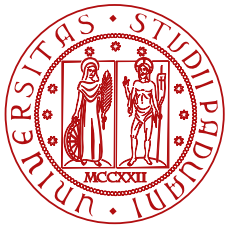

Università degli Studi di Padova

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

Abstract

Indice

Indice delle Figure

Indice delle Tabelle

Indice degli Algoritmi

Indice dei Frammenti di Codice

Lista degli Acronimi

CNN Rete neurale convoluzionale — *Convolutional Neural Network*

DNA Acido desossirbonucleico — *DeoxyriboNucleic Acid*

AI Intelligenza artificiale — *Artificial Intelligence*

ATP Adenosintrifosfato — *Adenosine TriPhosphate*

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 all'interno del *citoplasma*. 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 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].

[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.

[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.

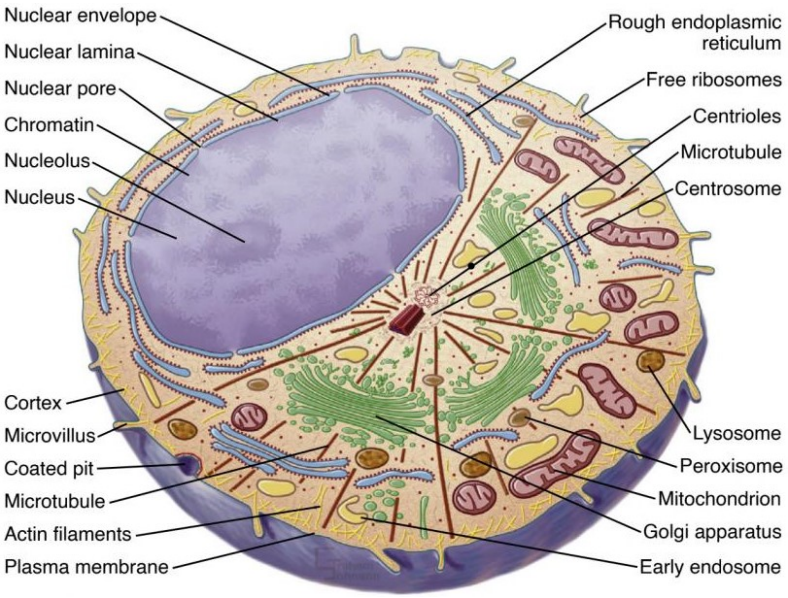

Figura 1.1: Rappresentazione schematica della cellula eucariote; si possono notare i principali organuli tra cui i mitocondri, lisosomi e perossiomi, il reticolo endoplasmatico, e il nucleo [2].

I *ribosomi* si occupano di accelerare la sintesi delle proteine usando le sequenze di nucleotidi del *RNA messaggero* (mRNA) per specificare la sequenza di aminoacidi

Parla dei nucleoli [27]

parla dei geni, delle parti del gene in modo da spiegare bene le varianti non codificanti. Se vuoi parla anche della funzione del dna del creare proteine descrivendo in breve il DNA

parla anche di come il DNA è salavto dei cromosmi in agglomarati

DNA replication [28]

1.2 CENNI STORICI

C'è sto bell'articolo che racconta per bene la situa[29–31]. 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[32]

