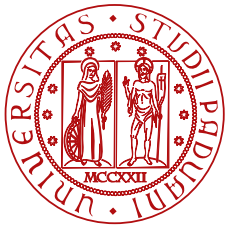

Università degli Studi di Padova

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

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

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

Indice delle Figure

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 — branca 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 codificianti*. Le mutazioni in queste zone del genoma, che apparentemente svolgono funzioni marginali, sono responsabili dello sviluppo di disturbi importanti, come le *malattie mendeliane*[1] [4, 5], l'epilessia [6], malattie cardiovascolari [3, 7] e soprattuto tumori — tra cui il cancro del colon-retto e tumore al seno [8–11].

Risulta quindi vitale continuare a studiare gli effetti che le varianti non codificanti in sequenze genomiche hanno sugli individui. Proprio a questo proposito, con l'avvento dell'intelligenza artificiale, in particolare del *deep learning*, si continuano a trovare e perfezionare soluzioni che permettano di delineare sempre con più 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 [12] — ha avuto un forte impulso nell'approfondire le varianti non codificanti ma rimangono ancora significative lacune nella comprensione riguardante la relazione tra mutazioni genetiche ed espressione genica. L'utilizzo di tecniche di deep learning quindi cruciale per continuare la ricerca; a questo proposito, nel presente elaborato accademico verranno descritti e paragonati tre *tool* che utilizzano le *reti neurali convoluzionali* per predire l'effetto delle varianti non codificanti su sequenze genomiche: DeepSEA [13], Basset [14] e DeepSATA [15].

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

1.1 BACKGROUND

La cellula è l'unità fondamentale della vita. La cellula è una piccola miscela acquosa con componenti chimici, racchiusi in una mambrana, 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 [16].

All'interno della cellula eucariote (illustrazione 1.1), 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 maggior parte delle attività cellulari[2]; questo processo è anche chiamato *respirazione cellulare* perchè consumando l'ossigeno viene rilasciata anidride carbonica [16, 17]. 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 [18].

Il *reticolo endoplasmatico* è invece un organulo molto esteso e svolge molteplici funzioni. Tra queste compiti rientrano quelli di traslocazione di proteine e il ripiegamento delle proteine (*protein folding*) [16, 19]. I *lisosomi* si occupano di degradare e riciclare gli scarti cellulari e

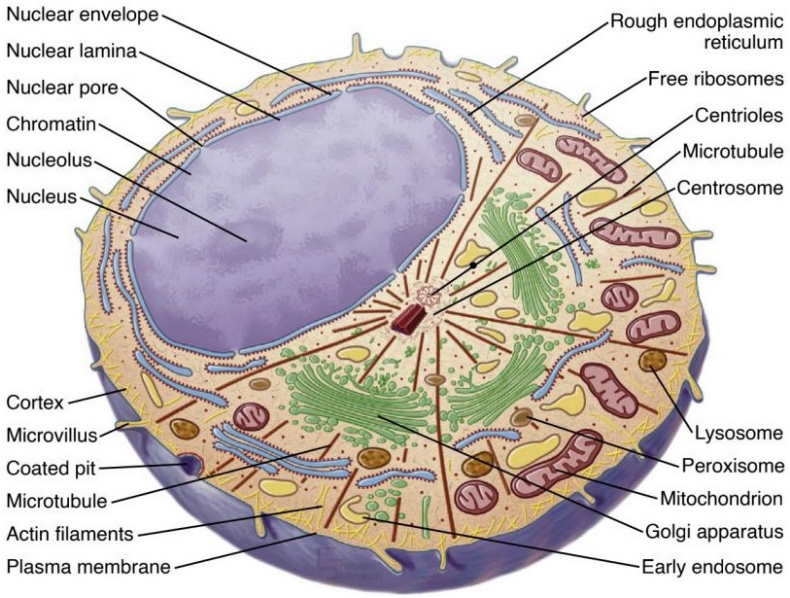

Figura 1.1: Rappresentazione schematica della cellula eucariote; si possono notare i principali organuli tra cui i mitocondri, lisosomi e perossiomi, il reticolo endoplasmatico, e il nucleo [2].

