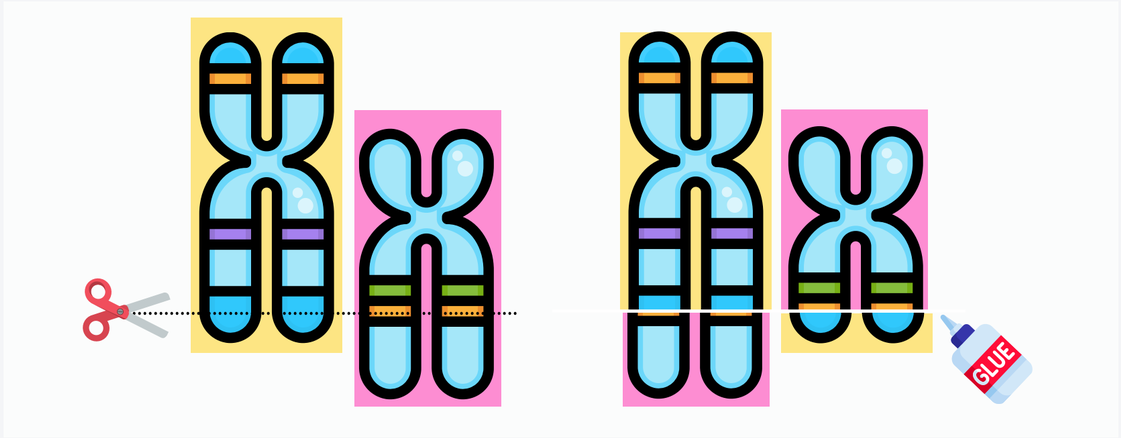
BIOINFORMATIC PROJECT

GENE FUSION 3D PRIORIZATION

A. Arcidiacono, M. D'Amato, A. de Hoffer

Project Overview



Translocation

Interstitial deletion

Chromosomal inversion

Hybrid gene ⋯⋯⋯⋯⋯⋯⋯⋯ **Hybrid protein**

Translation

Onco

Not Onco

The tool simulates gene fusion and assign the probability that pair is an oncogene exploiting its protein's 3d structure

DATASET

DEEPrior

Chromosome number of **5p** gene

Breakpoint coordinate of **5p** gene

Chromosome number of **3p** gene

Breakpoint coordinate of **3p** gene

Label (1 Onco, 0 not Onco)

