UNIVERSITÀ
DEGLI STUDI
FIRENZE

Da un secolo, oltre.

HR EXCELLENCE IN RESEARCH

Nucleotide Transformer for biological sequences

Laureando: Gio Formichella

Relatore: Prof. Paolo Frasconi

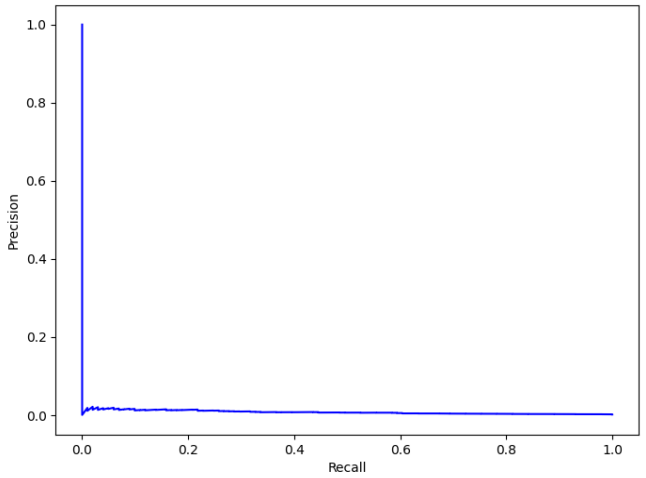
Firenze, 19 Febbraio 2024

Edited C prediction

- La proteina APOBEC1 sostituisce alcuni nucleotidi con citosine.
- Il criterio con cui la proteina sceglie il nucleotide da editare non è ancora stato scoperto e, ad oggi, non esistono studi algoritmico-predittivi di questo fenomeno.
- Task: Predire per ogni citosina della sequenza di DNA se è derivante dall'editing biologico o meno. Si tratta quindi di un problema di classificazione binaria di tipo sequence-to-sequence.

Edited C prediction

- Un primo approccio: utilizzo dei k-mers, quindi dei nucleotidi vicini alle C, come variabili.
- Risultato con Gradient boosting:



- Ritenendo di poter fare meglio con una diversa rappresentazione del contesto biologico, ho provato ad utilizzare gli embeddings di un modello Transformer.

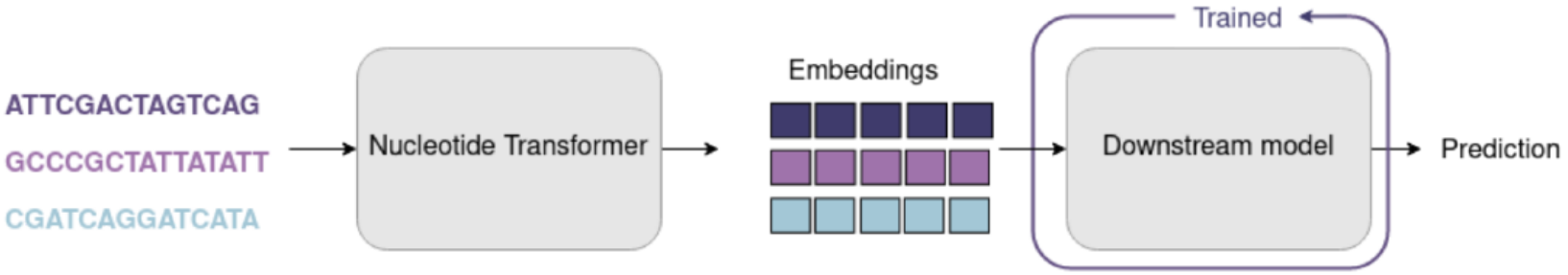
Nucleotide Transformer

Modelli "Nucleotide Transformer" (*Dalla-Torre et al. 2023*):

- Modelli tra 50M – 2.5B di parametri allenati su DNA proveniente da 500 – 4052 diversi genomi.
- Modelli allenati sul supercomputer Nvidia Cambridge–1: 128 A100 GPU con tempi tra 1 e 28 giorni in base al modello.
- 6-mers tokens come trade-off tra lunghezza della sequenza e dimensione degli embeddings.
- Lunghezza massima delle sequenze processata pari a 1000.
- 20-esimo layer più performante dei 24.