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

giocano un ruolo fondamentale per l'omeostasi della cellula[3], il suo sviluppo e il suo invecchiamento [20–22]. Infine, i *perossiomi* 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 [16, 23, 24].

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, 16]. La molecola di DNA è una struttura a doppia elica formata da *nucleotidi*. Osservando l'illustrazione 1.2, i nucleotidi sono composti a loro volta da tre elementi fondamentali: una *base azotata*, uno *zucchero* e un *gruppo fosfato*[4]. Le basi azotate 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, 25]. 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

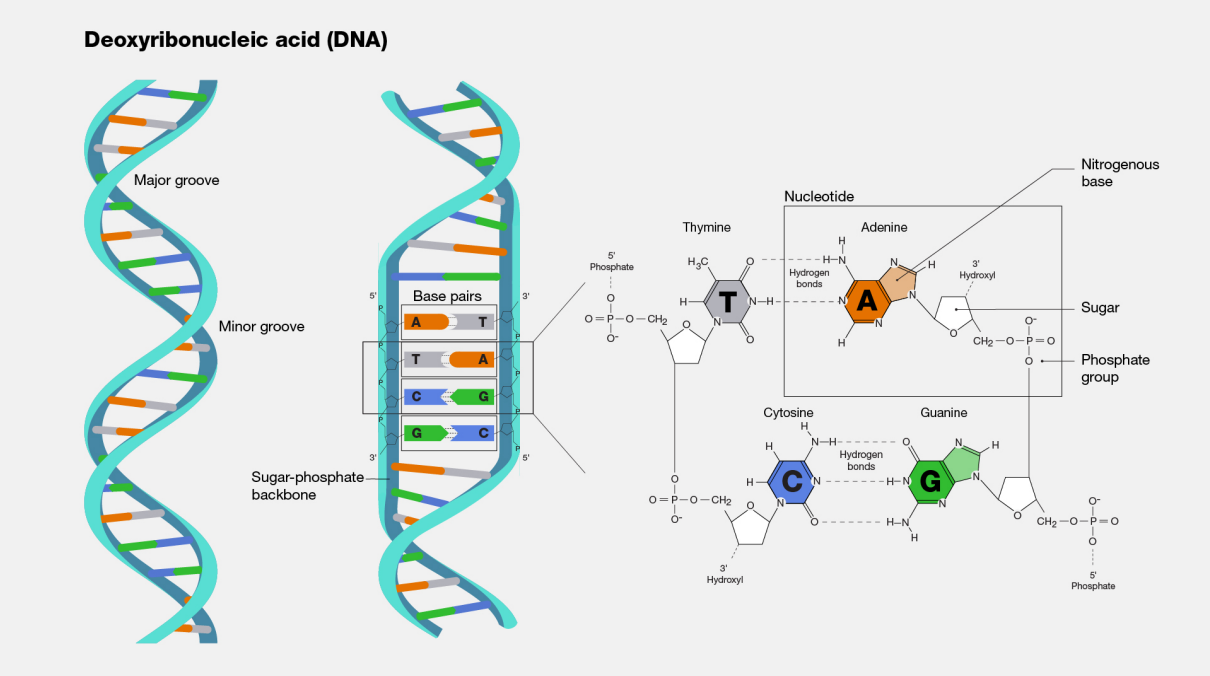

Figura 1.2: Rappresentazione schematica del DNA in cui si possono osservare le coppie di basi azotate, legate tra loro attraverso gli zuccheri e i gruppi fosfati [26].

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

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

a raggomitolarsi in un cromosoma di grandezza inferiore a 2 micron (figura 1.3). Il processo di *DNA-packaging* inizia avvolgendo la doppia elica di DNA attorno a delle proteine dette *istoni* e formado 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 [27, 28].

