Report di Tirocinio svolto presso il Laboratorio di Bioinformatica dell'Università di Trento

May 19, 2020

Supervisore: Alessandro Romanel

Tirocinante: Ermano Buikis 197761

1 Introduzione generale

Molti studi hanno dimostrato che i rischi di sviluppare cancro sono associati ai singoli polimorfismi a singolo nucleotide (SNP) presenti in zone non codificanti del genoma. Le varianti geniche in regioni non-codificanti sono state descritte di fondamentale importanza per la predizione di disfunzioni geniche, ma difficile e' la loro individuazione. In questo studio sono riportati due diversi studi sugli SNPs in regioni non codificanti. Nel primo studio vedremo la descrizione di SNPs in regioni non codificanti associate ad un long non coding RNA responsabile della regolazione genica del gene del recettore della prostata(PGR). Nel secondo studio illustro un modello di rete neurale predittiva del fenotipo cellulare (canceroso o normale) basandomi su singoli record di SNP in regioni non codificanti. Entrambi gli studi condividono lo stesso dataset.

1.1 Dataset

- RS_ID : ID dello SNPs
- chr : numero del cromosoma
- pos : posizione all'interno del cromosoma
- ref : nucleotide di riferimento (o ancestrale)
- alt : nucleotide alternativo
- scoreA : MAF score 1000genomes
- scoreB : MAF score TOPMed
- functional_element : codice dell'elemento funzionale
- n_experiment : numero di esperimenti che riportano tale SNP
- file_type : tipo di file dal quale e'stato estratto lo SNP
- cell_line : codice (o nome, a seconda dei casi) della linea cellulare

- cancer_type : questo elemento si riferisce se il campione apparteneva ad una biopsia o da una coltura cellulare derivante da tessuto della prostata
  - valori possibili :
    * 'Nan' per biopsia
    * 'Prostate' per linea cellulare
- cell_line_cancer : presenza od assenza di tumorgenicita' nel fenotipo cellulare
  - valori possibili :
    * 'normal'
    * 'cancer'

2 PARTE PRIMA - SNPs su lncRNA (PGR-AS1) prossimale al gene del recettore del progesterone (PGR).

2.1 Introduzione

Gli SNPs nelle regioni non codificanti del gene del recettore del progesterone (PGR) si pensa che siano correlati ad un aumento del rischio di carcinoma all'endometrio,all' ovario ed alla prostata. Tuttavia, nessuno studio ha valutato sistematicamente il ruolo delle regioni non codificanti del gene PGR nella carcinogenesi in questi tessuti. In questo studio, sono stati individuati SNPs associati a regioni non codificanti del gene PGR. Si pensa siano correlati ad un aumento del rischio per lo sviluppo tumorale.

2.2 Metodi

Per la ricerca delle varianti geniche sono state utilizzate diverse risorse online gratuite, le quali contengono un interfaccia web per la ricerca degli SNPs all'interno di Database pubblici governativi e non. Al fine di ottenere un informazione quanto piu' completa possibile, sono state confrontate ed integrate le informazioni relative alle varianti, provenienti dalle diverse fonti. Quando possibile e' stata preferita l'annotazione proveniente dal genoma umano di riferimento di ultima generazione (GRCh38).

- Database utilizzati
  - dbSNP https://www.ncbi.nlm.nih.gov/snp/
  - Clinvar https://www.ncbi.nlm.nih.gov/clinvar/
  - Ensembl https://www.ensembl.org/index.html
  - UCSC Genome Browser http://genome.ucsc.edu/index.html
  - GeneCard https://www.genecards.org/
  - Gene browser NCBI https://www.ncbi.nlm.nih.gov/gene/
  - Genome locator (GeneCard) https://genecards.weizmann.ac.il/geneloc/index.shtml

2.3 Codice Python per l'elaborazione dei dati degli SNPs trovati in associazione con Elementi Funzionali in Linee cellulari tumorali della prostata

2.3.1 Preparazione dei dati

[2]:
```
import pandas as pd
path = '/home/user/Desktop/Tirocinio/ML_python/
 ↪functionalElementDataARBindingProstate.csv'
raw_data = pd.read_csv(path) # load csv file
```

