



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

1	Introduzione	1
1.1	Background	2
1.2	Cenni storici	2
1.3	Stato dell'arte	3
2	Reti neurali	5
2.1	Cenni storici	5
2.2	CNN	5
3	Dettagli Implementativi	7
4	Conclusioni	9
	Bibliografia	11

Indice delle Figure

1.1	Rappresentazione schematica del DNA [18].	3
-----	---	---

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*



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]. Ciò nonostante, la quasi totalità dei disturbi genomici è dovuta alle mutazioni nelle regioni non codificanti [2] — 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 [3, 4], l'epilessia [5], malattie cardiovascolari [2, 6] e soprattutto tumori — tra cui il cancro del colon-retto e tumore al seno [7–10].

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 [11] — 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 [12], Basset [13] e DeepSATA [14].

¹Le 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 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 [15].

All'interno della cellula eucariote 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, trasformano; questo processo è chiamato respirazione cellulare.

Ciò nonostante, l'organulo più importante è il nucleo.

Plurifunctional nucleous [16]

il DNA si trova nel nucleo in forma compatta, i cromosomi.

Il DNA è una molecola che si trova all'interno del nucleo di ciascuna cellula ed è composta, in una struttura a doppia elica, da *nucleotidi*. Osservando l'illustrazione 1.1, i nucleotidi sono composti a loro volta da tre elementi fondamentali: una *base azotata*, uno *zucchero* e un *gruppo fosfato*. Le basi azotate sono quattro — Adenina (A), Citosina (C), Guanina (G) e Timina (T) — e si legano tra loro 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, 17].

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

DNA replication [19]

1.2 CENNI STORICI

C'è sto bell'articolo che racconta per bene la situa[20–22]. Qua ci puoi buttare dentro anche la questione meme del junk DNA così pushi per bene le citazioni goliardiche.

Sto libro parla anche della scopetrta delle cellule[15]

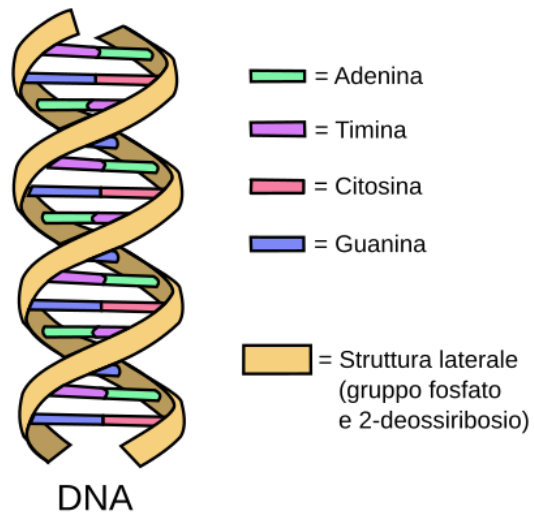


Figura 1.1: Rappresentazione schematica del DNA [18].

1.3 STATO DELL'ARTE

parla dei tool e quanti ne sono usciti per letsgoscare le cose. Parla delle diversi vantaggi che AI ha portato nel letsgosky. Butta dentro deepvirfinder a anche alphafold perche fa figo[23]



Reti neurali

quando fai CNN dici che ML fa cagare per questi

2.1 CENNI STORICI

video Enkk per storia degli LLM

Reti neurali, teoricamente formulate negli anni '40 [24] e sviluppate solo nell'ultimo decennio

2.2 CNN



Dettagli Implementativi



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] Feng Zhang e James R Lupski. «Non-coding genetic variants in human disease». In: *Human molecular genetics* 24.R1 (2015), R102–R110.
- [3] JD French e SL Edwards. «The role of noncoding variants in heritable disease». In: *Trends in Genetics* 36.11 (2020), pp. 880–891.
- [4] Heidi Chial. «Mendelian genetics: patterns of inheritance and single-gene disorders». In: *Nature Education* 1.1 (2008), p. 63.
- [5] 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.
- [6] 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.
- [7] Ekta Khurana et al. «Role of non-coding sequence variants in cancer». In: *Nature Reviews Genetics* 17.2 (2016), pp. 93–108.
- [8] 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.
- [9] 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.
- [10] Kyriaki Michailidou et al. «Association analysis identifies 65 new breast cancer risk loci». In: *Nature* 551.7678 (2017), pp. 92–94.
- [11] Claudia Caudai et al. «AI applications in functional genomics». In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 5762–5790.
- [12] 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.

- [13] 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.
- [14] 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.
- [15] Bruce Alberts et al. *Essential cell biology*. Garland Science, 2015.
- [16] Thoru Pederson. «The plurifunctional nucleolus». In: *Nucleic acids research* 26.17 (1998), pp. 3871–3876.
- [17] 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.
- [18] Forluvoft. *Struttura a doppia elica del DNA*. 2012. URL: [https://commons.wikimedia.org/wiki/File:DNA_simple2_\(it\).svg](https://commons.wikimedia.org/wiki/File:DNA_simple2_(it).svg).
- [19] Stephen P Bell e Anindya Dutta. «DNA replication in eukaryotic cells». In: *Annual review of biochemistry* 71.1 (2002), pp. 333–374.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] Jie Ren et al. «Identifying viruses from metagenomic data using deep learning». In: *Quantitative Biology* 8.1 (2020), pp. 64–77.
- [24] 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.