Utilizzo della BWT per la generazione di features per migliorare algoritmi di machine learning per la classificazione di sequenze genomiche che presentano gene fusion

Vincenzo Tarantino, Bartolomeo Mazzeo

Corso di Strumenti formali per la Bioinformatica 2024/2025

Contents

1 Introduzione - Descrizione del problema

La Burrows-Wheeler Transform (BWT) è un algoritmo ampiamente utilizzato in diversi ambiti, in particolare nella compressione dei dati e nella bioinformatica nel contesto dell'analisi e l'elaborazione efficiente di sequenze genomiche. Una delle sue applicazioni più rilevanti è l'estrazione di caratteristiche (features) significative da sequenze di DNA o RNA, facilitando la classificazione e il riconoscimento di particolari strutture biologiche.

Nell'articolo [1], la BWT viene utilizzata per estrarre i b-mers, ovvero segmenti di sequenza che condividono determinate proprietà strutturali e che vengono identificati sfruttando la trasformazione BWT della sequenza originale. I b-mers rappresentano caratteristiche distintive delle sequenze genomiche e possono essere utilizzati come input per modelli di apprendimento automatico al fine di individuare schemi e correlazioni nei dati utili per la classificazione.

Nel nostro progetto, vogliamo applicare metodologie simili per verificare se le features estratte tramite la BWT possono essere utili per individuare sequenze che presentano **gene fusion**, ovvero eventi di fusione genica in cui segmenti di due geni distinti si uniscono per formare un gene ibrido.

Questi eventi sono particolarmente rilevanti in ambito oncologico, in quanto possono contribuire alla progressione di diversi tipi di tumori.

L'obiettivo è quindi verificare se, attraverso l'estrazione di b-mers ottenutidalla BWT delle sequenze genomiche e l'addestramento di modelli di machine learning, sia possibile distinguere efficacemente tra sequenze con e senza gene fusion.

2 Burrows-Wheeler Transform e B-mers

2.1 Burrows-Wheeler Transform (BWT)

La *Burrows-Wheeler Transform* (BWT) è un algoritmo di trasformazione delle stringhe introdotto da Burrows e Wheeler nel 1994. Il suo principale utilizzo è nella compressione dei dati e nell'analisi di sequenze testuali; trova applicazione in bioinformatica nell'allineamento di sequenze genomiche e nella classificazione di sequenze biologiche.

Il vantaggio della BWT è che l'algoritmo trasforma una stringa S in una nuova rappresentazione che raggruppa i caratteri simili, rendendo più efficiente la compressione e la ricerca di pattern. L'algoritmo svolge principalmente i seguenti passaggi:

1. Generazione di tutte le possibili rotazioni cicliche della stringa S.
2. Le rotazioni vengono ordinate tenendo conto dell'ordinamento lessicografico.
3. Estrazione dell'ultima colonna della matrice ordinata, che rappresenta la trasformata BWT di S.

2.2 Estrazione dei B-mers

L'estrazione dei *b-mers* è un processo che sfrutta la proprietà della BWT di raggruppare caratteri simili. A differenza dei k-mers, che sono tutte le possibili sottosequenze di lunghezza k di una stringa, i b-mers sono sequenze di lunghezza variabile estratte direttamente dalla rappresentazione BWT.

Il procedimento di estrazione avviene seguendo questi passaggi:

1. Si individuano i caratteri ripetuti consecutivamente nella BWT.
2. Per ciascuna ripetizione, si seleziona l'intervallo corrispondente nelle rotazioni ordinate.
3. Si identifica il prefisso comune a tali rotazioni e si costruisce il b-mer concatenando il carattere ripetuto con il prefisso.

Per avere la certezza di considerare solo features importanti, solitamente si selezionano solo b-mers che appaiono almeno due volte in una sequenza.

3 Gene Fusion: Definizione, Cause e Importanza nella Diagnosi Genomica

Il genoma è l'insieme completo del DNA di un organismo, che contiene tutte le informazioni genetiche necessarie per il suo sviluppo e funzionamento. Esso è composto da geni, ovvero segmenti di DNA che codificano per proteine. In alcuni casi, può verificarsi una **fusione genica**, un fenomeno in cui due geni distinti si uniscono, portando alla produzione di proteine anomale. Queste proteine possono alterare i processi cellulari e contribuire allo sviluppo di malattie, come il cancro.