La rilevanza del DNA è data delle informazioni essenziali che questa molecola contiene. Tali informazioni risiedono nei geni, che sono delle sequenze genomiche che codificano uno o più prodotti biologici operativi [30]. 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 [31]. L'espressione genica è divisa in due fasi principali: la *trascrizione* — 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 — anzichè la Timina è presente l'Uracile (U) — e per lo zucchero — da desossiribosio a ribosio [32]. La trascrizione del DNA in RNA inizia quando delle proteine, chiamate *fattori di trascrizione*, riconoscono la regione che delimita l'inizione della molecola del gene da esprimere, detta *zona promotrice*. Dopo aver riconosciuto l'inizio della sequenza, queste proteine permettono ad un enzima chiamata *RNA polimerasi* di attaccarsi ed aprire la doppia elica del DNA [33]. 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

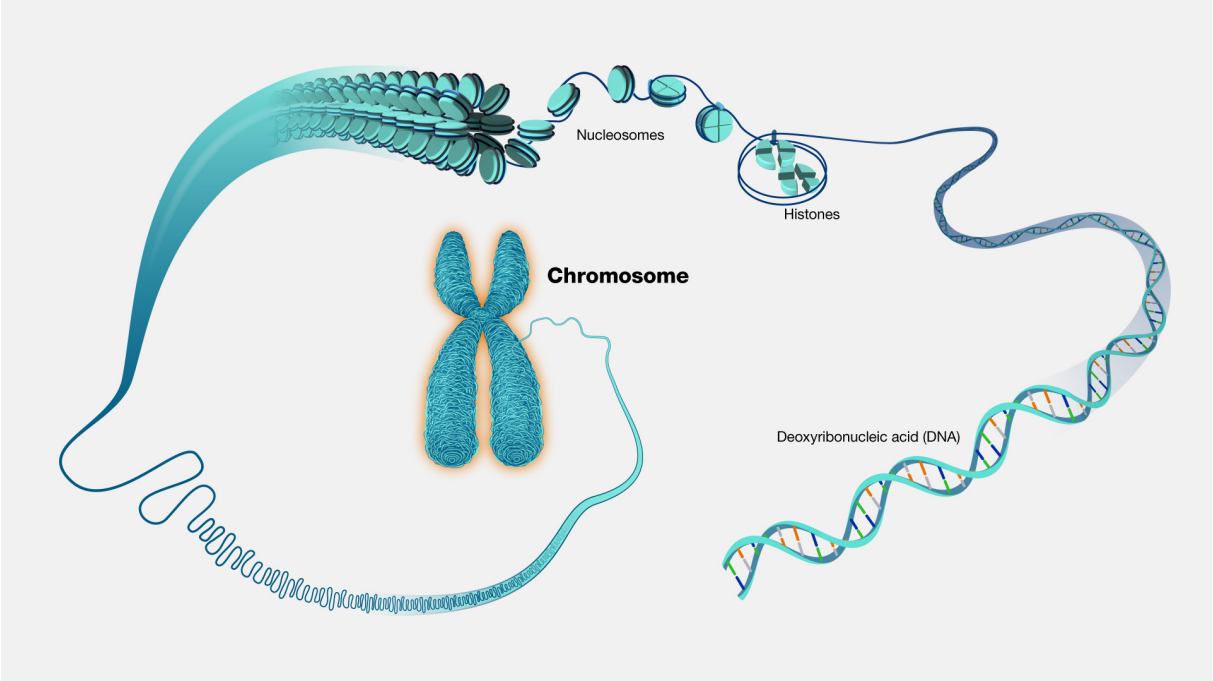

Figura 1.3: Il processo di impacchettamento del DNA che permette di compattare la struttura a doppia eilca nel cromosoma [29].

a quello del DNA (di conseguenza $A \to U$, $C \to G$, $G \to C$ e $T \to A$). Così facendo l'acido ribonucleico viene creato un nucleotide alla volta, analizzando quello del DNA [32]. 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 1.4.

