Assessing the Efficiency of DNABERT in the Binary Classification of Transposons and Retrotransposons: A Comparative Analysis

Eduardo Scarpa
Università degli Studi di Salerno
Dipartimento di Informatica
A.A. 2023/2024

Alfredo Cannavaro
Università degli Studi di Salerno
Dipartimento di Informatica
A.A. 2023/2024

Abstract

Questo studio presenta un'analisi comparativa tra DNABERT e il "Gene Fusion Classifier" per la classificazione binaria di trasposoni e retrotrasposoni, evidenziando l'applicabilità e l'efficacia di tecniche avanzate di machine learning nel campo della genetica. Attraverso un'approfondita metodologia di ricerca, che include la selezione e la preparazione dei dati, il processo di addestramento e di ottimizzazione del modello, e una rigorosa strategia di valutazione, abbiamo esplorato le potenzialità di questi modelli per decifrare la complessa struttura genetica dei trasposoni. I risultati dimostrano che DNABERT, adattato dall'architettura di BERT per analizzare sequenze di DNA, supera in termini di accuratezza, precisione e capacità analitica il "Gene Fusion Classifier", specificamente progettato per identificare fusioni geniche. Nonostante le limitazioni metodologiche e la necessità di risorse computazionali significative, questo confronto offre preziose intuizioni sul potenziale dei modelli di apprendimento automatico nel migliorare la comprensione e la classificazione di elementi genetici critici, ponendo le basi per future ricerche nel settore della bioinformatica e della biologia computazionale.

1 INTRODUZIONE

L'architettura dei Transformer ha rivoluzionato l'elaborazione del linguaggio naturale (NLP), offrendo una base robusta per lo sviluppo di modelli di apprendimento profondo capaci di catturare complesse relazioni nei dati sequenziali [1]. BERT (Bidirectional Encoder Representations from Transformers) ha esteso questa innovazione, presentando una metodologia pre-addestrata che migliora significativamente la comprensione del contesto nel testo [2]. Questi avanzamenti hanno trovato applicazioni oltre l'NLP, estendendosi al dominio della genetica, dove DNABERT è stato introdotto come strumento per l'analisi di sequenze di DNA, pur non essendo stato originariamente sviluppato per la classificazione specifica di trasposoni [3]. I trasposoni, elementi genetici mobili che contribuiscono alla diversità genetica e hanno impatti significativi sulla regolazione genica, presentano una sfida unica per l'analisi genomica data la loro natura variabile e la loro abbondanza nel genoma. La nostra ricerca si propone di colmare questa lacuna, estendendo l'applicazione di DNABERT alla classificazione binaria dei trasposoni. Questo approccio sfrutta la profondità e la flessibilità dei modelli basati su Transformer per interpretare le complesse sequenze di DNA associate ai trasposoni, aprendo nuove vie per la comprensione della loro funzione e distribuzione nel genoma.

1.1 Stato dell'arte

La classificazione dei trasposoni sfrutta tecnologie avanzate di apprendimento automatico, tra cui l'applicazione di modelli come DNABERT, originariamente concepiti per il riconoscimento e l'analisi di sequenze di DNA. Questi modelli, basati sull'architettura dei Transformer, hanno aperto la strada all'elaborazione e interpretazione avanzata del DNA, analogamente al trattamento del linguaggio naturale. Mentre DNABERT stesso non si concentra direttamente sulla classificazione dei trasposoni, il suo approccio rappresenta una base promettente per future ricerche in questo ambito, suggerendo che tecniche simili potrebbero essere adattate per migliorare significativamente la precisione e l'efficacia nella classificazione dei trasposoni, un'area critica per la comprensione della genetica e dell'evoluzione.

1.2 Contributo di questo lavoro

Nel corso di questo studio, si è proceduto con la rielaborazione e la personalizzazione di un dataset di trasposoni proveniente dalla banca dati TREP[1], affiliata all'Università di Zurigo, per l'utilizzo con DNABERT. Questa operazione ha incluso la trasformazione delle sequenze estratte in k-mer di lunghezza 4, un passaggio critico per rendere il dataset compatibile con DNABERT, che interpreta le sequenze di DNA con metodi ispirati all'elaborazione del linguaggio naturale. Tale adattamento facilita l'applicazione di modelli basati su Transformer per una classificazione precisa dei trasposoni, apportando un contributo significativo alla comprensione della loro distribuzione e funzione genetica.

2 LAVORI CORRELATI

Nel panorama attuale della ricerca, si assiste a un'intensa attività investigativa finalizzata all'identificazione di firme genomiche specifiche per diversi tipi di cancro. Questo sforzo comprende l'esplorazione di metodologie avanzate per l'analisi del DNA, tra cui spicca l'uso di DNABERT per la classificazione precisa di elementi genetici quali retrotrasposoni e trasposoni. La presente sezione si dedicherà alla presentazione di studi correlati al nostro lavoro, evidenziando non solo la convergenza degli obiettivi di ricerca ma anche l'impiego di strumenti analoghi nel processo analitico. L'accento sarà posto su come queste ricerche parallele contribuiscano alla nostra comprensione della complessità genomica e alla sua relazione con le patologie oncologiche.

[1] https://trep-db.uzh.ch/downloadFiles.php

2.1 Deep Learning-based Clustering Approaches for Bioinformatics

Il documento [4] esamina l'uso del Deep Learning (DL) per il clustering in bioinformatica, evidenziando come questa tecnologia possa migliorare l'analisi di dati non strutturati e ad alta dimensionalità come sequenze, espressioni, testi e immagini. Si concentra su come il DL possa essere utilizzato per scoprire pattern nascosti nei dati biologici, contribuendo a una comprensione più profonda dei processi biologici.

2.1.1 Workflow. Il workflow descritto nel paper include l'impiego di algoritmi di clustering basati su DL per analizzare vari casi d'uso bioinformatici, come l'imaging bio-medico, la genomica del cancro e il clustering di testi biomedici. Questo approccio sfrutta il DL per estrarre caratteristiche rappresentative dai dati, che vengono poi utilizzate per migliorare l'accuratezza del clustering.

2.1.2 Metodologia. La metodologia impiegata combina tecniche di apprendimento profondo per l'estrazione di caratteristiche e algoritmi di clustering tradizionali. Il processo si articola in due fasi principali: l'inizializzazione dei parametri e l'apprendimento delle rappresentazioni mediante reti neurali profonde, seguito dall'ottimizzazione dei parametri per migliorare l'obiettivo di clustering.

