GENEFUSION BWT

FEATURE EXTRACTION E CLASSIFICAZIONE SU DATI GENOMICI REALI

Un approccio **alignment-free** basato su **Machine Learning**

Carmine Calabrese 0522501853
Daniele Dello Russo 0522501766

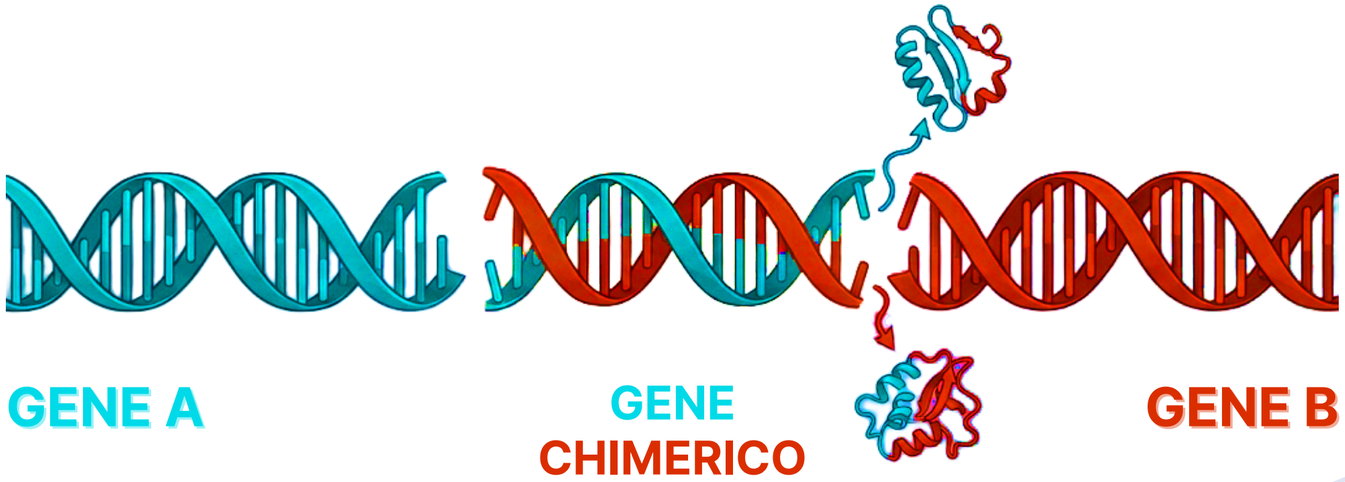
Prof. **Rocco Zaccagnino**

FLaTN²BIO Lab

COSA SONO LE FUSIONI GENICHE?

Geni **ibridi** generati dall'unione di due o più geni originari [1]

Driver molecolari della *carcinogenesi*



[1] MJ Annala et al. "Fusion genes and their discovery using high throughput sequencing." In: Cancer letters 340.2 (2013)

RILEVANZA CLINICA

PRECISIONE DIAGNOSTICA

Marker per confermare sottotipi di cancro

PROGNOSI

Forniscono indicazioni sull'**aggressività** della malattia.

TARGET TERAPEUTICO

Monitoraggio della **risposta** alle cure e delle resistenze.

[1] MJ Annala et al. “Fusion genes and their discovery using high throughput sequencing.” In: Cancer letters 340.2 (2013)

OBIETTIVO DELLO STUDIO

Creare un classificatore robusto usando la BWT

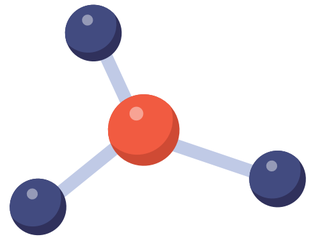
Sequenza DNA grezza ⟶ Trasformazione ⟶ Classificazione

Superare i limiti dell’analisi standard

Utilizzare la **Burrows-Wheeler Transform (BWT)** per far emergere pattern nascosti