[3]:
```
tot =raw_data['RS_ID'].count() # get number of rows
print('totale record =',tot)
```

```
totale record = 457933
```

[4]:
```
raw_data.dtypes # check data type
```

[4]:
```
RS_ID                 object
chr                   object
pos                    int64
ref                   object
alt                   object
scoreA               float64
scoreB               float64
functional_element    object
n_experiment           int64
file_type             object
cell_line             object
cancer_type           object
cell_line_cancer      object
dtype: object
```

2.3.2 Dati grezzi del dataset

[5]:
```
raw_data.head(1)
```

[5]:
```
        RS_ID chr        pos ref alt  scoreA   scoreB functional_element  \
0  rs10201930   2  189874460   G   A  0.0719  0.06594     H3K36me3-human

   n_experiment file_type cell_line cancer_type cell_line_cancer
0             1    narrow  prostate         NaN           normal
```

2.3.3 Prendo solo gli SNPs associati a forme tumorali

[6]:
```
df = raw_data[(raw_data.cell_line_cancer == 'cancer')] # TAKE only cancer type
```

2.3.4 Estrazione dei dati di PGR Progesterone receptor gene

Estrazione dei dati degli SNPs su PGR. Vado a selezionare solo gli SNPs che cadono nella prossimita' del gene. In particolare tutti gli SNPs che sono fino a 10Kbp (kilobasi) a monte o a valle rispetto al gene d'interesse.

[7]:
```
CHR = '11' # chromosome
margin = 10_000 # bp
start = 101_029_624 - margin # bp
end = 101_129_813 + margin    # bp
```

[8]:
```
def extract_SNPs(df,pos_start,pos_stop,chr):
    '''
    extract SNPs near the position of interest
    insert: start position, stop position and chromosome
    '''
    df_filtered = df[(df.chr == chr) & (df.pos >= pos_start) & (df.pos <=␣
 ↪pos_stop)]
    return df_filtered
```

[9]:
```
df = extract_SNPs(df , start, end , CHR )
df
```

[9]:
```
             RS_ID chr        pos ref alt  scoreA    scoreB  \
100652  rs12281137  11  101076876   C   T  0.0120  0.011420
165457  rs17096381  11  101058518   A   G  0.0130  0.017656
253585    rs522930  11  101057855   T   C  0.0787  0.114591
350101  rs73583818  11  101123743   A   G  0.0116  0.011309
350102  rs73583818  11  101123743   A   G  0.0116  0.011309

       functional_element  n_experiment file_type cell_line cancer_type  \
100652         EZH2-human             1    narrow      PC-3    prostate
165457      H3K27ac-human             1    narrow      VCaP    prostate
253585      H3K27ac-human             1    narrow      VCaP    prostate
350101         CTCF-human             1    narrow     22Rv1    prostate
350102         CTCF-human             1    narrow     C4-2B    prostate

       cell_line_cancer
100652           cancer
165457           cancer
253585           cancer
350101           cancer
```

```
350102          cancer
```

2.4 Annotazione delle caratteristiche degli SNPs trovati in associazione con le regioni non codificanti di PGR

2.4.1 rs522930

```
SNPs in : PGR antisense RNA 1
SNPs ref: T  alt: C
maf1000genomes: 0.078674

type: lncRNA (long-non-coding)

Approved symbol : PGR-AS1
Approved name : PGR antisense RNA 1
RefSeq status: VALIDATED
Locus type : RNA, long non-coding forward strand.

Name variants
     PGR-AS1-202 : 1545 bp
     PGR-AS1-201 : 429 bp

Function: Non-coding exon

Region: Splice region, UTR

Functional element found associated:

    - H3K27ac-human

    - POLR2AphosphoS5-human
        This gene encodes the largest subunit of RNA polymerase II,
        the polymerase responsible for synthesizing messenger RNA in eukaryotes.
        https://www.genecards.org/cgi-bin/carddisp.pl?gene=POLR2A

Cell line: VCaP, prostate gland

Gene Enhancer
    ID: GH11J101126
    pos: chr11: 101,126,888-101,129,371

Expression level experiments (in prostate):
    RPKM (reads per kilo base per million reads placed ) : 0.549 ± 0.308
    Counts : 10368

HGNC ID  HGNC:52650
```

```
Symbol status Approved
Alias symbols: AT1, AT2, AT3
Alias names : PR-antisense-transcripts
Chromosomal location : 11q22.1

Variant type: SNV

Alleles:
    T>C [Show Flanks]

Chromosome:
    11:101187124 (GRCh38)
    11:101057855 (GRCh37)

Validated:
    by frequency,by cluster

MAF:
    C=0.018519/4 (Vietnamese)
    C=0.078674/394 (1000Genomes)
    C=0.114591/14389 (TOPMED)
    C=0.126667/76 (NorthernSweden)
    C=0.130977/4109 (GnomAD)
    C=0.157758/608 (ALSPAC)
    C=0.167745/622 (TWINSUK)
    C=0.185938/833 (Estonian)
```