Le fusioni geniche avvengono quando (COMBINATORICS ON WORDS AND MACHINE LEARNING FOR COMPUTATIONAL BIOLOGY: THE GENE FUSION - Eduardo Autore - 2024) il DNA si riorganizza in modo anomalo, portando alla combinazione di due geni distinti. Questo può accadere principalmente a causa di:

- **Riarrangiamenti cromosomici**: cambiamenti nella struttura dei cromosomi, tra cui:
  - *Traslocazioni*: spostamento di segmenti tra cromosomi diversi;
  - *Delezioni*: perdita di parti del DNA;

  - *Inversioni*: segmenti di DNA reinseriti in orientamento invertito;
  - *Duplicazioni*: copie multiple di un gene che possono favorire la fusione.

- **Errori durante la divisione cellulare**: anomalie nella separazione dei cromosomi durante la mitosi o la meiosi possono generare fusioni geniche.
- **Errori nella riparazione del DNA**: quando il meccanismo di correzione del DNA unisce segmenti da cromosomi diversi, può formarsi un gene fuso.

Questi eventi possono alterare l'espressione genica e contribuire all'insorgenza di malattie.

4 Svolgimento progetto

4.1 Dati a disposizione

Per lo svolgimento del progetto sono stati utilizzati due dataset contenenti sequenze genomiche di DNA:

- **all_transcripts.fasta**: contiene sequenze genomiche **senza gene fusion**, con lunghezza compresa tra 52.000 e 250.000 nucleotidi. Il file comprende 400 sequenze.
- **fusim.bench.fasta**: contiene sequenze **con gene fusion**, con lunghezza compresa tra circa 2.000 e 5.000 nucleotidi. Il file comprende 2561 sequenze.

Entrambi i file sono in formato FASTA, un formato standard per la rappresentazione di sequenze genomiche che associa un id ad ogni sequenza. In particolare, è possibile distinguere facilmente tra sequenze con e senza gene fusion basandosi sul loro id: le sequenze contenenti fusioni geniche presentano un identificatore che inizia con `ref|...`, mentre le sequenze senza fusioni geniche hanno un identificatore che inizia con `ENST...`.

Poiché la lunghezza delle sequenze senza gene fusion è notevolmente superiore a quella delle sequenze con gene fusion, si è reso necessario un preprocessing per rendere i dati confrontabili.

Un terzo file, **genes_panel.txt**, elenca alcuni geni di interesse, ma non è stato utilizzato direttamente nell'analisi.

4.2 Preprocessing dei dati e applicazione BWT

La grande differenza di lunghezza tra le sequenze genomiche dei due dataset rendeva difficile un confronto diretto. Per ovviare a questo problema, è stato applicato un metodo di **segmentazione** alle sequenze di `all_transcripts.fasta`. Innanzitutto, è stata calcolata la lunghezza media delle sequenze presenti nel file `fusimbench`, risultata pari a circa 2500 nucleotidi. Per gestire la differenza di lunghezza tra le sequenze dei due dataset, le stringhe di lunghezza maggiore sono state suddivise in blocchi di circa 2500 nucleotidi.

Oltre alla disparità nella lunghezza, i due dataset presentavano anche una differenza nel numero di sequenze disponibili. Per bilanciare il numero di campioni tra le due classi, sono stati selezionati più blocchi per ciascuna sequenza senza gene fusion, in modo da ottenere un numero di elementi comparabile tra i due dataset.

Queste operazioni hanno permesso di ottenere due insiemi di dati equilibrati in termini di lunghezza e quantità dei dati. I due dataset sono stati uniti in un unico file CSV, aggiungendo una colonna `label` per identificare la tipologia di sequenza. In questo modo, la distinzione tra sequenze con e senza gene fusion risulta più immediata, senza dover fare affidamento sulla struttura dell'ID. A questo punto è stato possibile applicare la **Burrows-Wheeler Transform** per poi procedere con l'estrazione dei b-mers.