anche esmpi di DNA folding

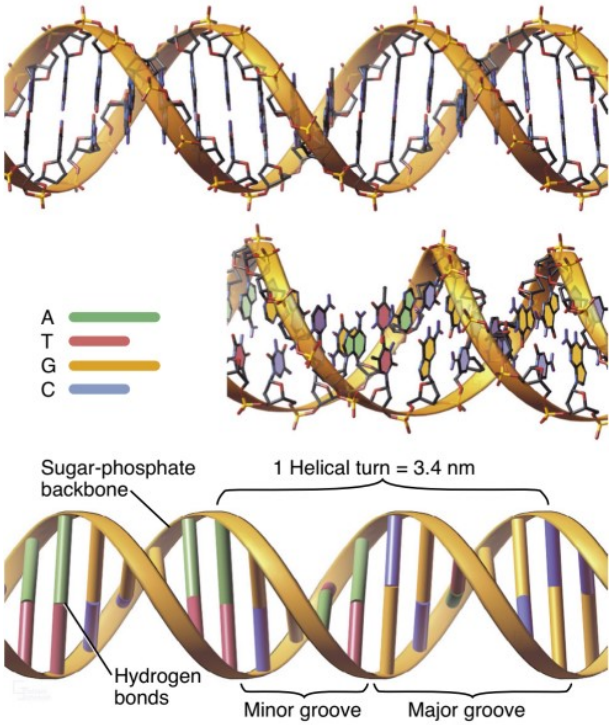

Figura 1.2: Rappresentazione schematica del DNA; si possono osservare le coppie di basi azotate, legate tra loro attraverso gli zuccheri e i gruppi fosfati [2].

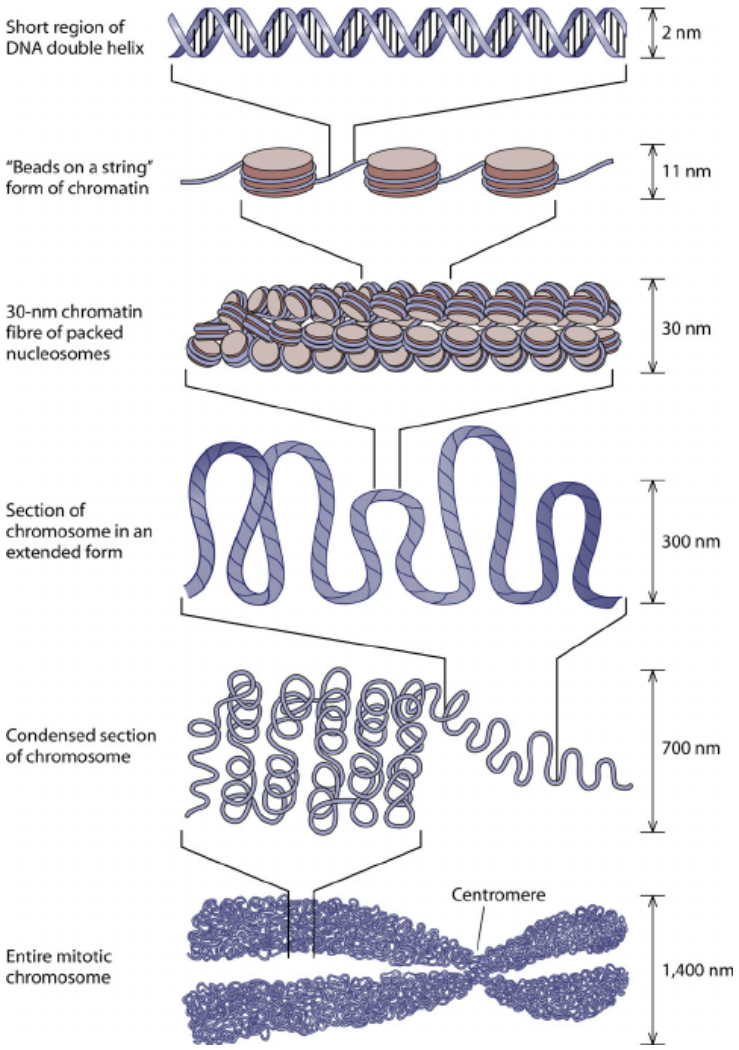

Figura 1.3: Il processo di impacchettamento del DNA [26].

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 [33] e sviluppate solo nell'ultimo decennio

2.2 CNN

3

Dettagli Implementativi

4

Conculsioni

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

[27] Thoru Pederson. «The plurifunctional nucleolus». In: *Nucleic acids research* 26.17 (1998), pp. 3871–3876.

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

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.