Nucleotide Transformer

- Modello scelto: "500M_1000G"
- Ho utilizzato il modello pre-addestrato come una black box mediante API ed ho costruito la seguente pipeline:



Pipeline: input

C A G T C A C A T C T G T A

Pipeline: tokenizzazione

C A G T C A C A T C T G T A

Tokenizer

<CLS>, <CAGTCA>, <CATCTG>, <T>, <A>, <pad>, ...

- Start token
- Token di nucleotidi
- Padding tokens

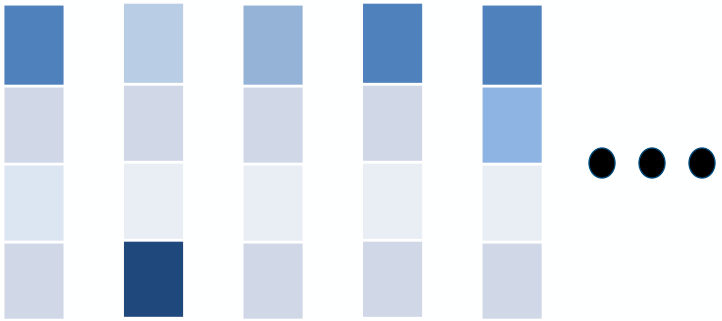
Pipeline: inferenza

C A G T C A C A T C T G T A

Tokenizer

<CLS>, <CAGTCA>, <CATCTG>, <T>, <A>, <pad>, ...

Inferenza: layer 20

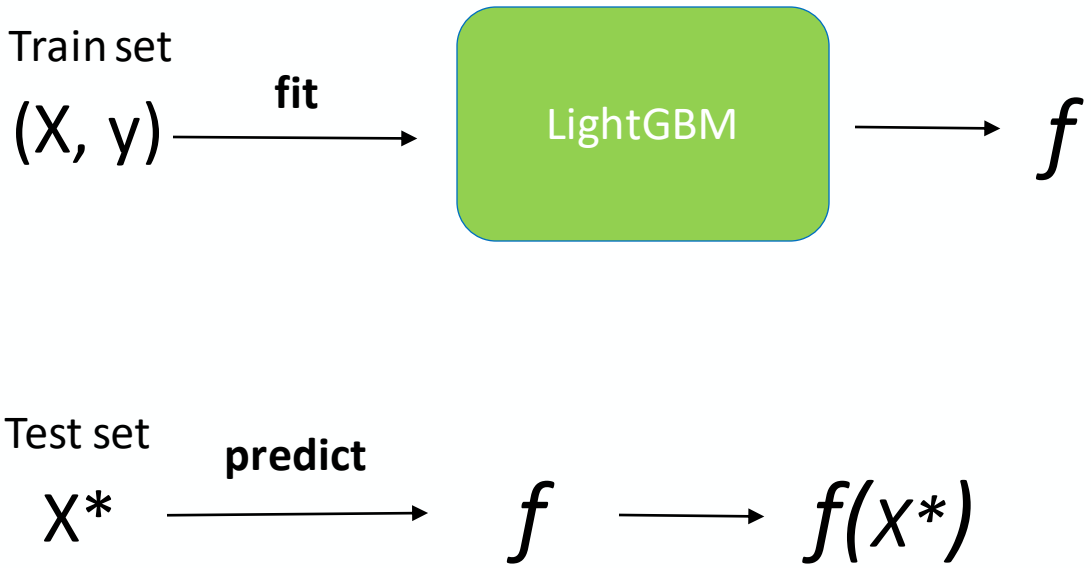


Embedding:

- Rappresentazione del token all'interno della sequenza
- Array di float di lunghezza 1280
- Scartate le rappresentazioni dei padding token e, per task sequence-to-sequence, anche dei CLS