Addestrare modelli di **ML** per distinguere *Fusion vs Non-Fusion*

LA BWT



mississippi$	⟹	$ mississipp i
ississippi$m		i $mississip p
ssissippi$mi		i ppi$missis s
sissippi$mis		i ssippi$mis s
issippi$miss		i ssissippi$ m
ssippi$missi		m ississippi $
sippi$missis		p i$mississi p
ippi$mississ		p pi$mississ i
ppi$mississi		s ippi$missi s
pi$mississip		s issippi$mi s
i$mississipp		s sippi$miss i
$mississippi		s sissippi$m i

La BWT è una **permutazione reversibile** dei caratteri di una sequenza **S** [2]

↓

Calcolo della matrice delle **rotazioni cicliche** di *S$*

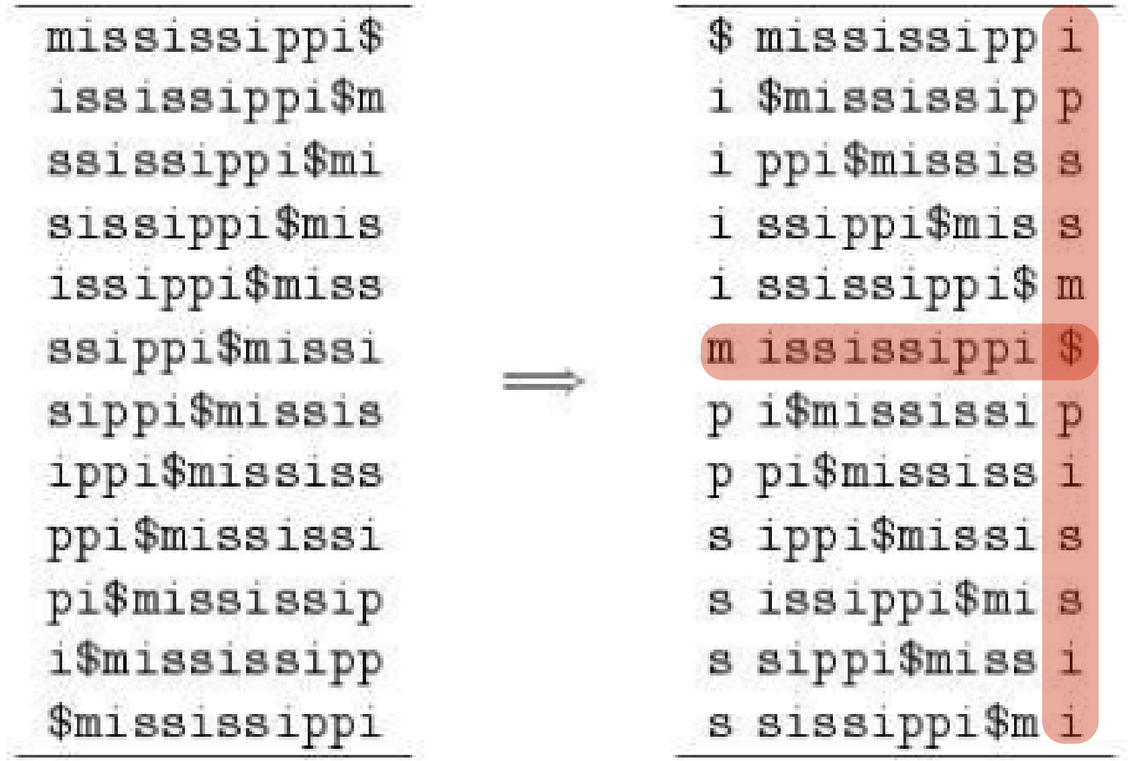
↓

Ordinamento lessicografico della matrice

[2] Elad Verbin Shir Landau. “The Burrows-Wheeler compression algorithm is even better than what you have thought.” In: (Apr. 2005).