Prima di uscire dal nucleo l'RNA messaggero subisce una serie di elaborazioni necessarie per rendere le informazioni immagazzinae sicure: diverse sono le malattie che emergono per mutazioni presenti nell'mRNA tra cui la distrofia miotonica[5] [35]. La prima eleborazione viene chiamata $5'$-*end capping* e si occupa di aggiungere alla terminazione $5'$ dell'mRNA una Guanina 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 *introni*— 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 — datta anche *poly*A *tail* — che, in maniera molto simile al $5'$-*end capping* garantisce una stabilità del filamento di acido ribonucleico [36, 37].

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 anzichè codifiare

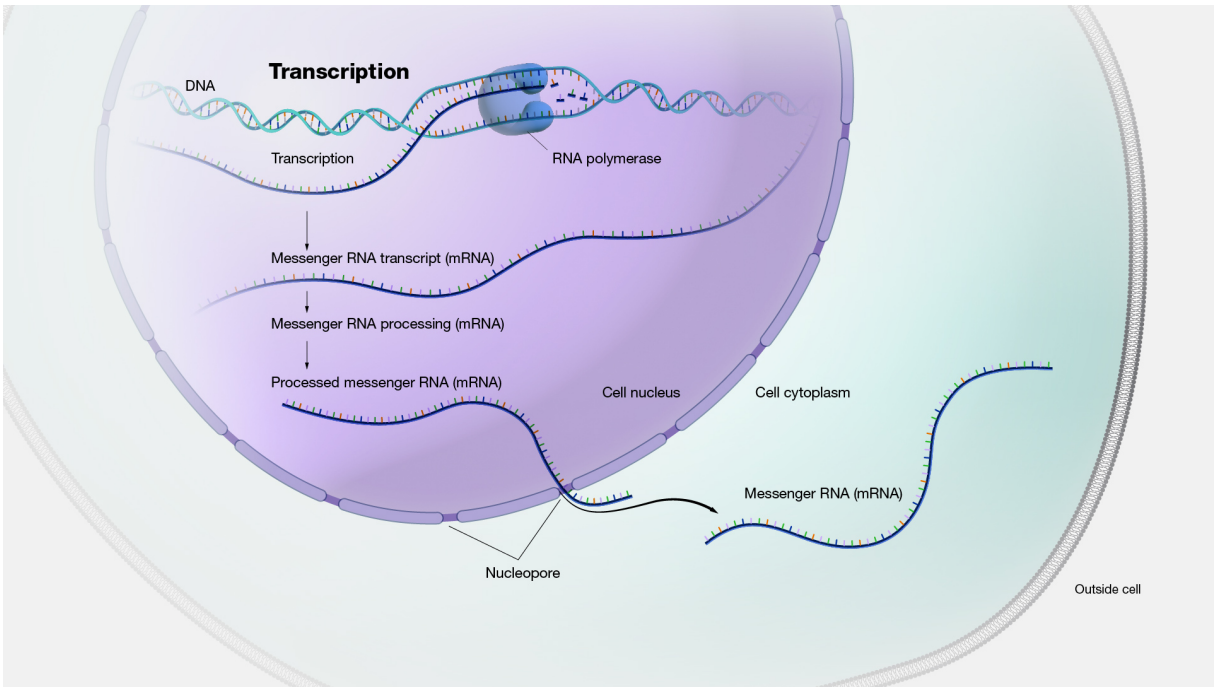

Figura 1.4: Il processo di trascrizione del DNA del gene in RNA mediante la RNA polimerasi [34].

[5]Le distrofie miotoniche sono patologie che colpiscono principalmente l'apparato muscolo-scheletrico.

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; 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 [37–39]. In questo modo i ribosomi, analizzando codone dopo codone riescono a crare la catena polipeptidica mediante l'RNA di trasporto, come mostrato nella figura 1.5.