Pipeline: classificatore

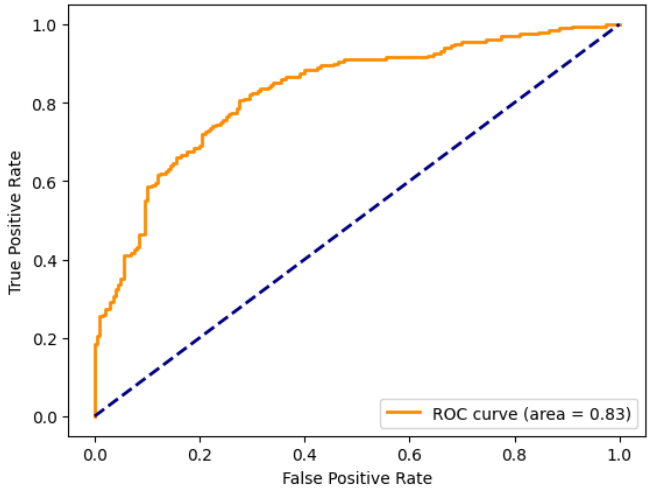
- Utilizzato LightGBM (*Ke et al. 2017*) invece che le NN.



- X,X*: embeddings
- y: labels
- *f*: trained LightGBM
- *f*(X*): predictions

Enhancer type prediction

- Verifica del corretto funzionamento della pipeline.
- Problema di classificazione binaria di tipo sequence-to-class: ogni sequenza di DNA è "enhancer" o "non enhancer".
- Rappresentazione di sequenza = media aritmetica degli embedding dei token della sequenza.
- Risultato:
  - Accuratezza del 75%



Protein secondary structure prediction

- Task di tipo sequence-to-sequence
- Proteina = sequenza di aminoacidi

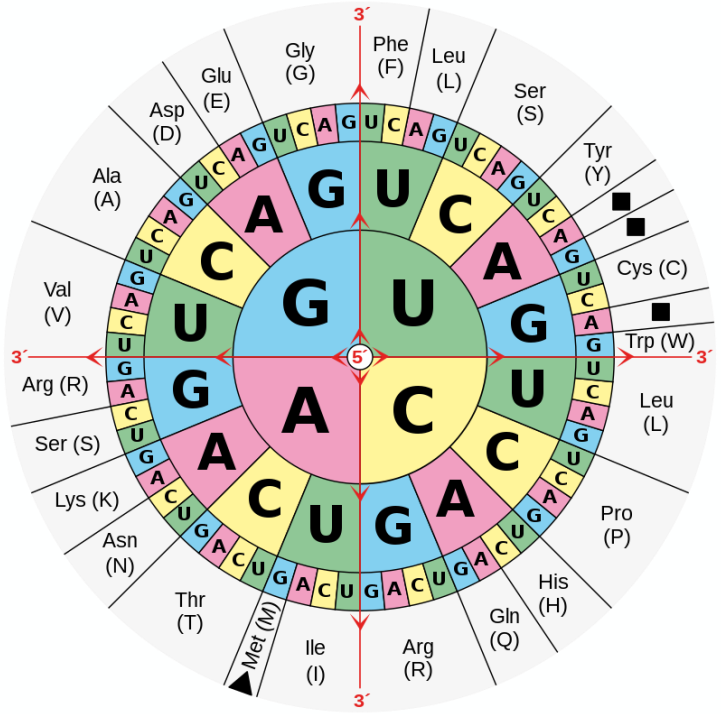
Sequenza di aminoacidi	Classi aminoacidi
M	C
K	C
T	C
A	H
Y	H
M	H
I	E
K	E
Q	C
X	C
S	C

- H: α-helix
- E: β-sheet
- C: coil

Protein secondary structure prediction

- Conversione degli aminoacidi in proteine
- Non sempre c'è una corrispondenza biunivoca

Sequenza di aminoacidi	Sequenza di RNA
M	A
K	U
T	G
A	A
Y	U
M	A
I	A
K	U
Q	A
X	.
S	.