LA BWT

Il risultato della BWT sarà **l'ultima colonna (L)** della matrice ordinata e l'**indice (I)** della riga contenente la stringa che termina per il terminatore $



VARIANTI BWT

STANDARD BWT

Applicazione dell'algoritmo sulla **sequenza originale**

Catturare le **regolarità** e i **pattern ripetuti** della sequenza principale

REVERSE BWT

Applicazione dell'algoritmo alla **sequenza invertita**

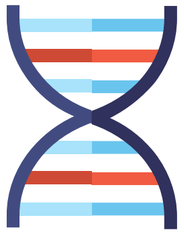
Gestire la **natura a doppio filamento del DNA**, intercettando pattern di fusione indipendenti dall'orientamento di lettura

POSITIONAL BWT

Variante che conserva informazioni sulla **posizione originale** dei nucleotidi

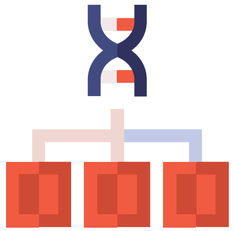
Utilizzata come baseline di confronto per valutare se la l**ocalizzazione spaziale** offrisse vantaggi rispetto all'approccio statistico globale

PIPELINE E DATASET



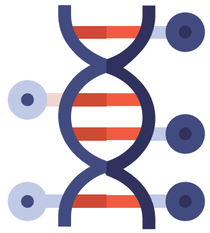
RAW DATA

Dataset reale contenente sequenze *Fusion* e *Non-Fusion*



PREPROCESSING

Rimozione sequenze duplicate



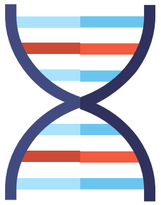
WINDOWING

Riduzione lunghezza delle sequenze

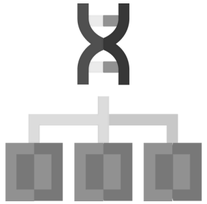
FEATURE ENGINEERING

Estrazione delle features dalle sequenze processate

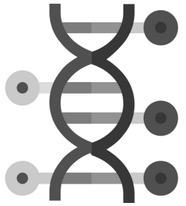
PIPELINE E DATASET



RAW DATA

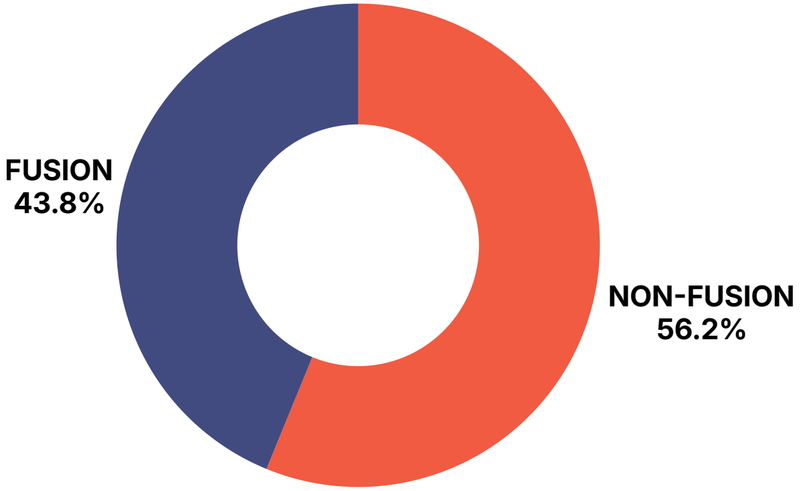


PREPROCESSING



WINDOWING

FEATURE ENGINEERING

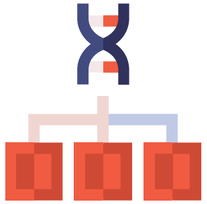


Struttura Dataset

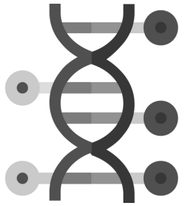
- sequence
- gene1
- gene2
- junction_point
- label