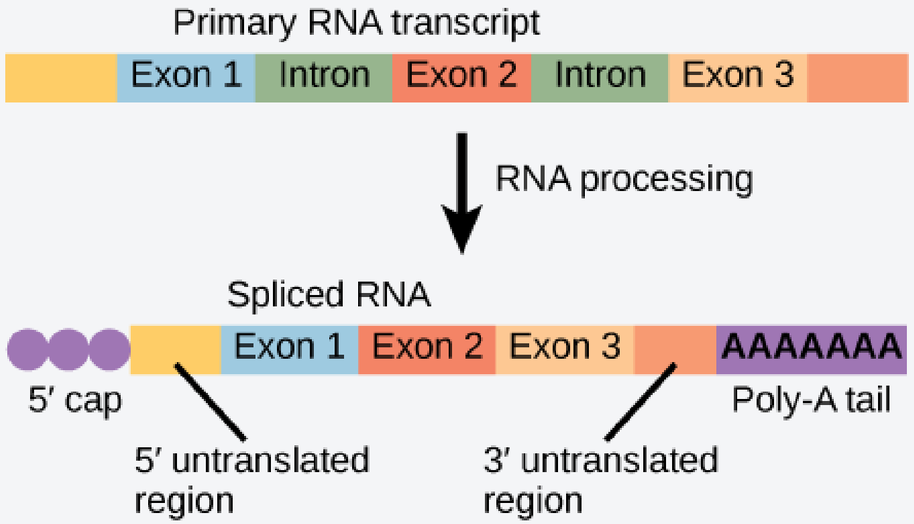
456 546

Gene Fusion class

Dataset Filtering

Annotation: Ensebl

Regions: CDS

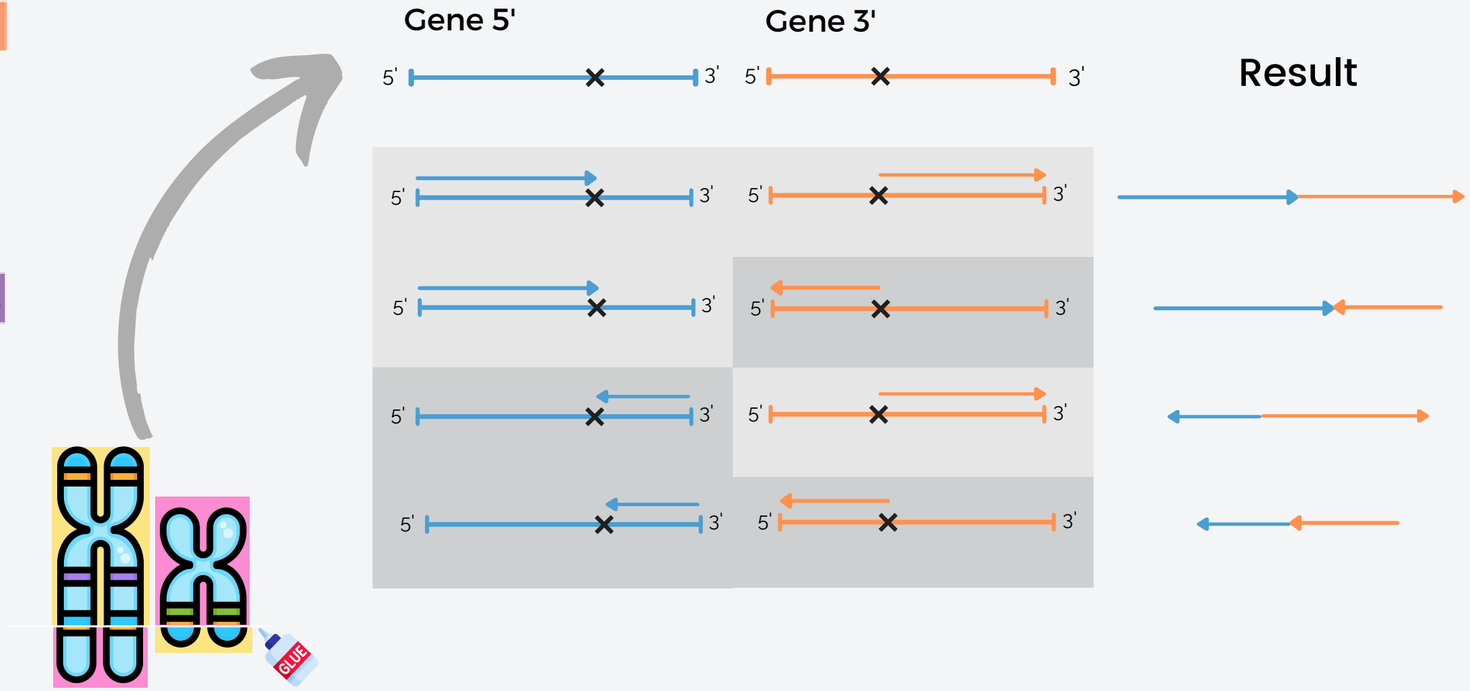


Simulation

Intron

Exon

UTR

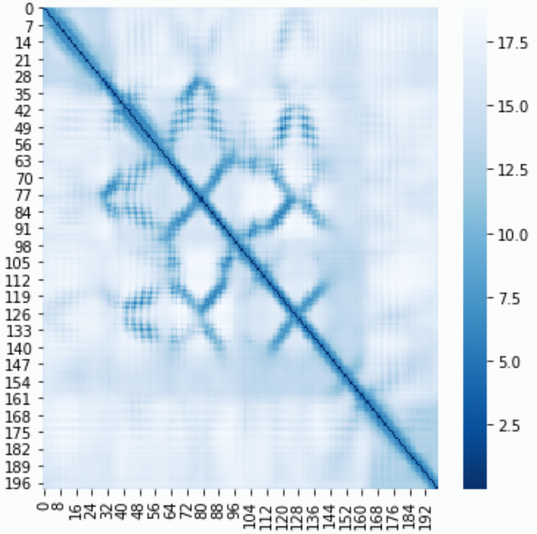


PROTEIN STRUCTURE

Proteins' interaction is guided by their 3d structure!