2.1.3 Risultati. I risultati mostrati evidenziano come l'impiego del DL nel clustering possa portare a miglioramenti significativi nell'analisi di dati bioinformatici, superando i limiti dei metodi tradizionali di clustering. In particolare, si osserva un miglioramento nell'accuratezza del clustering e nella capacità di trattare dati ad alta dimensionalità, contribuendo a scoperte significative in vari campi della bioinformatica.

2.2 MeShClust: An Intelligent Tool for Clustering DNA Sequences

MeShClust [5] introduce un approccio innovativo al clustering di sequenze di DNA utilizzando l'algoritmo mean shift, adattato dalle sue applicazioni di successo in campi come l'elaborazione delle immagini e la visione artificiale. Questo rappresenta una delle poche applicazioni bioinformatiche dell'algoritmo mean shift, migliorando significativamente l'accuratezza e l'affidabilità del processo di clustering.

2.2.1 Workflow. Il workflow di MeShClust incorpora uno strumento software composto da un classificatore e l'algoritmo mean shift. Il classificatore predice la somiglianza tra sequenze, mentre l'algoritmo mean shift, adattato per il clustering di sequenze di DNA, rifinisce iterativamente i centri dei cluster basandosi sulla somiglianza delle sequenze, determinando automaticamente il numero di cluster senza richiedere un conteggio predefinito.

2.2.2 Metodologia. La metodologia comporta la rappresentazione delle sequenze come istogrammi di k-mer, che vengono poi elaborati dal classificatore per prevedere la somiglianza tra le sequenze. L'algoritmo mean shift ricalcola iterativamente i centri dei cluster basandosi su queste somiglianze, fondendo i cluster sovrapposti e perfezionando così il processo di clustering senza la necessità di parametri definiti dall'utente per il numero di cluster.

2.2.3 Risultati. MeShClust ha dimostrato un'alta accuratezza nel clustering di sequenze di DNA, superando gli strumenti basati su algoritmi greedy tradizionali. Si è dimostrato capace di produrre cluster ottimali anche quando gli sono stati forniti parametri di somiglianza di sequenza inaccurati dall'utente, evidenziando la sua flessibilità ed efficacia nel gestire dati di sequenze di DNA.

3 BACKGROUND

In questa sezione verranno fornite alcune definizioni per capire meglio le tecniche e gli strumenti utilizzati che verranno citati nei capitoli successivi.

3.1 Introduzione ai Trasposoni

I trasposoni, noti anche come "elementi genetici mobili", sono sequenze di DNA in grado di cambiare la loro posizione all'interno del genoma. Questa mobilità li rende attori chiave nella variazione genetica e nell'evoluzione, influenzando la struttura genetica e la funzionalità genomica attraverso l'inserimento in nuove locazioni, potenzialmente modificando l'espressione genica o causando riarrangiamenti genetici. Esistono diversi tipi di trasposoni, inclusi quelli che si muovono direttamente attraverso il meccanismo "taglia e incolla" e quelli che utilizzano un intermediario di RNA nel processo "copia e incolla" [6]. La loro presenza ubiquitaria nei genomi di numerosi organismi sottolinea il loro significativo ruolo evolutivo, contribuendo alla diversità biologica e all'adattamento.

3.2 Storia della Classificazione dei Trasposoni

La storia della classificazione dei trasposoni è profondamente radicata nella scoperta di Barbara McClintock sui "geni saltatori" negli anni '40 [7], un evento che ha aperto la strada alla comprensione dell'importanza biologica e evolutiva dei trasposoni. Con l'evolversi delle tecnologie di sequenziamento e delle tecniche bioinformatiche, è stata possibile una classificazione più raffinata dei trasposoni, basata sui loro meccanismi di azione e sulle loro sequenze. Questi progressi hanno rivelato la diversità dei trasposoni, suddivisi in classi principali come i trasposoni di classe I, che utilizzano un intermediario di RNA, e quelli di classe II, che si spostano direttamente come segmenti di DNA [7]. La ricerca ha anche messo in luce il ruolo cruciale dei trasposoni nella diversificazione genetica, nella plasticità genomica e nell'adattamento, sottolineando il loro impatto dinamico nei genomi attraverso le specie. Le scoperte continuano a promettere nuove intuizioni sul complesso rapporto tra trasposoni e genomi ospiti, enfatizzando la loro importanza nel modellare i paesaggi genetici e contribuire alla biodiversità.

3.3 Apprendimento Profondo nella Bioinformatica

L'adozione dell'apprendimento profondo nella bioinformatica ha significativamente ampliato le capacità di analisi delle sequenze di DNA, portando a una comprensione più profonda dei meccanismi genetici sottostanti. Questi modelli avanzati, tra cui le reti neurali convoluzionali (CNN) e le reti neurali ricorrenti (RNN), hanno reso possibile l'esplorazione di dati genetici su una scala senza precedenti [8]. La precisione migliorata nella predizione delle funzioni geniche e l'identificazione di pattern genetici ha accelerato progressi in vari campi, tra cui la genomica comparativa e l'epigenetica. Questa

evoluzione metodologica non solo ha accelerato la scoperta scientifica ma ha anche aperto nuove vie per applicazioni cliniche, come la medicina personalizzata, dove la comprensione dettagliata delle basi molecolari delle malattie sta guidando lo sviluppo di terapie più mirate ed efficaci [6]. Le terapie mirate sono trattamenti personalizzati basati sulla comprensione genetica e molecolare delle malattie. Sfruttando le scoperte dell'apprendimento profondo nella bioinformatica, gli scienziati possono identificare con precisione le anomalie genetiche specifiche di un individuo che contribuiscono alla malattia. Questo permette lo sviluppo di farmaci che colpiscono specificamente queste anomalie, migliorando l'efficacia del trattamento riducendo al contempo gli effetti collaterali rispetto ai metodi terapeutici più tradizionali. Questo approccio rappresenta una svolta nella cura di molte malattie, inclusi diversi tipi di cancro come il Carcinoma mammario, il Carcinoma polmonare, il Melanoma, il Carcinoma gastrico e la Leucemia.

3.4 Architettura dei Transformer

L'architettura dei Transformer, introdotta da Vaswani [1], ha segnato una svolta nell'elaborazione del linguaggio naturale (NLP) grazie alla sua capacità unica di gestire sequenze di dati in parallelo e di focalizzare l'attenzione su contesti distanti nel testo. Questo meccanismo di attenzione multi-testa consente ai Transformer di valutare l'importanza di ogni parola all'interno di una frase, migliorando la comprensione del contesto e la generazione di testo. La flessibilità e l'efficienza dei Transformer hanno trovato applicazioni anche nella genetica, dove possono essere utilizzati per analizzare sequenze di DNA, identificando relazioni funzionali complesse e pattern nascosti. Un esempio significativo di questa applicazione è illustrato dal lavoro di Rives [9], che dimostra come i Transformer possano essere adattati per scoprire strutture biologiche e funzionalità a partire da vasti set di dati di sequenze proteiche. Questi progressi aprono nuove prospettive per la ricerca genetica, sottolineando il potenziale dei Transformer non solo nell'NLP ma anche come strumenti potenti per l'analisi dei dati biologici e genetici.