PIPELINE E DATASET

RAW DATA

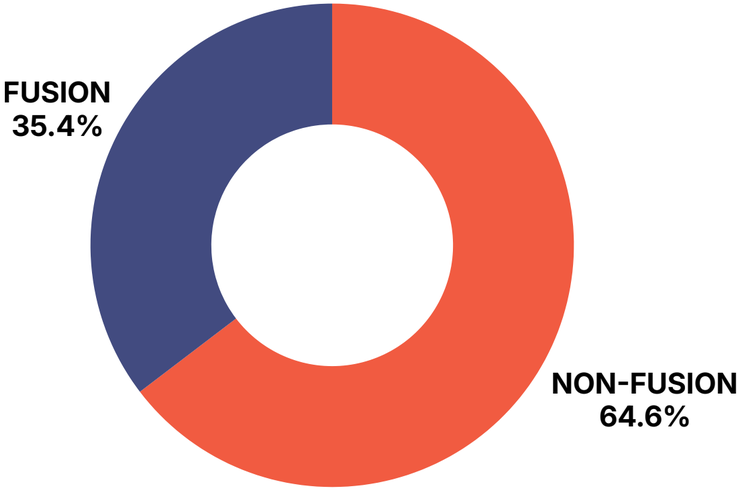


PREPROCESSING



WINDOWING

FEATURE ENGINEERING



Rimozione dei duplicati

FUSION

116126

↓

81481

NON-FUSION

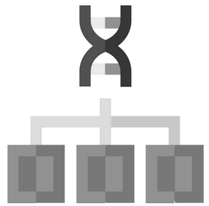
148954

↓

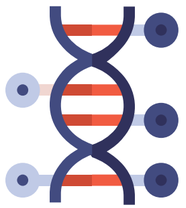
148944

PIPELINE E DATASET

RAW DATA



PREPROCESSING



WINDOWING

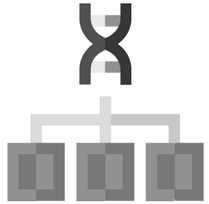
FEATURE ENGINEERING

Riduzione della lunghezza delle sequenza a 1000bp

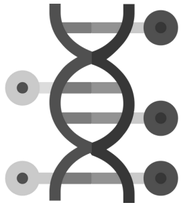
- **Sequenze Fusion:** Finestra centrata sul *junction_point*
- **Sequenze Non-Fusion:** Finestra centrata su un *punto casuale*
- **Gestione bordi**

PIPELINE E DATASET

RAW DATA



PREPROCESSING



WINDOWING

FEATURE ENGINEERING

Estrazione delle features in quattro gruppi

Statistiche di base

GC-Content
Entropia Shannon

Caratteristiche strutturali BWT

Entropia BWT
Run-Lengths
Compression Ratio

Analisi Bio-Linguistica

KL-Divergence
Stop Codon Ratio

Features Ensemble

Gruppo ibrido di features

MODELLI DI CLASSIFICAZIONE

Random Forest
(RF)

Molti alberi decisionali, alta stabilità

XGBoost
(XGB)

Boosting veloce, massime performance

Multilayer Percperton
(MLP)

Architettura a strati per pattern complessi

METRICHE DI VALUTAZIONE

Accuracy

Misura la percentuale complessiva di predizioni corrette (sia Fusioni che Normali) rispetto al totale dei campioni analizzati.

Precision

Indica l'affidabilità delle predizioni positive

Recall

Misura la capacità del modello di individuare tutti i casi positivi

F1-Score

La media armonica tra Precision e Recall

STATISTICHE DI BASE

Features ricavate dalle **sequenze originali**

GC-Content

Rappresenta la percentuale di basi **G** e **C** presenti nella sequenza.

Shannon Entropy

Misura il grado di **incertezza** o **disordine** dell'informazione nella sequenza originale.