- Ensembl https://www.ensembl.org/Homo_sapiens/Transcript/Sequence_cDNA?db=core;g=ENSG00000281187624;t=ENST00000632820;v=rs522930;vdb=variation;vf=83510084
- Clinvar https://www.ncbi.nlm.nih.gov/clinvar?LinkName=gene_clinvar&from_uid=101054525
- GeneCard https://www.genecards.org/cgi-bin/carddisp.pl?gene=PGR-AS1
- Expression level https://www.ncbi.nlm.nih.gov/gene/101054525/?report=expression

2.4.2 rs73583818

```
SNPs in : LINE Long interspersed nuclear elements
SNPs ref: A  alt: G
maf1000genomes: 0.01 (G)| Highest population MAF: 0.08

Chromosome:
    11:101253012 (GRCh38)

This variant has 2504 sample genotypes.
This variant has no desease correlated.

Type: intergenic variant
```

```
Code name site LINE element: L1ME3F
```

- UCSC Genome Browser: Visualizzazione del contesto genomico

http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&vir

- UCSC Genome Browser: Caratteristiche della LINE (Long interspersed nuclear elements) su cui cade lo SNPs

http://genome.ucsc.edu/cgi-bin/hgc?hgsid=837126199_aSaJyCH19PE7mm3iRvY5dXASNZoN&c=chr11&l=101

- Ensembl: contesto genico

https://www.ensembl.org/Homo_sapiens/Variation/Context?db=core;r=11:101252512-101253512;v=rs73583818;vdb=variation;vf=89100875

2.4.3 rs17096381

```
overlap element: lncRNA
    name: PGR-AS1-201
    name: PGR-AS1-202

MAF
    G=0.000223/1 (Estonian)
    G=0.000539/2 (TWINSUK)
    G=0.001297/5 (ALSPAC)
    G=0.012979/65 (1000Genomes)
    G=0.014811/465 (GnomAD)
    G=0.017656/2217 (TOPMED)

type: intron variant
    non coding transcript variant

    overlap with LINE element
    Family: L2


SNPs ref: A  alt: G
maf1000genomes: MAF: 0.012979 (G)|Highest population MAF: 0.09

Chromosome: 11:101187787 (GRCh38)

Functional Element: H3K27ac-human
   Acetylation at the 27th lysine residue of the histone H3 protein

Cell Line : VCaP
   Androgen-sensitive human prostate adenocarcinoma cells
```

- Ensembl: Dettagli sullo SNP

https://www.ensembl.org/Homo_sapiens/Variation/Mappings?db=core;r=11:101187287-101188287;v=rs17096381;vdb=variation;vf=87223580#ENST00000531772_87223580_G_tablePanel

- Genome Data viewer: Contesto biologico

https://www.ncbi.nlm.nih.gov/genome/gdv/browser/gene/?id=4524

- Genome Browser

http://genome.ucsc.edu/cgi-bin/hgc?hgsid=837126199_aSaJyCH19PE7mm3iRvY5dXASNZoN&c=chr5&l=1018