3.5 BERT e DNABERT

BERT (Bidirectional Encoder Representations from Transformers) ha segnato un punto di svolta nell'apprendimento profondo e nella preparazione del linguaggio naturale grazie alla sua innovativa architettura basata sui Transformer, introdotta da Devlin nel loro influente lavoro del 2018 [2]. Questo modello ha rivoluzionato il modo in cui le macchine comprendono il linguaggio umano, utilizzando un approccio bidirezionale per analizzare il contesto di ogni parola all'interno di un testo, permettendo un'interpretazione più ricca e accurata rispetto ai precedenti modelli unidirezionali o contestualmente ciechi.

La potenza di BERT risiede nella sua pre-addestrazione, che avviene su un vasto corpus di testo e consente al modello di imparare una rappresentazione linguistica profonda prima di essere fine-tunato su compiti specifici di NLP. Questo approccio ha dimostrato di migliorare significativamente le prestazioni in una vasta gamma di compiti, dalla comprensione del testo alla generazione linguistica, stabilendo nuovi standard di riferimento per la comunità scientifica.

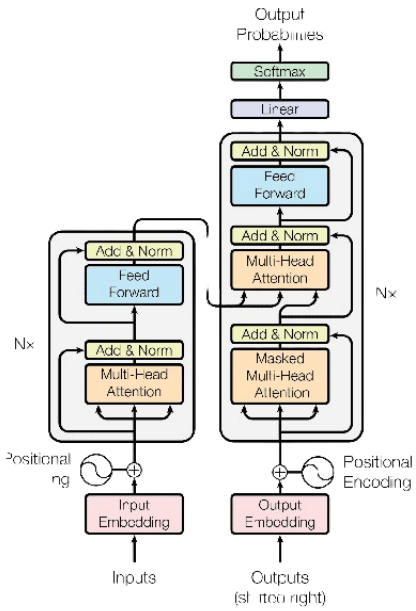

Figure 1: L'architettura del modello Transformer

DNABERT [3] estende l'applicabilità di questa tecnologia sovversiva al campo della genetica, adattando il concetto di BERT per analizzare sequenze di DNA con una precisione senza precedenti. Pre-addestrato su sequenze genetiche, DNABERT sfrutta la sua capacità di catturare relazioni complesse e contestuali tra nucleotidi per identificare elementi funzionali cruciali all'interno del genoma, come promotori, enhancer e siti di legame per i fattori di trascrizione. Quest'innovazione rappresenta un salto qualitativo nell'analisi genetica, offrendo strumenti potenti per la ricerca biomedica, riconoscimento o l'identificazione di biomarcatori per le malattie e lo sviluppo di terapie personalizzate.

L'adattamento di BERT per l'analisi del DNA apre la strada a una nuova era nella bioinformatica, dove le tecniche di apprendimento profondo possono essere applicate per decifrare la complessità del genoma umano e di altri organismi. DNABERT illustra come le metodologie avanzate di NLP possano essere trasferite con successo ad altri ambiti scientifici, sottolineando l'importanza dell'interdisciplinarità nella ricerca moderna e la convergenza tra biologia computazionale e intelligenza artificiale.

3.6 Applicazioni e Sfide Correnti

L'impiego di DNABERT e tecnologie correlate nell'analisi genetica, in particolare nella classificazione dei trasposoni, rappresenta un'avanzata significativa che sfrutta la potenza dell'apprendimento profondo per decodificare la complessità del genoma. Questi modelli, addestrati per interpretare sequenze di DNA, hanno dimostrato un'eccezionale capacità di identificare e classificare trasposoni, elementi genetici mobili che giocano un ruolo cruciale nell'evoluzione del genoma e nella regolazione genica. L'applicazione di DNABERT ha portato a scoperte importanti, come la mappatura precisa di

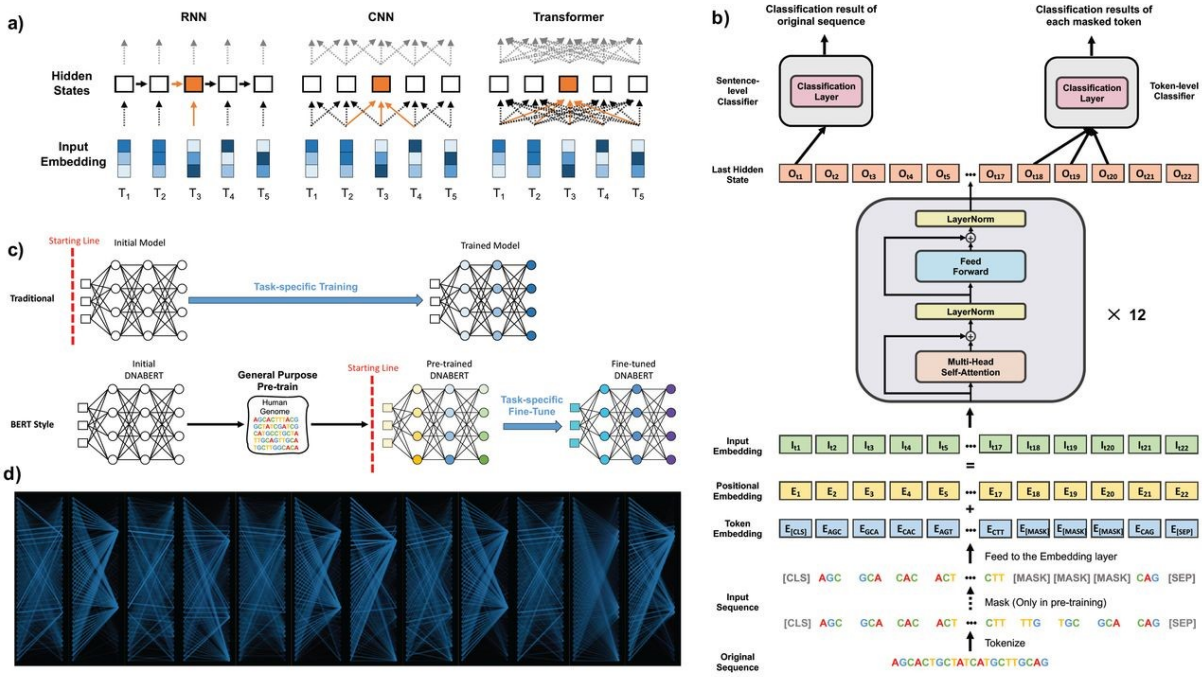

Figure 2: Informazioni dettagliate sull'architettura e sulle caratteristiche del modello DNABERT

sequenze trasponibili e la comprensione delle loro dinamiche evolutive [3].