Protein secondary structure prediction

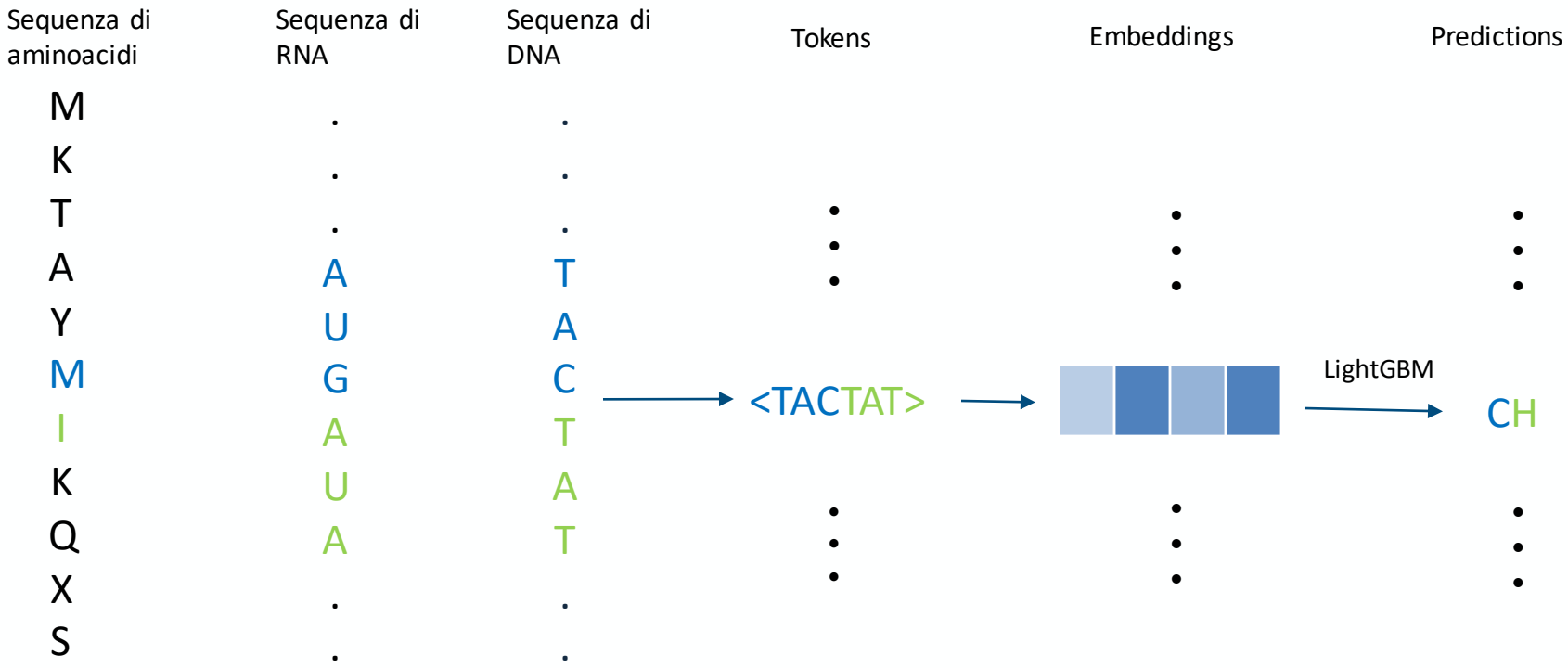
- Conversione di RNA in DNA

Sequenza di aminoacidi	Sequenza di RNA	Sequenza di DNA
M	.	.
K	.	.
T	.	.
A	A	T
Y	U	A
M	G	C
I	A	T
K	U	A
Q	A	T
X	.	.
S	.	.