STATISTICHE DI BASE

GC-Content

Shannon Entropy

	ACCURACY	PRECISION	RECALL	F1-SCORE
RF	**0.7967**	**0.6997**	**0.7445**	**0.7214**
XG-BOOST	0.5917	0.4508	0.7084	0.5510
MLP	0.6464	0.0000	0.0000	0.0000

CARATTERISTICHE STRUTTURALI BWT

Features ricavate dalle **sequenze processate** con la BWT

Shannon Entropy

Entropia di Shannon calcolata sulle sequenze processato

Run Lenghts

Analisi della lunghezza delle **sequenze con caratteri identici**

↓

MIN MAX MEAN

Compression Ratio

Rapporto di compressione della stringa BWT.

CARATTERISTICHE STRUTTURALI BWT

Shannon Entropy Run Lenghts Compression Ratio

	ACCURACY	PRECISION	RECALL	F1-SCORE
RF	**0.8415**	**0.7981**	0.7387	**0.7672**
XG-BOOST	0.6320	0.4875	**0.7921**	0.6036
MLP	0.6545	0.5224	0.2674	0.3537

ANALISI BIO-LINGUISTICA

Analisi delle sequenze come un **vocabolario di triplette**

KL Divergence

Misura quanto due **distribuzioni di probabilità** divergono tra loro.

Stop Codon Ratio

Conta la presenza di **segnali di stop prematuri.**

ANALISI BIO-LINGUISTICA

KL Divergence

Codon Stop Ratio

	ACCURACY	PRECISION	RECALL	F1-SCORE
RF	**0.8096**	**0.7120**	**0.7749**	**0.7421**
XG-BOOST	0.6334	0.4876	0.7210	0.5818
MLP	0.6483	0.5211	0.0683	0.1208

FEATURES ENSEMBLE

Gruppo ibrido di features contenente le **features più significative** calcolate in precedenza