4.3 Applicazione ECCD per la riduzione del numero di features

Successivamente, il dataset è stato riorganizzato per ottenere i vettori numerici: per ogni sequenza, vengono elencati tutti i b-mers estratti dal dataset insieme alla loro frequenza di comparsa all'interno della sequenza stessa.

Infine, è stato applicato l'algoritmo ECCD per selezionare solo i b-mers più significativi, ovvero quelli che influenzano maggiormente il livello di incertezza nel dataset, contribuendo così a migliorare la qualità del dataset con l'obiettivo di fornire dati migliori ai modelli di machine learning.

5 Addestramento degli algoritmi di Machine Learning

Una volta ottenuto il dataset finale con i b-mers selezionati, sono stati addestrati diversi modelli di Machine Learning per verificare se le features estratte tramite la BWT fossero efficaci nella classificazione delle sequenze con e senza gene fusion. I dati sono stati divisi in questo modo:

- **Training set**: pari all'80% del dataset totale, utilizzato per l'apprendimento dei modelli.
- **Test set**: pari al restante 20%, utilizzato per valutare le prestazioni dei modelli su dati mai visti prima.

Per garantire che entrambe le classi (*sequenze con gene fusion* e *sequenze senza gene fusion*) siano rappresentate equamente nei due insiemi, la suddivisione è stata effettuata con stratificazione delle classi. Questo significa che la proporzione tra sequenze con e senza gene fusion è mantenuta ugualmente sia nel training set che nel test set.

Sono stati testati quattro algoritmi principali: **Support Vector Machine (SVM)**, **Naïve Bayes**, **Random Forest** e **Reti Neurali Artificiali (ANN)**. In questa sezione vengono descritti brevemente questi algoritmi e vengono riportati i risultati ottenuti nel nostro caso.

5.1 Support Vector Machine (SVM)

Le Support Vector Machine (SVM) sono algoritmi di Machine Learning supervisionati utilizzati per problemi di classificazione e regressione. L'obiettivo principale di una SVM è trovare un **iperpiano ottimale** che separi al meglio le classi nello spazio delle features. Se i dati non sono linearmente separabili, la SVM può applicare delle trasformazioni tramite funzioni chiamate *kernel*, che proiettano i dati in uno spazio a dimensioni superiori, rendendo la separazione più semplice.

Nel nostro caso, è stato utilizzato un kernel **lineare**, in quanto si è osservato che i b-mers generano dati che possono essere separati con un semplice iperpiano. L'addestramento è stato effettuato utilizzando il dataset ridotto dopo l'applicazione dell'algoritmo ECCD.

Risultati: Il modello SVM ha ottenuto un'accuratezza del 91%, dimostrando di essere molto efficace nella classificazione delle sequenze genomiche. La matrice di confusione ha mostrato un buon bilanciamento tra precision e recall, con una chiara distinzione tra sequenze con e senza gene fusion. Notiamo che le sequenze con gene fusion classificate erroneamente sono 28, lo stesso vale per quelle senza gene fusion.

