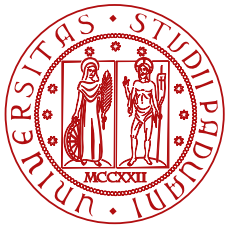

Università degli Studi di Padova

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*

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]. Ciò nonostante, la quasi totalità dei disturbi genomici è dovuta alle mutazioni nelle regioni non codificanti [2] — 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] [3, 4], l'epilessia [5], malattie cardiovascolari [2, 6] e soprattuto 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 e quindi di reti neurali rimane quindi cruciale per continuare la ricerca; a questo proposito in questo documento verranno descritti e paragonati tre *tool* che utilizzano le *reti neurali convoluzionali* per predire l'effetto delle varianti non codificanti su seguenze genomiche: DeepSEA [12], Basset [13] e DeepSATA [14].

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

1.1 VARIANTI NON CODIFICANTI

1.2 STATO DELL'ARTE

STORIA

2

Reti neurali

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] 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.