GC-Content

Entropia BWT

Run Lengths

MIN MAX MEAN

Compression Ratio

KL-Divergence

Stop Codon Ratio

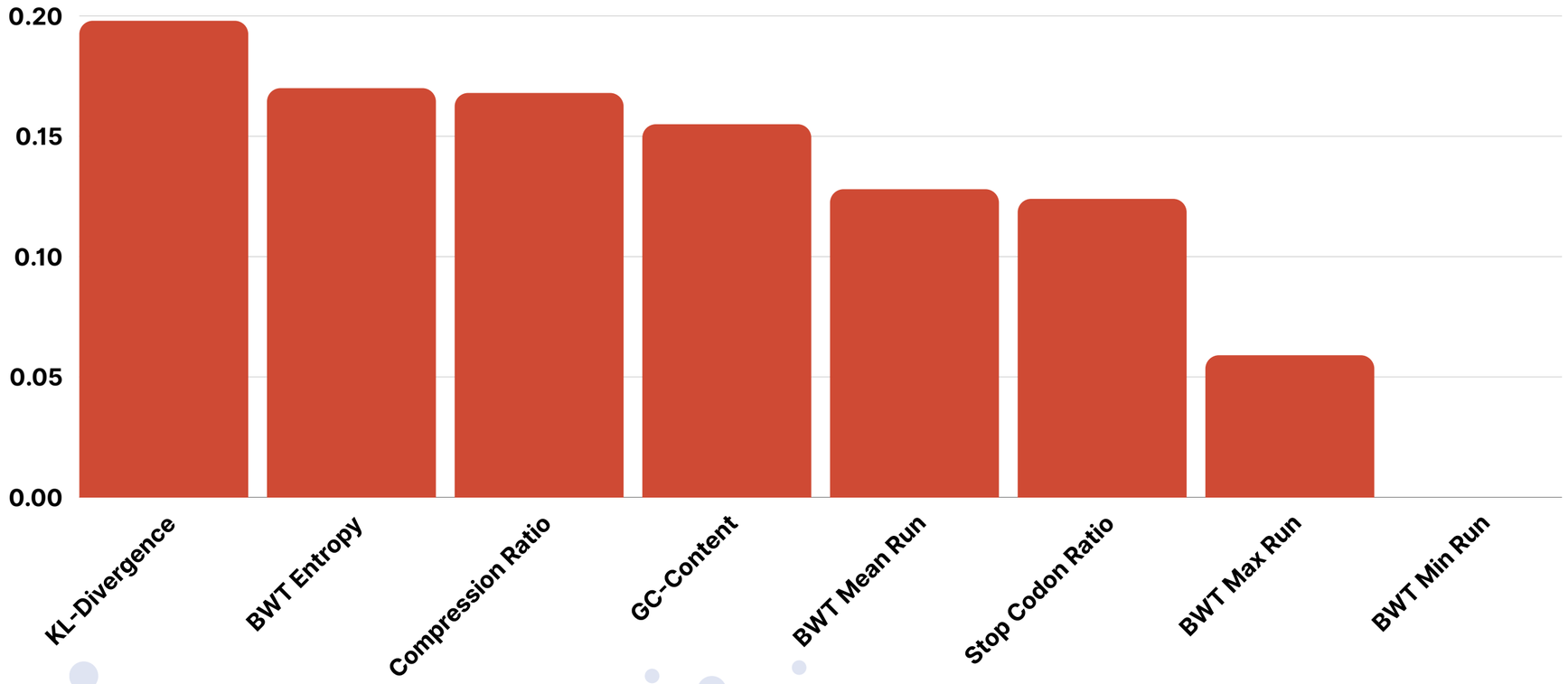
FEATURES ENSEMBLE

GC-Content
Entropia BWT
Run Lengths
Compression Ratio
KL-Divergence
Stop Codon Ratio

	ACCURACY	PRECISION	RECALL	F1-SCORE
RF	**0.8904**	**0.8969**	**0.7796**	**0.8341**
XG-BOOST	0.7124	0.5683	0.7776	0.6566
MLP	0.7193	0.6417	0.4670	0.5406

FEATURES IMPORTANCE

Studio dell'**explainability** del modello, per identificare le features che hanno avuto **più impatto** sulle scelte

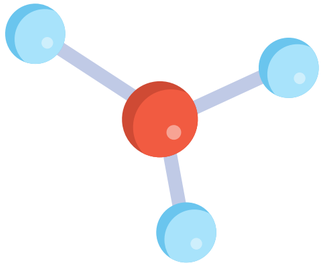


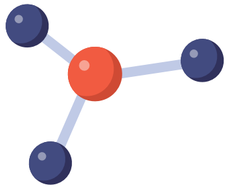
CONCLUSIONI E SVILUPPI FUTURI

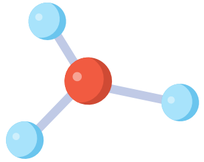
Superiorità delle **features strutturali** e dei modelli **tree-based**

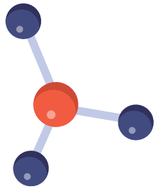
Limiti delle reti neurali su pattern strutturali

Successo del **modello ibrido**

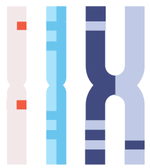






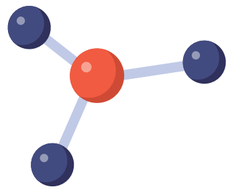


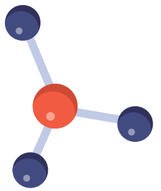
CONCLUSIONI E SVILUPPI FUTURI



Studio approfondito delle **varianti della BWT**

Validazione su **dataset patologici**





GRAZIE!

DO YOU HAVE ANY QUESTIONS?

C.CALABRESE31@STUDENTI.UNISA.IT
D.DELLORUSSO1@STUDENTI.UNISA.IT