Una volta creata la sequenza polipeptidica, inizia il processo di ripiegamento della proteina. 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 [41]. 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*.

geni, parti del gene in modo da spiegare bene le varianti non codificanti

Come accennato all'inizio del capitolo, la cellula possiede la notevole capacità di replicarsi; il processo replicazione cellulare è detto *mitosi*.

DNA replication [42] durante il ciclo della mitosi (non descriverla tutta) [43]

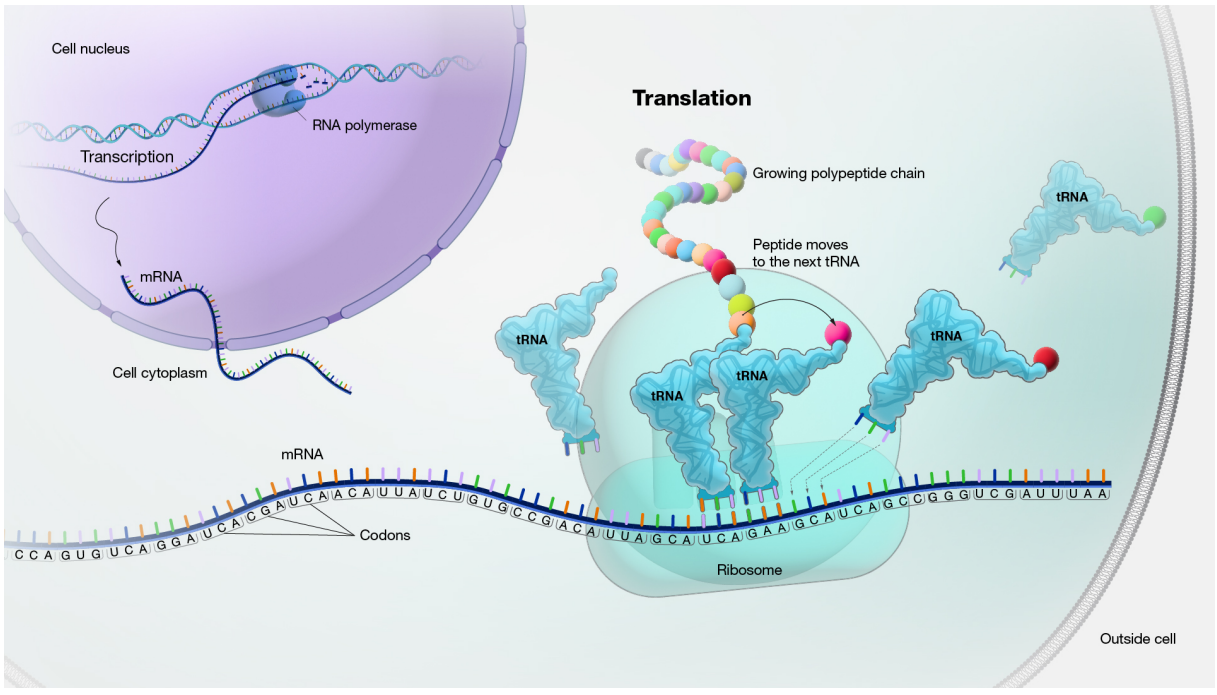

Figura 1.5: Il processo di traduzione da RNA messaggero a polipeptide attraverso il tRNA e i ribosomi [40].