- dbSNP

https://www.ncbi.nlm.nih.gov/snp/?term=rs17096381

2.4.4 rs12281137

```
overlap element: lncRNA
   name: PGR-AS1-201
   name: PGR-AS1-202

MAF
   T=0.000259/1 (ALSPAC)
   T=0.00027/1 (TWINSUK)
   T=0.010385/325 (GnomAD)
   T=0.01142/1434 (TOPMED)
   T=0.011981/60 (1000Genomes)

type: intron variant
  non coding transcript variant

SNPs ref C: alt: G/T
maf1000genomes: MAF: 0.01 (T)|Highest population MAF: 0.08

Chromosome: 11:101206145 (GRCh38)

Functional Element: EZH2-human
    EZH2 function: maintaining the transcriptional
        repressive state of genes
        over successive cell generations

Cell line: PC-3
    PC3 is a human prostate cancer cell line used in
        prostate cancer research and drug development
```

- dbSNP : dati sulla variante genica

https://www.ncbi.nlm.nih.gov/snp/?term=rs12281137

- Ensembl: dati sulla variante genica

https://www.ensembl.org/Homo_sapiens/Variation/Mappings?db=core;r=11:101205645-101206645;v=rs12281137;vdb=variation;vf=86831890

2.5 Caratteristiche degli SNPs trovate su PGR-AS1

Nello studio su PGR sono stati trovati 5 SNPs, due idendici (con stessa rs ID). Questi due SNPs (rs73583818 A>G) sono trovati associati a due diverse linee cellulari, ed elementi funzionali. Su quattro SNPs univoci trovati (ovvero con la stessa rs ID), tre di essi (rs12281137, rs17096381, rs522930) sono stati trovati associati ad una regione non codificante del gene PGR. Nella regione colpita dai suddetti SNPs e' stato trovato l' elemento PGR-AS1.

2.6 Descrizione di PGR-AS1

PGR-AS1 è una trascritto a sette esoni di RNA non codificante lungo (long-non-coding RNA lncRNA), lungo 1545 bps (Ensembl). Esso si trova sul cromosoma umano 11, la posizione e' la seguente 100,999,808-101,080,322 nella versione GRCh37 del genoma umano. E' stato annotato manualmente da Havana Group nel Vertebrate Genome Annotation (VEGA) basato su Ensembl. Il gene ha tre sinonimi: AT1, AT2, AT3 e risiede nel forward strand. Le trascrizioni PGR-AS1 sono giuntate e poliadenilate, contengono elementi ripetitivi (elementi nucleari intervallati lunghi e (LINEs e SINE) e lunghe ripetizioni terminali (LTR) e sono trascritte su una regione di 70 kb di DNA genomico. PGR-AS1 e' composto da due isoforme: PGR-AS1-201, PGR-AS1-202, rispettivamente 429 bp e 1545 bp. Tutti e 3 gli SNPs trovati sono in regioni comuni ad entrambe le isoforme.

2.7 Letteratura su PGR-AS1

Nell' ultima annotazione per il gene PGR-AS1 riportata su Clinvar(Aug 12, 2011), la variante genica associata al trascritto e' stata identificata come 'Uncertain significance' ovvero di significato incerto. Nel dataset fornito, sono 4 su 5 SNPs associati a PGR-AS1, trovati in linee cellulari carcinogeniche. Si ritiene dunque, che questo lncRNA abbia una significativa importanza nella regolazione genica di PGR, e dunque varianti geniche associate ad esso possano contribuire a sviluppare disfunzioni geniche, avendo un possibile impatto nella associazione a forme tumorali. In particolare questo gene e' stato trovato altamente espresso nei seguenti tessuti: nella mucosa dell'endometrio, nell'ovario e nella ghiandola della prostata. In uno studio su PGR-AS1, l'RNA antisenso e' stato trovato in associazione con il complesso: Argonauta, ribonucleoproteina-k, RNA polimerasi II e con la proteina 1 gamma dell'eterocromatina. Lo studio dimostra come l'RNA antisenso, legato al complesso proteico, e' determinante per l'attivazione genica del promotore di PGR.