RNA	DNA
A	T
U	A
C	G
G	C

Protein secondary structure prediction

- Embeddings di coppie di aminoacidi



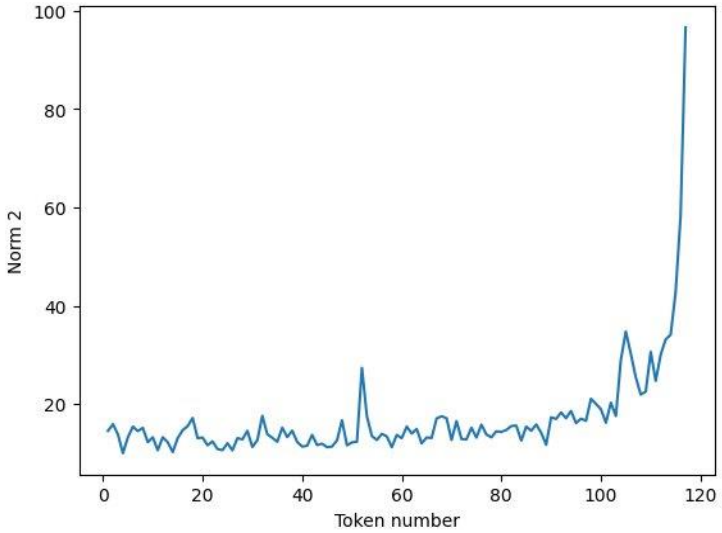
Protein secondary structure prediction

- Sequenze di aminoacidi scomposte in sottosequenze:
  - In presenza del carattere X
  - Quando la lunghezza supera la soglia ammessa dal modello
- Scartate sottosequenze di lunghezza < 200 caratteri.

- Risultato:
  - Accuratezza del 50% nella predizione delle 3 classi.

Protein secondary structure prediction

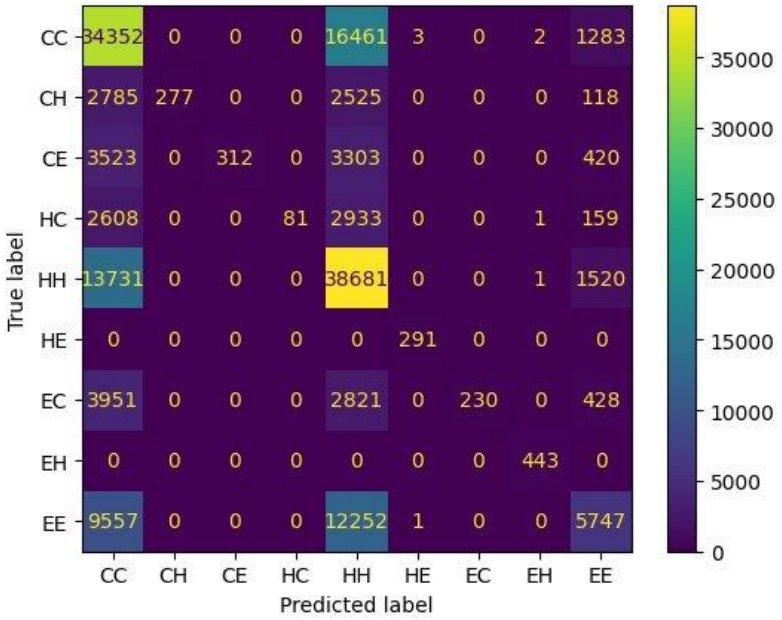
- L'accorciamento delle sequenze provoca perdita di contesto e, di conseguenza, peggiora gli embeddings.
- Verificato confrontando embeddings dei token comuni ad una sequenza e a quella ottenuta rimuovendo gli aminoacidi finali.
- Risultato:



- Modifica 1: Rimossi gli embeddings dei token ai margini delle sottosequenze. L'accuratezza, però, non è migliorata.

Protein secondary structure prediction

- Verifica della confusion matrix: modello predice eccessivamente "CC" e "HH".



- Causa: le sequenze da predire sono composte da lunghe stringhe dello stesso carattere ripetuto molteplici volte, con prevalenza di C ed H mentre le E sono più rare.

Protein secondary structure prediction

- Modifica 2: impiego di 2 classificatori: uno per la label di sinistra e uno per la label di destra. Seppur rimosso il bias, le performance non sono cambiate.