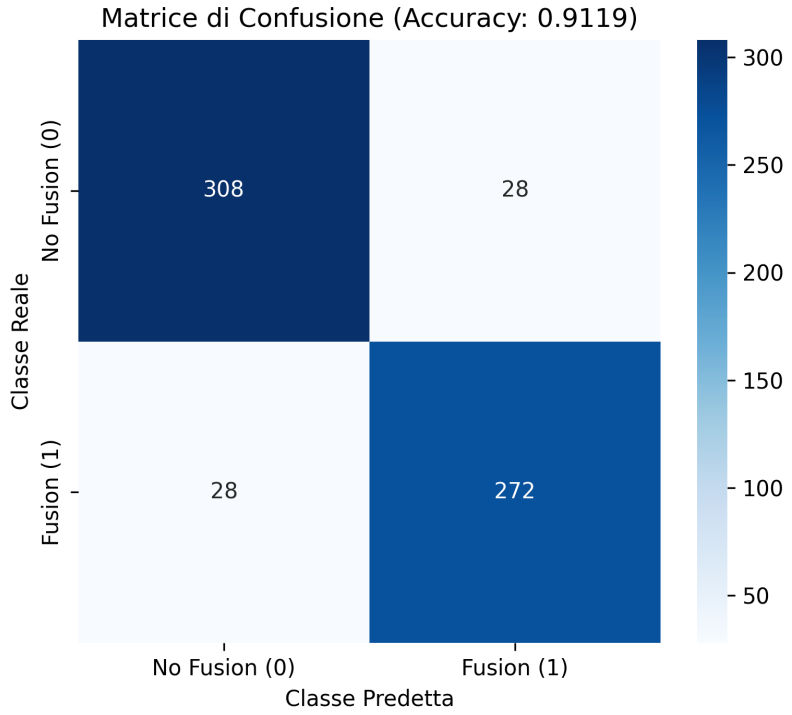

Figure 1: Confusion matrix SVM

5.2 Naïve Bayes

L'algoritmo **Naïve Bayes** è un classificatore probabilistico basato sul *Teorema di Bayes*, che assume che tutte le feature siano indipendenti tra loro (assunzione di indipendenza condizionale). Nonostante questa assunzione sia spesso irrealistica in molti contesti, il modello funziona molto bene in problemi di classificazione testuale e bioinformatica, dove i pattern possono essere approssimati con questa ipotesi.

Nel nostro caso, Naïve Bayes è stato addestrato per stimare la probabilità che una sequenza appartenga alla classe **gene fusion** o **non gene fusion**, utilizzando la frequenza dei b-mers come feature.

Risultati: L'accuratezza ottenuta con Naïve Bayes è stata inferiore rispetto alla SVM, attestandosi intorno all'81%. Questo risultato è in linea con le aspettative, dato che il modello fa forti assunzioni di indipendenza che potrebbero non essere completamente valide nel nostro dataset. Visualizzando la matrice di confusione, notiamo un aumento delle sequenze classificate erroneamente per entrambe le classi.

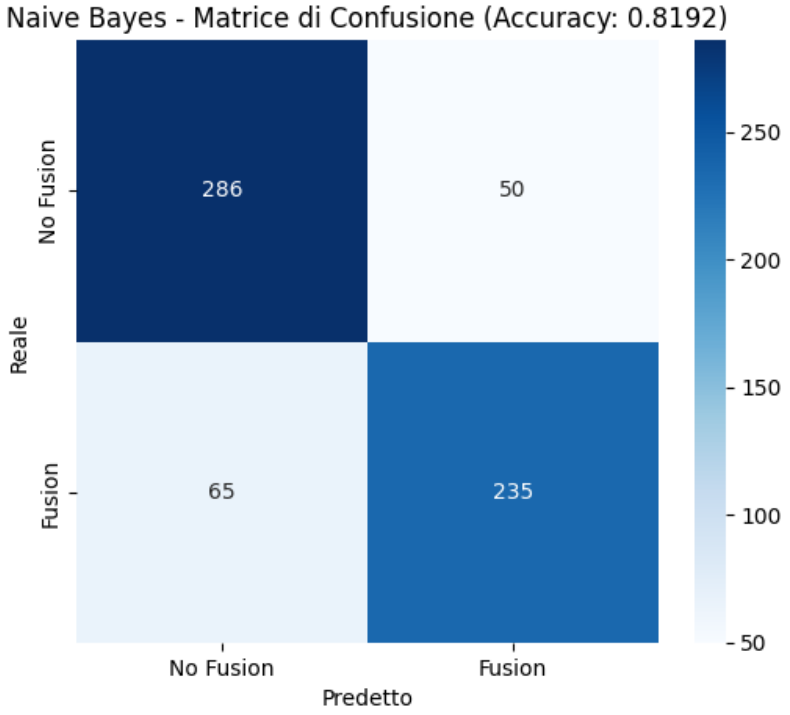

Figure 2: Confusion matrix Naive Bayes

5.3 Random Forest

Il modello **Random Forest** è un classificatore basato su insiemi di alberi decisionali (*decision trees*). Funziona generando molteplici alberi decisionali su sottoinsiemi casuali dei dati e aggregando le loro predizioni per ottenere un risultato più robusto e meno incline all'overfitting. Ogni albero è costruito basandosi su un insieme casuale di feature, rendendo il modello meno sensibile a variazioni nei dati di training.