2.8 Dati su PGR-AS1 : Progesterone Receptor antisense RNA 1

- Clinvar: dati su PGR-AS1

https://www.ncbi.nlm.nih.gov/clinvar/?term=PGR-AS1%5Bgene%5D

- Ensembl: dati su PGR-AS1

http://www.ensembl.org/Homo_sapiens/Transcript/Summary?g=ENSG00000282728;r=11:101129077-101198910t=ENST00000632820

- Studio su PGR-AS1 correlato ad attivita' regolatoria

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2574822/

- Enseml: Elementi regolatori associati al gene PGR e PGR-AS1

http://sep2019.archive.ensembl.org/Homo_sapiens/Regulation/Context?db=core;fdb=funcgen;r=11:101125001-101131001;rf=ENSR00000440315

2.9 Conclusione

Grazie ai precedenti studi sull'elemento funzionale di PGR-AS1, il lncRNA e'stato trovato in associazione la proteina Argonauta, facente parte del complesso proteico RISC (complesso silenziatore indotto da RNA) con il quale ha attivita' regolatoria per la regione genica del promotore di PGR. L'elemento PGR-AS1 stato trovato responsabile dell'attivita' regolatoria: sia inibitoria, sia attivatoria sul promotore gene del recettore del progesterone. Le varianti geniche che sono state riportate sono state trovate nel Datset sono associate a un fenotipo canceroso. La maggiorparte delle MAF degli SNPs trovati si aggira attorno a poco piu'dell'1%.

In letteratura e'stata trovata l'associazione di PGR-AS1 con il complesso proteico di regolazione genica. Le varie MAF a basso score che indicano un impatto biologico forte. Tutti gli SNPs riportati sono associati sono trovati con fenotipo canceroso. Per tutti i motivi precedentemente detti, si ritiene che le varianti geniche (SNPs) che vanno a colpire l'elemento PGR-AS1, influiscono enormemente sulla regolazione del gene recettore del progesterone, andando a contribuire allo sviluppo di forme tumorali associate alla mucosa dell'endometrio, al ovario ed alla prostata.

3 PARTE SECONDA - Applicazione di un modello di Deep Learning per la predizione di una classe binaria su dati inerenti a polimorfismi a singolo nucleotide (SNPs).

3.1 Introduzione

Grazie all'aumento della potenza di calcolo dei computer negli ultimi anni, lo sviluppo di modelli predittivi attraverso il Machine Learning e' diventato maggiormente permissivo. Assieme a questo vantaggio dal punto di vista dell'hardware, nei recenti anni sono stati sviluppati pacchetti e librerie capaci di rendere l'implementazione di questi modelli attraverso l'utilizzo di poche righe di codice. In particolare, in questa applicazione ho utilizzato Tensorflow, una libreria sviluppata da Google per facilitare l'utilizzo dell'apprendimento automatico. Ho provato ad implementare un modello di Machine Learning che sfrutta le Reti Neurali, in particolar modo le Reti di Deep Learning, ovvero reti multistrato. Ho utilizzato i dati forniti per costruire un modello che cercasse di utilizzare tutti gli attributi possibili biologicamente e matematicamente sensati disponibili sul Dataset per la predizione di una categoria binaria, ovvero con solo due possibili valori associati a quell'attributo.

3.2 Elaborazione dei dati

Il modello usato ha il fine di utilizzare un sottoinsieme degli attributi di input, per predirre la classe binaria di output. Poiche' alcuni attributi sono stati ritenuti inutili, ognuno per motivi diversi, il modello e' stato basato su di un sottoinsieme di attributi esistenti. Alcuni attributi sono stati scartati per la loro presunta inutilita' dal mio personale punto di vista, poiche' poco predittivi per la classe di output, altri sono stati scartati in quanto l'attributo aveva un numero di valori univoci simile al numero di righe del dataset. Preciso il fatto che il modello da me scelto e' quello avere la migliore accuratezza fra diversi modelli testati. Detto questo, sono comunque possibili tutte le combinazioni desiderate di input di attributi e attributo da predirre. Il fine ultimo e' quello di predirre una classe binaria, ovvero con solo due possibili valori. E' stata scelta la classe attributo del Dataset che indica se lo SNPs e' in presenza di caratteristiche fenotipiche tumorgeniche.

Al fine di poter utilizzare gli attributi non numerici del dataset, ho convertito ogni valore di attributo in un numero intero. Questo procedimento e' essenziale in quanto la Rete Neurale e' capace solo di elaborare numeri interi e reali.