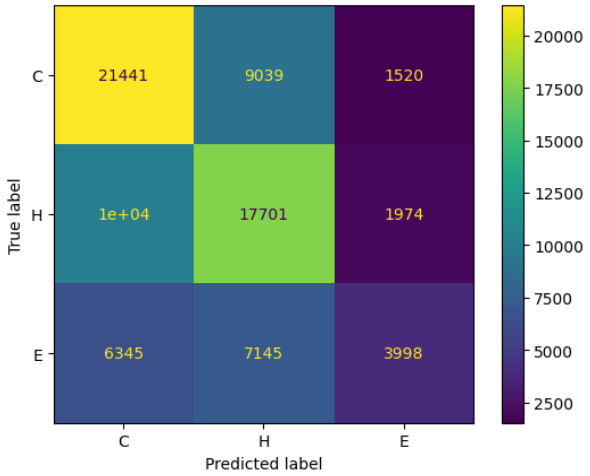
- Il problema risiede negli embeddings.

Protein secondary structure prediction

- Modifica 2: impiego di 2 classificatori: uno per la label di sinistra e uno per la label di destra. Seppur rimosso il bias, le performance non sono cambiate.

- Il problema risiede negli embeddings.

- Modifica 3: rimossa randomicità nella scelta dei codoni.
- Risultato:
  - Accuratezza al 55%.



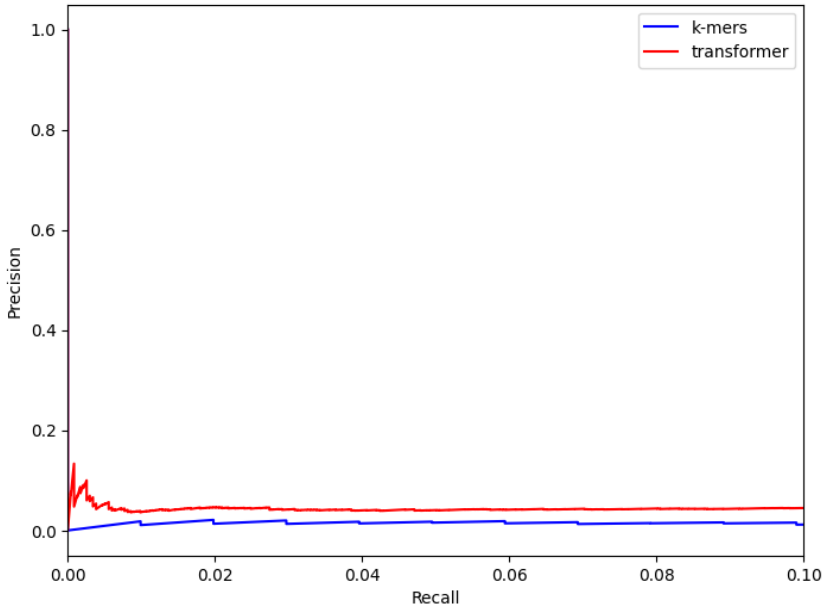
Protein secondary structure prediction

- Risultato ragionevole considerando le seguenti limitazioni:
  - Conversione non sempre esatta tra aminoacidi e codoni.
  - Rigidità della struttura dei token: non è possibile creare una singola rappresentazione per un singolo aminoacido della sequenza.
  - Assenza di informazione evolutiva: non sono stati utilizzati gli allineamenti multipli.
  - Assenza di fine-tuning.

Edited C predicion

- K-mers:
  - Finestra di 48 nucleotidi a destra e 48 a sinistra della C.
  - Considerati i 2,3,4-mers.
- Data la struttura dei token del NT:
  - Token positivo: se contenente almeno una C editata.
  - Token negativo: se contenente solo C non editate.
  - Token non considerato: se non contenente alcuna C.
- Dato lo sbilanciamento del dataset (le citosine editate sono molto meno rispetto alle non editate) è stato fatto undersampling sul training set del classificatore e sono state confrontate le curve precision-recall.

Edited C predicion



- NT è leggermente più preciso dei k-mers.
- Data la complessità biologica del problema, neanche il NT ha colto il relativo segnale biologico.

UNIVERSITÀ
DEGLI STUDI
FIRENZE
Da un secolo, oltre.

HR EXCELLENCE IN RESEARCH

Grazie per l'attenzione