Nonostante questi successi, le sfide persistono, evidenziando i limiti delle attuali tecnologie di sequenziamento e analisi. Una delle principali difficoltà riguarda l'interpretazione dei dati generati: la classificazione dei trasposoni richiede non solo un'analisi computazionale accurata ma anche una profonda comprensione biologica delle loro funzioni e meccanismi d'azione. Inoltre, la vastità e la complessità del genoma umano e di altri organismi pongono sfide significative in termini di elaborazione dati e necessità di risorse computazionali.

Ulteriori sfide includono l'adattabilità dei modelli a diversi tipi di organismi e la generalizzazione dei risultati ottenuti. Mentre DNABERT e modelli simili offrono approcci promettenti, la loro efficacia può variare a seconda della specificità delle sequenze di DNA analizzate e della disponibilità di dati annotati per l'addestramento. Questo sottolinea l'importanza di un lavoro continuo per migliorare le tecniche di apprendimento profondo, sviluppando algoritmi più sofisticati che possano gestire la diversità genetica e fornire intuizioni più precise sulla funzione e l'evoluzione dei trasposoni.

In conclusione, mentre DNABERT e tecnologie affini rappresentano un passo avanti nell'analisi genetica, la loro applicazione nella classificazione dei trasposoni sottolinea sia il potenziale che le sfide dell'integrazione dell'intelligenza artificiale nella ricerca genetica. La superazione di queste sfide richiederà un approccio multidisciplinare, che combini competenze in bioinformatica, biologia molecolare e intelligenza artificiale, per sbloccare pienamente il potenziale di queste tecnologie nell'illuminare i misteri del genoma.

4 METODOLOGIA

4.1 Introduzione della Metodologia

Il presente capitolo delibera in dettaglio sulla metodologia impiegata per analizzare sequenze di DNA tramite l'utilizzo di DNABERT, una variante del modello BERT (Bidirectional Encoder Representations from Transformers) specificamente adattata per interpretare sequenze genetiche. L'impiego di questa tecnologia s'inquadra nell'ambito di un approccio innovativo alla bioinformatica, mirato a sfruttare le capacità dei modelli di deep learning per classificare i trasposoni nelle sequenze di DNA. Questi elementi genetici mobili, capaci di cambiare posizione all'interno del genoma, hanno un ruolo cruciale nella variazione genetica e nell'evoluzione, rendendo la loro identificazione e classificazione di primaria importanza per la comprensione della dinamica genetica.

Gli obiettivi specifici di questo capitolo si concentrano sulla descrizione dettagliata dei materiali e dei dati utilizzati per l'analisi, degli strumenti e delle tecnologie impiegate, nonché delle procedure metodologiche adottate per condurre lo studio. Attraverso una narrazione chiara e precisa, si mira a fornire al lettore una comprensione approfondita del processo di analisi, dalla preparazione dei dati fino alla loro elaborazione e interpretazione mediante DNABERT.

L'approccio generale adottato si articola in diverse fasi, ciascuna delle quali è volta a ottimizzare l'efficacia dell'analisi e la precisione dei risultati. In primo luogo, è stata prestata particolare attenzione alla selezione e preparazione delle sequenze di DNA da analizzare, assicurandosi che i dati fossero di alta qualità e rappresentativi delle varie tipologie di trasposoni. Successivamente, si è proceduti con

l'addestramento e il fine-tuning di DNABERT, configurando il modello per massimizzare la sua capacità di riconoscere e classificare accuratamente i trasposoni basandosi sulle caratteristiche uniche delle sequenze genetiche.

Questo capitolo si propone, dunque, di delineare un quadro metodologico solido e replicabile, capace di guidare ricerche future nel campo dell'analisi genetica mediante l'impiego di modelli di deep learning avanzati come DNABERT. Attraverso la descrizione dettagliata dell'approccio metodologico adottato, si intende non solo illustrare la procedura seguita per il raggiungimento degli obiettivi di ricerca ma anche fornire spunti per ulteriori sviluppi nel campo della bioinformatica e della genetica computazionale.

4.2 Materiale e Dati

4.2.1 Fonti dei Dati. Le sequenze di DNA analizzate in questo studio provengono principalmente dalla banca dati TREP (Transposable Element Platform) versione 1, un database pubblico affiliato all'Università di Zurigo. TREP è specializzato nella catalogazione dei trasposoni nelle piante, offrendo un'ampia gamma di sequenze genetiche annotate. Questa piattaforma è stata scelta per la sua ricchezza di dati e per la specificità del suo focus sui trasposoni, rendendola ideale per il nostro obiettivo di classificazione.

I criteri di selezione dei dati si sono concentrati sulla qualità e sulla rilevanza delle sequenze genetiche per lo studio dei trasposoni. Sono state selezionate sequenze rappresentative di diversi tipi di trasposoni, con l'intento di coprire un'ampia varietà di classi e famiglie. Questa selezione mirata ha permesso di assicurare che il dataset fosse sufficientemente variegato per testare l'efficacia di DNABERT nella classificazione dei trasposoni, mantenendo al contempo una gestione efficiente delle risorse computazionali durante l'addestramento del modello.

4.2.2 Preparazione dei Dati. La preparazione dei dati per l'analisi con DNABERT ha richiesto diversi passaggi critici, mirati a ottimizzare la compatibilità e l'efficacia del dataset per il modello. Inizialmente, è stata condotta una fase di pulizia dei dati, durante la quale le sequenze di DNA sono state esaminate per rimuovere eventuali errori o incongruenze, come sequenze incomplete o errate. Questo passaggio ha garantito che solo dati di alta qualità fossero inclusi nell'analisi.

Successivamente, è stata applicata una normalizzazione delle sequenze di DNA per assicurare che tutti i dati fossero presentati in un formato uniforme. Questo processo ha incluso l'allineamento delle sequenze e l'adattamento delle loro lunghezze per soddisfare i requisiti di input di DNABERT, facilitando un'elaborazione coerente e senza intoppi.