Esempio di associazione classe:intero con l'attributo di output.

```
cancer_type ha due possibili valori: normal e cancer

  Associazione dei valori con interi
      0 = 'normal'
      1 = 'cancer'
```

Per le associazioni fra { valore classe attributo= numero intero corrispondente } vedesi il codice del progetto.

3.3 Feature engineering: scoreC come la media fra le due MAF

E'stata "ingegnerizzata" uno nuovo attributo con l'obbiettivo di ridurre la dimensione degli attributi di input, questo al fine di ottimizzare la predizione del modello. Gli attributi ScoreA e ScoreB sono rappresentanti i valori di MAF(minor allele frequency) per lo SNP del record. Questi due valori provengono da fonti autorevoli, quali i database di 1000genomes e TOPMed. Le due MAF sono state unite in una unica con il nome 'ScoreC' che rappresenta la media delle due. Al fine di utilizzare tutti i record possibili, nei record dove non vi era una delle due MAF, e' stato preso il valore solamente di una delle due. I record che non presentavano almeno una delle due MAF sono stati scartati.

3.4 Cambiare gli attributi di input e l'attributo di output

Nel codice Python, alla linea "feature engineering" e' possibile scegliere le colonne che determinano gli attributi di input andando ad inserire manualmente gli attributi d'interesse ed andando a cambiare le colonne sul dataframe "df_data". E' possibile impostare manualmente anche la variabile "label_target" con la classe di output desiderata, nel mio caso era 'cell_line_cancer'.

3.5 Cambiare la struttura della rete neurale

Il modello e' stato costruito per avere un output con dominio di due elementi (appunto classificazione binaria). Va reso noto che e' possibile un suo riarrangiamento per la predizione di altre classi a maggiore numerosita', quali gli elementi funzionali correlati agli SNPs, o predirne la linea cellulare (queste possibili predizioni possono avere meno significato biologico). Nel caso in cui si voglia adattare il modello per la predizioni di classe a numerosita' maggiore di 2, va sottolineato che bisogna cambiare il numero di neuroni di output nel modello di Rete Neurale. Il numero di neuroni di output deve essere impostato pari al numero corrispondente al numero dei diversi valori possibili del attributo prescelto.

3.6 Eliminazione degli attributi inutili per la predittivita'

Sono stati eliminati i seguenti attributi:

n_experiment : numero di esperimenti relativi allo SNPs. E' stato visto che circa il 95% dei valori di questo attributo e' uguale ad 1, mentre il restante 5% e' uguale a 2. Avendo visto questa distribuzione univoca ho deciso di eliminare l'attributo dal modello.

file_type: il 100% dei valori di questo attributo e' uguale a 'narrow', dunque non essendoci variabilita' questo attributo e' ritenuto inutile per la predittivita' del modello.

3.7 Potenziali attributi utili, ma scartati per costruire il modello poiche' ritenuti poco predittivi

RSID : non rapprestativo per il modello di Machine Learning in quanto ogni rsID e' univoca ad un unico SNP, rendendo quindi impossibile generare pattern con valori univoci.

Chr: L'attributo dei cromosomi impostato come numero non porta ad alcuna correlazione con la possiblita' di fenotipo tumorale. Stesso ragionamento se consideraiamo i cromosomi come classi diverse, invece che come numeri.

Pos: La posizione, a mio parere, essendo raramente ripetuta, non e' sufficentemente predittiva in quanto classe.

Se considerato un dataset con maggiori ripetizioni di elementi (piu'esperimenti, maggiore ripetizioni di rsID), ma variabilita' ristretta (meno rsID), questi ultimi attributi possono essere utilizzati per la predizione di una classe. Anche il numero di esperimenti e' un potenziale attributo predittivo, ma scartato poiche' non ha una distribuzione sufficentemente uniforme fra i suoi valori.

3.8 Attributi di input finali

Al fine di predirre l'attributo 'cell_line_cancer' vengono utilizzati i seguenti attributi:

- `ref`
- `alt`

```
- functional_element
- cancer_type
- scoreC
```

3.9 Il modello di Rete Neurale

3.9.1 Keras

Keras è una API di alto livello per lo sviluppo di reti neurali scritta in Python e può essere utilizzata con Tensorflow, CNTK o Theano. Con Keras, i modelli si possono creare in due modi differenti: - API sequenziale: per la creazione livello per livello di modelli molto semplici. - API funzionale: per la creazione di modelli più complessi.