Nel nostro caso, la Random Forest è stata applicata per classificare le sequenze genetiche sulla base della presenza e frequenza dei b-mers, combinando le predizioni di più alberi per ottenere una decisione finale.

Risultati: L'accuratezza ottenuta è stata superiore a quella della SVM, attestandosi attorno al 95%. La Random Forest ha mostrato una buona capacità di classificazione, mostrando particolare efficacia nella classificazione corretta di sequenze con gene fusion.

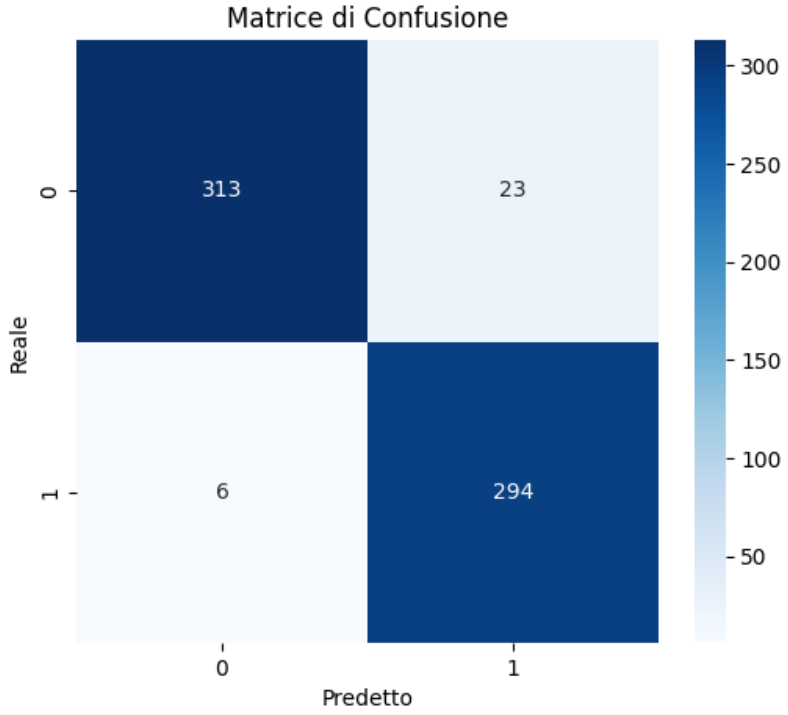

Figure 3: Confusion matrix Random Forest

5.4 Reti Neurali Artificiali (ANN)

Le **Reti Neurali Artificiali** (*Artificial Neural Networks*) sono modelli di apprendimento ispirati alla struttura del cervello umano. Sono composti da strati di neuroni interconnessi, suddivisi in:

- **Strato di input:** riceve i dati in ingresso (nel nostro caso, la frequenza dei b-mers).
- **Strati nascosti:** elaborano le informazioni applicando funzioni di attivazione non lineari per catturare pattern complessi nei dati.
- **Strato di output:** restituisce la previsione finale (classe 0 o 1).

Nel nostro caso, è stata utilizzata una rete neurale con **un singolo strato nascosto** e funzione di attivazione *ReLU*, seguita da un livello di output con attivazione *softmax* per la classificazione binaria.

Risultati: L'accuratezza ottenuta con la rete neurale è risultata simile a quella della Random Forest, raggiungendo il 94%. Possiamo notare infatti che il numero di elementi classificati erroneamente è simile a quello della Random Forest.