Il passaggio finale della preparazione dei dati ha visto la trasformazione delle sequenze di DNA in k-mer di lunghezza 4. Questa trasformazione è stata fondamentale per rendere il dataset compatibile con l'approccio di DNABERT, che utilizza metodi ispirati all'elaborazione del linguaggio naturale per analizzare le sequenze genetiche. La conversione in k-mer ha permesso di suddividere le lunghe sequenze di DNA in unità più piccole e maneggevoli, facilitando l'identificazione di pattern e caratteristiche significative da parte del modello.

In sintesi, la sezione dei materiali e dei dati illustra l'approccio meticoloso adottato nella selezione e preparazione delle sequenze di DNA per l'analisi con DNABERT. Questo processo ha non solo assicurato l'alta qualità e la rilevanza dei dati utilizzati ma ha anche ottimizzato il dataset per l'efficace applicazione di tecniche avanzate di apprendimento automatico nella classificazione dei trasposoni

4.3 Descrizione degli Strumenti e delle Tecnologie

4.3.1 DNABERT. Rappresenta un significativo e importante avanzamento nell'analisi genetica, adattando l'avanzata architettura del modello BERT, inizialmente progettato per l'elaborazione del linguaggio naturale, per l'interpretazione e l'analisi di sequenze di DNA. Quest'adattamento ha permesso di impiegare tecniche di apprendimento profondo per identificare, classificare e approfondire la comprensione dei meccanismi genetici, come la distribuzione e la funzione dei trasposoni nel genoma.

L'architettura di DNABERT conserva la struttura fondamentale di BERT, sfruttando l'architettura dei Transformer per elaborare i dati in ingresso. Questi sono particolarmente efficaci nel gestire le dipendenze a lungo raggio nei dati, una capacità essenziale nell'analisi di sequenze di DNA, che sono spesso lunghe e complesse. Una delle principali caratteristiche di DNABERT è la sua elaborazione bidirezionale dei dati, che consente al modello di apprendere il contesto di ogni k-mer nella sequenza da entrambe le direzioni, fornendo un vantaggio sostanziale nell'identificazione di pattern genetici.

Oltre alla sua architettura bidirezionale, DNABERT introduce funzionalità specifiche per l'analisi genetica. Il modello è stato addestrato e ottimizzato per riconoscere le caratteristiche uniche delle sequenze di DNA, adottando k-mer di lunghezza variabile come unità fondamentale per l'analisi. Questo approccio consente a DNABERT di catturare la diversità dei motivi genetici nel DNA, agevolando una classificazione precisa e dettagliata dei trasposoni. DNABERT ha la capacità di essere addestrato su ampi set di dati di sequenze genetiche, acquisendo la capacità di apprendere un'ampia varietà di pattern genetici e di adattarsi a diversi contesti genetici.

4.3.2 Altri Strumenti e Tecnologie. In uno sforzo volto a valutare e confrontare diversi strumenti nell'ambito della classificazione genetica, è stato considerato il progetto "Gene Fusion Classifier", sviluppato dall'Università degli Studi di Salerno. Il progetto originale è stato concepito per identificare accuratamente le fusioni geniche, un elemento cruciale per la comprensione dello sviluppo di varie patologie.

Per estendere l'ambito di applicazione del "Gene Fusion Classifier", il modello è stato testato sulla classificazione di trasposoni e retrotrasposoni, un compito che presenta sfide analitiche distinte. Il

Metrica	Valore	%
Accuratezza (Accuracy)	0.8478	84.78%
Precisione (Precision)	0.8208	82.08%
Sensibilità (Recall)	0.9305	93.05%
F1 Score	0.8722	92.47%
Specificità (Specificity)	0.9333	87.22%

Table 1: Risultati delle metriche di valutazione per il modello Gene Fusion Classifier

dataset specifico per questi elementi genetici è stato processato attraverso il modello in questione per valutare la sua efficienza in una comparazione diretta con DNABERT, un modello precedentemente validato per tale scopo.

Il "Gene Fusion Classifier" ha impiegato algoritmi di apprendimento automatico e una rete neurale artificiale per processare il dataset. Le metriche di valutazione, quali accuratezza, precisione, recall e f1 score, sono state utilizzate per quantificare la performance del modello. Questi indicatori, cruciali per determinare la validità di un modello in contesti biomedici reali, sono stati riportati in una tabella per illustrare in modo chiaro le capacità del classificatore.

La tabella precedente riporta in dettaglio le prestazioni del "Gene Fusion Classifier" quando applicato al compito di distinzione tra trasposoni e retrotrasposoni, fornendo un quadro comparativo rispetto alle prestazioni del modello DNABERT. Questa analisi comparativa sottolinea l'importanza di selezionare lo strumento più appropriato in base alla specificità del compito analitico e agli obiettivi della ricerca.

4.4 Metodologia di Analisi

4.4.1 Pre-Addestramento e Fine-Tuning. Il dataset TREP, specificamente curato per la classificazione dei trasposoni, è stato fondamentale per il processo di addestramento di DNABERT. Questo dataset è stato diviso in tre parti, due delle quali sono state impiegate per il pre-addestramento e una per il fine-tuning, consentendo un'ottimizzazione specifica del modello per il compito di classificazione dei trasposoni.

Pre-Addestramento: Il modello DNA BERT è stato addestrato utilizzando un subset del dataset per apprendere rappresentazioni generali delle sequenze di DNA. Questa fase è stata eseguita per 5 epoche con un tasso di apprendimento di 2e-4, utilizzando una dimensione del batch di 32 per GPU per l'addestramento e la valutazione. Ulteriori dettagli del pre-addestramento includono la valutazione durante l'addestramento a intervalli di 100 step, il salvataggio del modello ogni 4000 step, l'applicazione di un warmup del 10% per il tasso di apprendimento, un dropout nascosto del 10%, e un weight decay di 0.01.

Fine-Tuning: Dopo il pre-addestramento, il modello è stato fine-tunato su un set di dati separato per specializzarlo nella distinzione tra trasposoni e retrotrasposoni. Si noti che il comando fornito sopra non specifica il fine-tuning come parte di un processo di addestramento, ma piuttosto per eseguire la predizione, suggerendo che il modello era già stato fine-tunato in precedenza. Il comando per la predizione utilizza una lunghezza massima di sequenza di 75 e una dimensione del batch per la GPU di 128 per le predizioni, indicando che il modello è stato valutato su un set di test per misurare le sue prestazioni.