3.9.2 Struttura della Rete

Il modello e'stato costruito costruito su una tipologia sequenziale, della categoria multistrato. Presenta 6 strati di neuroni in tutto, 1 di input, 1 di output e 4 strati di neuroni intermedi, chiamati anche (hidden layer). Il modello di rete neurale presentato ha la seguente struttura:

- 5 neuroni di input
- 8 neuroni hidden layer
- 16 neuroni hidden layer
- 32 neuroni hidden layer
- 16 neuroni hidden layer
- 2 neuroni di output

3.9.3 Dettagli teorici sulle reti neurali

- Le reti neurali vengono addestrate utilizzando un processo di ottimizzazione che richiede una funzione di perdita per calcolare l'errore del modello.
- Maximum Likelihood fornisce un framework per la scelta di una funzione di perdita durante l'addestramento di reti neurali e modelli di apprendimento automatico in generale.
- L'entropia incrociata (cross enropy) e' l'errore quadratico medio sono i due principali tipi di funzioni di perdita da utilizzare durante l'allenamento dei modelli di rete neurale.
- La funzione di costo riduce tutti i vari aspetti positivi e negativi di un sistema eventualmente complesso fino a un singolo numero, un valore scalare, che consente di classificare e confrontare le soluzioni candidate.
- Epoche rappresenta il numero di processi iterativi per ottimizzare il modello

3.9.4 Descrizione della struttura

I neuroni di input devono corrispondere al numero di attributi utilizzati per la predizione della classe finale. Il numero di neuroni di output deve corrispondere al numero di classi dell'attributo che si vuole predirre. Se si vuole cambiare modello bisogna inanzitutto adeguare la struttura della rete neurale, cambiando i neuroni di input con il numero di attributi di input, e il numero di

neuroni di output con i valori possibili dell'attributo di output. Nel mio caso, l'attributo di output presentava solo due tipologie di valore: 'normal' e 'cancer', per questo il numero di neuroni di output corrisponde a due.

3.9.5 Strati nascosti

La struttura intermedia dei strati nascosti(hidden layer) di neuroni e' molto flessibile, per questo e' possibile cambiarla. Non e' consigliabile aumentare enormemente il numero di neuroni in questi strati, poiche' la predittivita' potrebbe non alzarsi, ma potrebbe addirittura abbassarsi. Stesso discorso per l'aumento del numero di strati nascosti. Avendo provato diverse strutture di rete neurale, ho deciso questa serie di valori per gli strati nascosti (8x16x32x16). Sono arrivato a questa conclusione avendo visto strutture simili in esempi online, ed avendo riscontrato il fatto di aver visto che strutture eccessivamente diverse da questa avevano un calo nella predittivita'.

3.10 Modello

Dati gli andamenti delle curve di accuratezza e di perdita, si è scelto di fermare il processo di apprendimento a 10 epoche con il quale abbiamo ottenuto i seguenti risultati:

```
Test loss: 0.2235
Test accuracy: 0.8865

Record testati: 113489
```

3.11 Risultato

Calcolando l'accuratezza su un set di dati di test, pari a circa il 20% del dataset originale, è stato valutata un accuratezza della predizione della classe di output pari a 0.8865 (88%). Mentre la perdita del test (Test loss) derivata dalla funzione di costo, e' risulata essere pari a 0.224. Va sottolineato che ogni run-time del programma di rete neurale porta a soluzioni leggermente o drasticamente diverse in quanto i parametri iniziali sono assegnati casualmente. I risultati dello score per la predizione indicano una buona accuratezza, potendo cosi'associare circa l'88% in maniera corretta, uno SNPs con il fenotipo cellulare. Ritengo che questa rete neurale elaborata e' potenzialmente capace di discriminare varianti geniche determinanti per lo sviluppo tumorale da quelle non determinanti, di aplotipi che portano ad un fenotipo tumorale.