1.2 CENNI STORICI

C'è sto bell'articolo che racconta per bene la situa[44–46]. Qua ci puoi buttare dentro anche la questione meme del junk DNA così pushi per bene le citazioni goliardiche. Cita il libro [2] che descivre a cosa serve il dna (pagina 10)

Sto libro parla anche della scopetrta delle cellule[16]

1.3 STATO DELL'ARTE

parla dei tool e quanti ne sono uscit per letsgoscare le cose. Parla delle diversi vantaggi che AI ha portato nel letsgosky. Butta dentro deepvirfinder a anche alphafold perche fa figo[47]

anche esmpi di DNA folding

2

Reti neurali

quando fai CNN dici che ML fa cagare per questi

2.1 Cenni storici

video Enkk per storia degli LLM

Reti neurlai, teoricamente formulate negli anni '40 [48] e sviluppate solo nell'ultimo decennio

2.2 CNN

3

Dettagli Implementativi

4

Conclusioni

Bibliografia

[1] Sitanshu Sekhar Sahu e Ganapati Panda. «Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach». In: *Genomics, Proteomics and Bioinformatics* 9.1-2 (2011), pp. 45–55.

[2] Thomas D Pollard et al. *Cell Biology E-Book: Cell Biology E-Book*. Elsevier Health Sciences, 2022.

[3] Feng Zhang e James R Lupski. «Non-coding genetic variants in human disease». In: *Human molecular genetics* 24.R1 (2015), R102–R110.

[4] JD French e SL Edwards. «The role of noncoding variants in heritable disease». In: *Trends in Genetics* 36.11 (2020), pp. 880–891.

[5] Heidi Chial. «Mendelian genetics: patterns of inheritance and single-gene disorders». In: *Nature Education* 1.1 (2008), p. 63.

[6] Susanna Pagni et al. «Non-coding regulatory elements: Potential roles in disease and the case of epilepsy». In: *Neuropathology and Applied Neurobiology* 48.3 (2022), e12775.

[7] Ashish Kapoor et al. «An enhancer polymorphism at the cardiomyocyte intercalated disc protein NOS1AP locus is a major regulator of the QT interval». In: *The American Journal of Human Genetics* 94.6 (2014), pp. 854–869.

[8] Ekta Khurana et al. «Role of non-coding sequence variants in cancer». In: *Nature Reviews Genetics* 17.2 (2016), pp. 93–108.

[9] Jianbo 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». In: *The American Journal of Human Genetics* 105.1 (2019), pp. 29–47.

[10] Stig E Bojesen et al. «Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer». In: *Nature genetics* 45.4 (2013), pp. 371–384.

[11] Kyriaki Michailidou et al. «Association analysis identifies 65 new breast cancer risk loci». In: *Nature* 551.7678 (2017), pp. 92–94.

[12] Claudia Caudai et al. «AI applications in functional genomics». In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 5762–5790.

[13] Jian Zhou e Olga G Troyanskaya. «Predicting effects of noncoding variants with deep learning–based sequence model». In: *Nature methods* 12.10 (2015), pp. 931–934.

[14] David R Kelley, Jasper Snoek e John L Rinn. «Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks». In: *Genome research* 26.7 (2016), pp. 990–999.

[15] Wenlong 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». In: *International Journal of Molecular Sciences* 24.15 (2023), p. 12023.

[16] Bruce Alberts et al. *Essential cell biology*. Garland Science, 2015.

[17] Patrick F Chinnery e Eric A Schon. «Mitochondria». In: *Journal of Neurology, Neurosurgery & Psychiatry* 74.9 (2003), pp. 1188–1199.

[18] Heidi M McBride, Margaret Neuspiel e Sylwia Wasiak. «Mitochondria: more than just a powerhouse». In: *Current biology* 16.14 (2006), R551–R560.

[19] Gia K Voeltz, Melissa M Rolls e Tom A Rapoport. «Structural organization of the endoplasmic reticulum». In: *EMBO reports* 3.10 (2002), pp. 944–950.

[20] Andrea Ballabio. «The awesome lysosome». In: *EMBO molecular medicine* 8.2 (2016), pp. 73–76.

[21] Chonglin Yang e Xiaochen Wang. «Lysosome biogenesis: Regulation and functions». In: *The Journal of cell biology* 220.6 (2021).

[22] Esteban C Dell'Angelica et al. «Lysosome-related organelles». In: *The FASEB Journal* 14.10 (2000), pp. 1265–1278.