4.4.2 Classificazione Binaria. La classificazione binaria eseguita con DNABERT ha avuto lo scopo primario di differenziare le sequenze di DNA dei trasposoni dai retrotrasposoni. Il modello, attraverso un processo di fine-tuning mirato, è stato affinato per riconoscere i marcatori distintivi e i pattern genetici che caratterizzano ciascuna delle due categorie.

I risultati preliminari hanno indicato una performance promettente di DNABERT, con metriche che riflettono un'alta capacità del modello di classificazione accurata. L'F1 score ottenuto, insieme ad altre metriche di valutazione, suggerisce un buon equilibrio tra precisione e sensibilità. Dettagli più approfonditi sui risultati e la loro interpretazione saranno discussi nel capitolo dedicato all'analisi dei risultati, dove verrà fornito un quadro completo dell'efficacia di DNABERT nella classificazione dei trasposoni.

4.5 Valutazione del Modello

4.5.1 Metriche di Valutazione. Per esaminare pienamente l'efficacia del modello DNABERT nella distinzione tra trasposoni e retrotrasposoni, è stata impiegata una serie di metriche di valutazione statistiche ben stabilite. Queste metriche forniscono una panoramica completa dell'affidabilità e della precisione del modello nelle previsioni di classificazione.

Nella valutazione del modello DNABERT, sono state considerate le seguenti metriche:

- **Accuratezza (Accuracy):** La proporzione di predizioni corrette, sia positive che negative, sul totale delle predizioni effettuate. La formula per il calcolo è:
$$\frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$
dove TP rappresenta i veri positivi e TN i veri negativi. Il modello ha raggiunto un'accuratezza del 92.09%, riflettendo un alto livello di precisione complessiva nelle previsioni di classificazione.
- **Precisione (Precision):** Misurata come:
$$\frac{TP}{TP + FP} \tag{2}$$
dove FP indica i falsi positivi. Questo indice riflette la proporzione di identificazioni positive che si sono rivelate corrette. DNABERT ha mostrato una precisione del 93.98%, indicando un alto grado di affidabilità nelle classificazioni positive.
- **Sensibilità (Recall o Sensitivity):** Calcolata come:
$$\frac{TP}{TP + FN} \tag{3}$$
con FN che rappresenta i falsi negativi. Questa metrica quantifica la capacità del modello di identificare tutte le sequenze effettivamente positive. Il modello ha evidenziato un recall dell'91.05%, dimostrando una notevole capacità di rilevamento dei trasposoni.
- **F1 Score:** La media armonica tra precisione e recall, utile in presenza di classi sbilanciate, calcolata come:
$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$
Il modello ha conseguito un F1 score del 92.47%, sottolineando un equilibrio ottimale tra precisione e sensibilità.
- **Specificità (Specificity):** La proporzione di veri negativi correttamente identificati, calcolata con la formula:
$$\frac{TN}{TN + FP} \tag{5}$$
Il modello ha dimostrato una specificità del 93.33%, confermando la sua efficienza nel riconoscere le sequenze non trasponibili.

- **Valore Predittivo Negativo (NPV):** La probabilità che le sequenze classificate come non trasponibili siano realmente negative, calcolata come:

$$\frac{TN}{TN + FN} \tag{6}$$

  Il valore NPV del 90.05% rafforza la validità delle predizioni negative del modello.
- **Tasso di Falsi Positivi (FPR o Fall-out):** La probabilità che una sequenza non trasponibile venga erroneamente classificata come trasponibile, espressa dalla formula:

$$\frac{FP}{FP + TN} \tag{7}$$

  Un FPR del 6.67% indica un basso tasso di falsi allarmi nel modello.

Queste metriche collettivamente offrono una valutazione dettagliata della performance di DNABERT, enfatizzando la sua efficacia nella classificazione delle sequenze di DNA e la sua applicabilità come strumento diagnostico e di ricerca in campo genetico.

I risultati delle metriche di valutazione per il modello DNABERT sono riassunti nella Tabella 2.

Metrica	Valore	%
Accuratezza (Accuracy)	0.9209	92.09%
Precisione (Precision)	0.9398	93.98%
Sensibilità (Recall)	0.9105	91.05%
F1 Score	0.9247	92.47%
Specificità (Specificity)	0.9333	93.33%
NPV	0.9005	90.05%
FPR (Fall-out)	0.0667	6.67%

Table 2: Risultati delle metriche di valutazione per il modello DNABERT

4.6 Considerazioni Etiche e Limitazioni

4.6.1 Considerazioni Etiche. Nell'ambito degli studi che impiegano dati genetici, emergono significative questioni etiche legate alla natura sensibile e potenzialmente identificabile di tali informazioni. La ricerca sulla classificazione binaria di trasposoni e retrotrasposoni mediante DNABERT ha richiesto l'adozione di misure etiche rigorose. La confidenzialità dei dati genetici utilizzati è stata garantita attraverso l'anonimato, precludendo la possibilità di ricondurre i dati agli individui di provenienza. I dataset, accessibili pubblicamente, sono stati trattati per assicurare la massima protezione delle informazioni personali. È stata verificata la presenza di un consenso informato per l'utilizzo dei dati genetici a fini di ricerca, in linea con le normative etiche vigenti, assicurando che ogni utilizzo rispettasse pienamente i diritti degli individui. Il trattamento dei dati è stato limitato esclusivamente agli scopi di ricerca dichiarati, evitando ogni possibile uso improprio o divulgazione non autorizzata. L'impegno verso la trasparenza e la riproducibilità si è concretizzato nella condivisione dei codici e delle metodologie utilizzate, facilitando la verifica dei risultati da parte della comunità scientifica. È stata considerata l'eventuale rilevanza delle scoperte a lungo termine, soprattutto in relazione a possibili applicazioni nel settore medico e biotecnologico, valutando attentamente le implicazioni etiche derivanti da tali sviluppi. La ricerca è stata condotta seguendo gli standard etici più elevati, garantendo il rispetto della dignità e della privacy di tutti gli individui coinvolti. È stata data priorità all'adesione alle linee guida etiche più recenti, in risposta agli sviluppi continui nel campo della genetica e della bioinformatica.