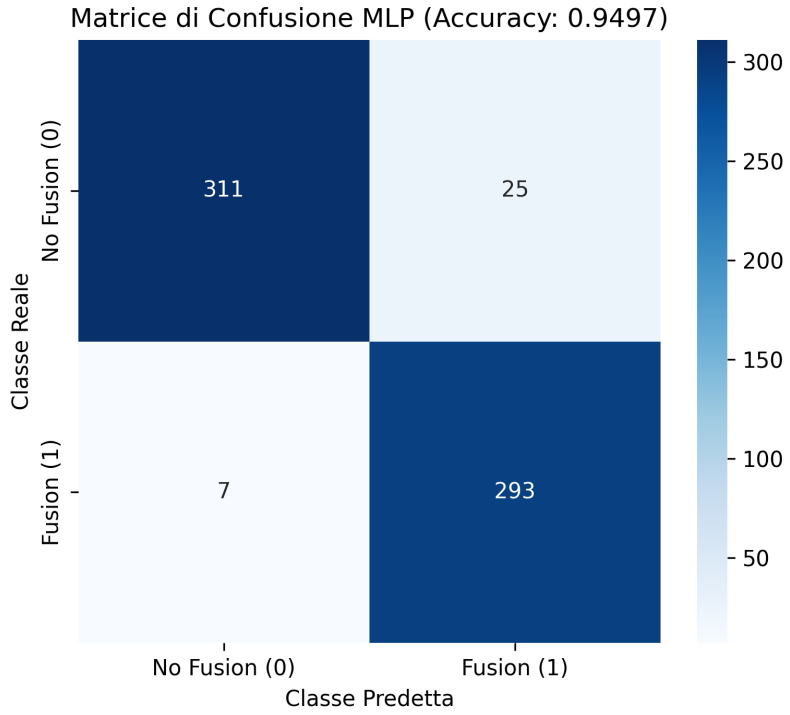

Figure 4: Confusion matrix Reti Neurali

5.5 Confronto tra i modelli

I risultati ottenuti mostrano che la **Random Forest** ha fornito le migliori prestazioni in termini di accuratezza e bilanciamento tra le classi, con un'accuracy del **95.44%**.

Le **reti neurali** hanno ottenuto prestazioni simili alla Random Forest, con un'accuracy del **94.97%**. La **Support Vector Machine (SVM)** ha raggiunto un'accuracy del **91.19%**, mentre il **Naive Bayes** ha registrato l'accuracy più bassa (**81.92%**), indicando che questo modello potrebbe non essere il più adatto per il problema analizzato.

Di seguito, una tabella riepilogativa delle prestazioni dei modelli di Machine Learning valutati:

Modello	Accuracy	Precision	Recall	F1-score
SVM	0.9119	0.91	0.91	0.91
Naive Bayes	0.8192	0.82	0.78	0.80
Random Forest	0.9544	0.96	0.95	0.95
Rete Neurale	0.9497	0.95	0.95	0.95

Table 1: Confronto delle prestazioni dei modelli di Machine Learning

6 Conclusioni

In questo lavoro abbiamo esplorato l'uso della *Burrows-Wheeler Transform* (BWT) per l'estrazione di features da sequenze genomiche al fine di distinguere tra sequenze con e senza *gene fusion*. Seguendo un approccio simile a quello presentato in letteratura, abbiamo estratto i *b-mers* come caratteristiche distintive e applicato tecniche di *feature selection* basate sull'entropia (*ECCD*) per ridurre la dimensionalità del dataset.

Successivamente, sono stati addestrati diversi algoritmi di *machine learning*, tra cui Support Vector Machine (SVM), Naive Bayes, Random Forest e Reti neurali, per valutare l'efficacia delle features estratte nella classificazione delle sequenze genomiche. I risultati hanno mostrato che i modelli di

Random Forest e reti neurali hanno ottenuto le migliori prestazioni in termini di accuratezza e bilanciamento delle classi, confermando l'utilità dell'approccio basato sulla BWT per questo tipo di problema. Tuttavia, potrebbero essere esplorate ulteriori ottimizzazioni e approcci, come l'uso di altri metodi di selezione delle features o l'integrazione di informazioni riguardanti la struttura delle sequenze genomiche per effettuare classificazioni più precise in base a quali geni si fondono nelle sequenze con gene fusion.

References

[1] Karthik Tangirala and Doina Caragea. Generating features using burrows wheeler transformation for biological sequence classification. In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms - Volume 1: BIOINFORMATICS, (BIOSTEC 2014)*, pages 196–203. INSTICC, SciTePress, 2014.