[23] Markus Islinger et al. «The peroxisome: an update on mysteries». In: *Histochemistry and cell biology* 137 (2012), pp. 547–574.

[24] Markus Islinger et al. «The peroxisome: an update on mysteries 2.0». In: *Histochemistry and cell biology* 150 (2018), pp. 443–471.

[25] Célia Fonseca Guerra et al. «Hydrogen bonding in DNA base pairs: reconciliation of theory and experiment». In: *Journal of the American Chemical Society* 122.17 (2000), pp. 4117–4128.

[26] National Human Genome Research Institute. *Deoxyribonucleic acid (DNA) Image*. `https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid`. 2024.

[27] An Jansen e Kevin J Verstrepen. «Nucleosome positioning in Saccharomyces cerevisiae». In: *Microbiology and molecular biology reviews* 75.2 (2011), pp. 301–320.

[28] Guohui Zheng. *The packaging of DNA in chromatin*. Rutgers The State University of New Jersey, School of Graduate Studies, 2010.

[29] National Human Genome Research Institute. *Chromosome Image*. `https://www.genome.gov/genetics-glossary/Chromosome`. 2024.

[30] Mark B Gerstein et al. «What is a gene, post-ENCODE? History and updated definition». In: *Genome research* 17.6 (2007), pp. 669–681.

[31] Robert J White. *Gene transcription: mechanisms and control*. John Wiley & Sons, 2009.

[32] Bruce Alberts et al. «From DNA to RNA». In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[33] Patrick Cramer. «Organization and regulation of gene transcription». In: *Nature* 573.7772 (2019), pp. 45–54.

[34] National Human Genome Research Institute. *Transcription Image*. `https://www.genome.gov/genetics-glossary/Transcription`. 2024.

[35] AV Philips e TA Cooper*. «RNA processing and human disease». In: *Cellular and Molecular Life Sciences CMLS* 57 (2000), pp. 235–249.

[36] Sami Hocine, Robert H Singer e David Grünwald. «RNA processing and export». In: *Cold Spring Harbor perspectives in biology* 2.12 (2010), a000752.

[37] Mark Livingstone et al. «Mechanisms governing the control of mRNA translation». In: *Physical biology* 7.2 (2010), p. 021001.

[38] Venki Ramakrishnan. «Ribosome structure and the mechanism of translation». In: *Cell* 108.4 (2002), pp. 557–572.

[39] Jérôme Lemonnier et al. «The Marathon of the Messenger». In: ().

[40] National Human Genome Research Institute. *Translation Image*. `https://www.genome.gov/genetics-glossary/Translation`. 2024.

[41] Georg E Schulz e R Heiner Schirmer. *Principles of protein structure*. Springer Science & Business Media, 2013.

[42] Stephen P Bell e Anindya Dutta. «DNA replication in eukaryotic cells». In: *Annual review of biochemistry* 71.1 (2002), pp. 333–374.

[43] J Richard McIntosh, Maxim I Molodtsov e Fazly I Ataullakhanov. «Biophysics of mitosis». In: *Quarterly reviews of biophysics* 45.2 (2012), pp. 147–207.

[44] Francis Harry Compton Crick e James Dewey Watson. «The complementary structure of deoxyribonucleic acid». In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 223.1152 (1954), pp. 80–96.

[45] James D Watson e Francis HC Crick. «Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid». In: *Nature* 171.4356 (1953), pp. 737–738.

[46] Yiwei Li, Wenhui Tang e Ming Guo. «The cell as matter: Connecting molecular biology to cellular functions». In: *Matter* 4.6 (2021), pp. 1863–1891.

[47] Jie Ren et al. «Identifying viruses from metagenomic data using deep learning». In: *Quantitative Biology* 8.1 (2020), pp. 64–77.

[48] Warren S McCulloch e Walter Pitts. «A logical calculus of the ideas immanent in nervous activity». In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.