4.6.2 Limitazioni del Metodo. Lo studio ha rivelato diverse limitazioni metodologiche nel modello DNABERT per la classificazione binaria di trasposoni e retrotrasposoni, che potrebbero influire sull'interpretazione dei risultati e sono essenziali per la comprensione dei contesti in cui i risultati possono essere generalizzati e per orientare future ricerche. La presenza di bias nei set di dati utilizzati rappresenta una delle principali preoccupazioni; nonostante gli sforzi per utilizzare dati rappresentativi, la disponibilità limitata di sequenze genetiche completamente annotate può portare a un bias di campionamento, influenzando la capacità del modello di generalizzare oltre i dati su cui è stato addestrato. Anche se DNABERT rappresenta un avanzamento significativo nell'analisi delle sequenze di DNA, emergono limitazioni tecniche intrinseche al modello, come la necessità di risorse computazionali significative e sfide legate all'interpretazione dei risultati. La complessità biologica dei trasposoni e dei retrotrasposoni può presentare sfide, in quanto la classificazione binaria potrebbe non catturare completamente la diversità funzionale e la variabilità tra differenti tipi e famiglie di elementi trasponibili. L'interpretazione dei risultati prodotti da DNABERT può essere complessa a causa della natura "scatola nera" dei modelli di deep learning, che pone sfide nella spiegabilità e nella trasparenza del processo decisionale. Il campo della genetica è in rapida evoluzione, richiedendo aggiornamenti continui del modello per rimanere allineato con l'attuale stato della ricerca. Le metriche utilizzate per valutare il modello sono standardizzate, ma la loro interpretazione può variare a seconda del contesto biologico, richiedendo un'attenta considerazione. Queste limitazioni enfatizzano l'importanza di procedere con cautela nell'interpretazione dei risultati e forniscono direzioni per future indagini, sia nel miglioramento dei modelli di machine learning in genetica che nella loro applicazione a problemi biologici complessi

4.7 Sommario della Metodologia

Incorporando le informazioni fornite con l'approccio innovativo utilizzato nel nostro studio, il capitolo presenta una sintesi esaustiva della metodologia adottata per la classificazione binaria di trasposoni e retrotrasposoni. Questo include l'utilizzo di DNABERT, complementato dalla nostra iniziativa di applicare il "Gene Fusion Classifier" per lo stesso scopo. La descrizione dettagliata copre la selezione e la preparazione dei dati, il processo di addestramento e di ottimizzazione del modello, oltre alle strategie di valutazione impiegate, che comprendono l'uso di metriche standard per garantire la precisione dei risultati ottenuti.

L'approccio metodologico scelto ha consentito di raggiungere gli obiettivi di ricerca, evidenziando come sia DNABERT che il "Gene Fusion Classifier", pur essendo stato originariamente sviluppato per identificare fusioni geniche, possano essere efficacemente utilizzati per analizzare e classificare sequenze genetiche relative a trasposoni e retrotrasposoni. La validazione incrociata e l'utilizzo

di un set di test separato hanno ulteriormente confermato l'efficacia e l'affidabilità di entrambi i modelli.

Nonostante le sfide metodologiche incontrate, l'implementazione di questi metodi ha portato a risultati promettenti, che arricchiscono il corpus di conoscenze esistente e aprono nuove strade per future ricerche. Questi risultati sostengono l'avanzamento scientifico nel campo e sottolineano l'importante ruolo dei modelli di apprendimento automatico in genetica. In definitiva, il capitolo dimostra come una metodologia ben definita e scrupolosamente applicata sia cruciale per acquisire nuove conoscenze affidabili e per promuovere ulteriori indagini nel settore della bioinformatica e della biologia computazionale.

5 ANALISI DEI RISULTATI

5.1 Confronto tra i modelli

Nel contesto dell'analisi genetica, è stata condotta una comparazione tra due modelli avanzati: DNABERT e Gene Fusion Classifier. L'obiettivo era determinare quale dei due modelli offrisse prestazioni migliori nella classificazione binaria di trasposoni e retrotrasposoni. Entrambi i modelli sono stati sottoposti a un'analisi approfondita, valutati su un insieme comune di metriche per garantire un confronto equo e oggettivo.

DNABERT, con le sue origini nell'elaborazione del linguaggio naturale, ha mostrato un'accuratezza del 92.09%, una precisione del 93.98%, un recall del 91.05% e un F1 score del 92.47%. La specificità registrata è stata del 93.33%, con un valore predittivo negativo (NPV) del 90.05% e un tasso di falsi positivi (FPR) del 6.67%. Queste metriche indicano che DNABERT ha una forte capacità di classificazione, sostenuta da un equilibrio notevole tra la capacità di rilevamento e la precisione.

D'altro canto, il Gene Fusion Classifier, progetto specificatamente focalizzato sulla classificazione delle fusioni geniche, ha ottenuto un'accuratezza dell'84.78%, una precisione dell'82.08%, un recall del 93.05% e un F1 score del 87.22%. La specificità misurata per questo modello è stata del 87.22%, dimostrando anche una buona capacità di classificazione, sebbene leggermente inferiore in termini di precisione e accuratezza rispetto a DNABERT.

Il confronto tra i due modelli rivela che, sebbene entrambi dimostrino prestazioni lodevoli, DNABERT supera il Gene Fusion Classifier in termini di accuratezza e precisione. Questi risultati non solo forniscono una chiara indicazione delle capacità di ciascun modello ma offrono anche spunti preziosi per future ricerche, indicando le aree in cui il Gene Fusion Classifier potrebbe essere ulteriormente ottimizzato.

Tali valutazioni comparative sono cruciali per guidare le scelte degli strumenti analitici in studi genetici futuri, dove la selezione di un modello rispetto a un altro può avere implicazioni significative per la scoperta e la comprensione di elementi genetici critici come trasposoni e retrotrasposoni.

5.2 Valutazione della Performance

La valutazione delle prestazioni ha preso in esame vari aspetti dei modelli DNABERT e Gene Fusion Classifier. Riguardo alla velocità di elaborazione, DNABERT ha mostrato una notevole efficienza, gestendo con destrezza dataset di ampie dimensioni, fattore critico nella genomica moderna. In termini di facilità di implementazione, entrambi i modelli beneficiano di architetture e librerie software ben supportate, consentendo un'integrazione relativamente semplice in flussi di lavoro esistenti.

5.3 Analisi dei Punti di Forza

L'analisi dei punti di forza ha rivelato che DNABERT eccelle nella comprensione del contesto bidirezionale, un vantaggio chiave nella classificazione genetica, mentre il Gene Fusion Classifier si distingue per la sua capacità di focalizzarsi specificatamente sulle fusioni geniche. Queste caratteristiche li rendono adatti a compiti specializzati di classificazione, con DNABERT che mostra versatilità attraverso vari contesti genetici e il Gene Fusion Classifier che offre precisione nelle sue aree di focalizzazione.

5.4 Identificazione delle Debolezze

In termini di debolezze, DNABERT può richiedere una quantità significativa di risorse computazionali per l'addestramento e può mostrare una certa sensibilità al rumore nei dati, richiedendo dataset di alta qualità. Il Gene Fusion Classifier, pur essendo meno esigente in termini di potenza computazionale, può risentire di limitazioni nell'identificazione di varianti genetiche meno comuni a causa di un possibile bias nel training set.

