] R. M. Bavle, «Mitosis at a glance,» *Journal of Oral and Maxillofacial Pathology*, vol. 18, n. Suppl 1, S2–S5, 2014.
- [45] C. E. Walczak, S. Cai e A. Khodjakov, «Mechanisms of chromosome behaviour during mitosis,» *Nature reviews Molecular cell biology*, vol. 11, n. 2, pp. 91–102, 2010.
- [46] X. Li, F. Yang e B. Rubinsky, «A theoretical study on the biophysical mechanisms by which tumor treating fields affect tumor cells during mitosis,» *IEEE Transactions on Biomedical Engineering*, vol. 67, n. 9, pp. 2594–2602, 2020.

- [47] M. Sullivan e D. O. Morgan, «Finishing mitosis, one step at a time,» *Nature reviews Molecular cell biology*, vol. 8, n. 11, pp. 894–903, 2007.
- [48] National Human Genome Research Institute, *Mitosis Image*, <https://www.genome.gov/genetics-glossary/Mitosis>, 2024.
- [49] R. A. Laskey, M. P. Fairman e J. J. Blow, «S phase of the cell cycle,» *Science*, vol. 246, n. 4930, pp. 609–614, 1989.
- [50] S. P. Bell e A. Dutta, «DNA replication in eukaryotic cells,» *Annual review of biochemistry*, vol. 71, n. 1, pp. 333–374, 2002.
- [51] A. Dutta e S. P. Bell, «Initiation of DNA replication in eukaryotic cells,» *Annual review of cell and developmental biology*, vol. 13, n. 1, pp. 293–332, 1997.
- [52] «Chapter 42 - S Phase and DNA Replication,» in *Cell Biology (Third Edition)*, T. D. Pollard, W. C. Earnshaw, J. Lippincott-Schwartz e G. T. Johnson, cur., Third Edition, Elsevier, 2017, pp. 727–741, ISBN: 978-0-323-34126-4. DOI: <https://doi.org/10.1016/B978-0-323-34126-4.00042-6>. indirizzo: <https://www.sciencedirect.com/science/article/pii/B9780323341264000426>.
- [53] National Human Genome Research Institute, *DNA Replication Image*, <https://www.genome.gov/genetics-glossary/DNA-Replication>, 2024.
- [54] W. S. McCulloch e W. Pitts, «A logical calculus of the ideas immanent in nervous activity,» *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.
- [55] A. M. Turing, *Computing machinery and intelligence*. Springer, 2009.
- [56] N. Muthukrishnan, F. Maleki, K. Ovens, C. Reinhold, B. Forghani, R. Forghani et al., «Brief history of artificial intelligence,» *Neuroimaging Clinics of North America*, vol. 30, n. 4, pp. 393–399, 2020.