5.5 Implicazioni Pratiche

Entrambi i modelli offrono implicazioni pratiche considerevoli. DNABERT, per la sua versatilità e profondità di apprendimento, si presenta come uno strumento potenziale per la diagnosi clinica e la ricerca di base, in grado di identificare varianti genetiche complesse. Il Gene Fusion Classifier, con la sua specificità, può rivelarsi prezioso per identificare con precisione le fusioni geniche patologiche, trovando applicazione in ambiti diagnostici mirati e in studi di oncogenomica.

6 CONCLUSIONI

In conclusione, l'importanza di un approccio metodologico rigoroso e dettagliatamente implementato per il raggiungimento di nuove conoscenze affidabili nel campo della bioinformatica e della biologia computazionale viene enfatizzata. Attraverso l'esplorazione della classificazione binaria di trasposoni e retrotrasposoni, sono stati implementati sia DNABERT che, in un contesto diverso, il "Gene Fusion Classifier", originariamente sviluppato per l'identificazione di fusioni geniche. Questa ricerca ha permesso non solo di valutare l'efficacia di questi modelli in applicazioni specifiche, ma anche di confrontare i risultati ottenuti, fornendo un quadro comparativo del loro potenziale analitico.

Il confronto tra i risultati ottenuti con DNABERT e quelli derivanti dall'uso innovativo del "Gene Fusion Classifier" apre prospettive interessanti su come differenti strumenti di machine learning possono essere sfruttati e, potenzialmente, integrati per affrontare le sfide poste dalla genetica moderna. Sebbene lo studio non abbia combinato direttamente i due modelli, l'analisi comparativa sottolinea l'importanza di continuare a esplorare sinergie tra diverse tecniche di apprendimento automatico per migliorare l'accuratezza e l'efficacia della classificazione genetica.

I risultati promettenti ottenuti, nonostante le limitazioni sulle metodologie incontrate, contribuiscono significativamente alla letteratura esistente e pongono le basi per ulteriori ricerche. L'impiego di queste tecnologie avanzate supporta l'avanzamento della comprensione scientifica in questo ambito e rafforza il concetto che l'innovazione tecnologica rappresenta una risorsa cruciale per affrontare le complesse questioni biologiche odierne.

In ultima analisi, lo studio evidenzia il potenziale insito nel confronto e nell'eventuale integrazione di diversi modelli di apprendimento automatico nel settore della genetica, sottolineando come tali approcci possano portare a scoperte significative e a un avanzamento concreto della scienza.

6.1 Prospettive per il Miglioramento

Il perfezionamento dei modelli DNABERT e Gene Fusion Classifier può procedere lungo diverse direttrici. L'integrazione di dataset più vasti e diversificati potrebbe aumentare la robustezza dei modelli di fronte a varianti genetiche rare o atipiche. Un affinamento degli algoritmi di apprendimento potrebbe ridurre la sensibilità al rumore e migliorare la gestione di dati non bilanciati. Inoltre, la sperimentazione con diverse configurazioni di architettura di rete e parametri di addestramento potrebbe ottimizzare ulteriormente le prestazioni. La collaborazione con esperti di dominio potrebbe anche portare a sviluppi innovativi, ad esempio attraverso l'inserimento di conoscenze biologiche a priori nel processo di apprendimento.

6.2 Proposte per Ricerche Future

I risultati comparativi offrono diversi spunti per ricerche future. È essenziale esplorare lo sviluppo di modelli che combinano le forze di DNABERT e Gene Fusion Classifier, potenzialmente creando un sistema ibrido che massimizza l'accuratezza e la specificità. La ricerca potrebbe anche indirizzarsi verso l'adattamento dei modelli a nuovi compiti, come la predizione delle conseguenze funzionali delle fusioni geniche o l'analisi di interazioni genomiche complesse. Infine, studi ulteriori potrebbero indagare l'applicazione di questi modelli in contesti clinici reali, valutando la loro efficacia nella diagnosi precoce o nel monitoraggio della progressione di malattie geneticamente correlate.

7 DATA AVAILABILITY

Il dataset utilizzato è reperibile sul sito curato dall'Univerisità di Zurigo. Reperibile su TREP, the TRansposable Elements Platform. La repository di DNABERT è reperibile su GitHub.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. (2023). arXiv: 1706.03762 [cs.CL].

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: pre-training of deep bidirectional transformers for language understanding. (2019). arXiv: 1810.04805 [cs.CL].

[3] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37, 15, (February 2021), 2112–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab083. eprint: https://academic.oup.com/bioinformatics/article-pdf/37/15/2112/56752258/btab083.pdf. https://doi.org/10.1093/bioinformatics/btab083.

[4] Md Rezaul Karim, Oya Beyan, Achille Zappa, Ivan G Costa, Dietrich Rebholz-Schuhmann, Michael Cochez, and Stefan Decker. 2020. Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, 22, 1, (February 2020), 393–415. ISSN: 1477-4054. DOI: 10.1093/bib/bbz170. eprint: https://academic.oup.com/bib/article-pdf/22/1/393/35934885/bbz170.pdf. https://doi.org/10.1093/bib/bbz170.

[5] Benjamin T James, Brian B Luczak, and Hani Z Girgis. 2018. MeShClust: an intelligent tool for clustering DNA sequences. *Nucleic Acids Research*, 46, 14, (May 2018), e83–e83. ISSN: 0305-1048. DOI: 10.1093/nar/gky315. eprint: https://academic.oup.com/nar/article-pdf/46/14/e83/25509702/gky315.pdf. https://doi.org/10.1093/nar/gky315.

[6] Alison B Hickman and Fred Dyda. 2016. Dna transposition at work. *Chemical reviews*, 116, 20, 12758–12784.

[7] Cédric Feschotte. 2023. Transposable elements: mcclintock's legacy revisited. *Nature Reviews Genetics*, 24, 11, 797–800.

[8] Imon Banerjee, Yuan Ling, Matthew C Chen, Sadid A Hasan, Curtis P Langlotz, Nathaniel Moradzadeh, Brian Chapman, Timothy Amrhein, David Mong, Daniel L Rubin, et al. 2019. Comparative effectiveness of convolutional neural network (cnn) and recurrent neural network (rnn) architectures for radiology text report classification. *Artificial intelligence in medicine*, 97, 79–88.

[9] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118, 15, e2016239118. DOI: 10.1073/pnas.2016239118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2016239118. https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.