Primary
Secondary
Tertiary
Quaternary

Retrive **3D structure** from primary structure and then using it for prediction



Due to limit in computational power
Distance matrix

HHBLITS E PROSPR

Primary Structure

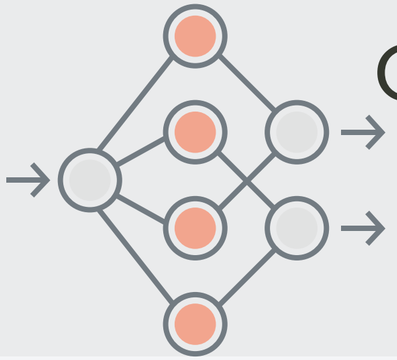
HHBlits

Sequence Profile

ProSPr

Distance matrix
Secondary Structure
Phi angle
Psi Angle
ASA

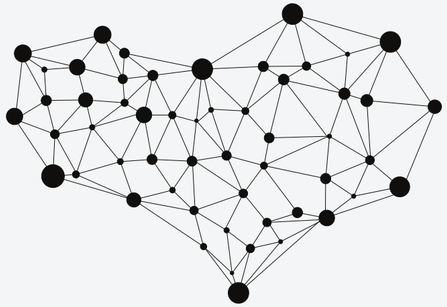
Fasta
MSA
HMM
.a3m



DCA
Resnet
Conv

Topological Data Analysis

Can we find a pattern?



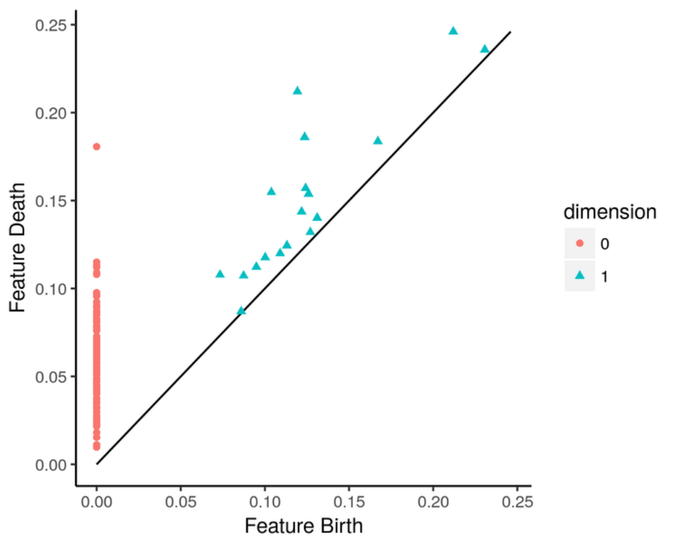
We can see the distance matrix as a graph where the nodes are the aminoacids and the weights are the pairwise distances between them. In this way, we can study the **topological features** of these spaces representing the objects as simplicial complexes and using **Topological Data Analysis** and **Persistent Homology**.

H0: zero-th homology group → detects connected components

H1: first homology group → detects holes

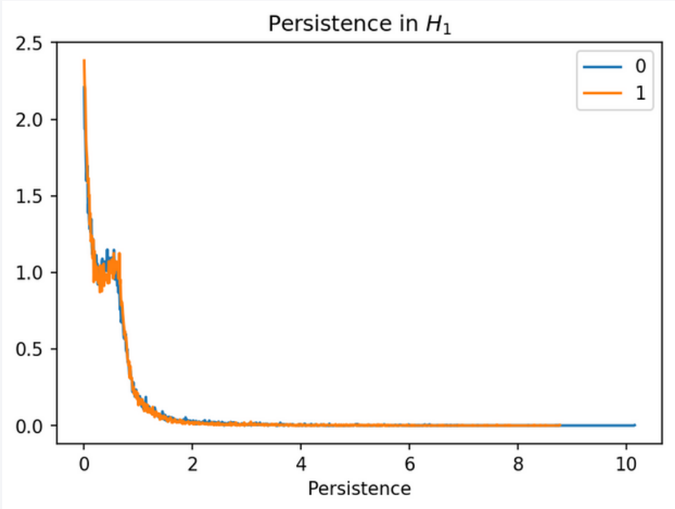
We can analyze persistence that keep tracks of when features appear and disappear.

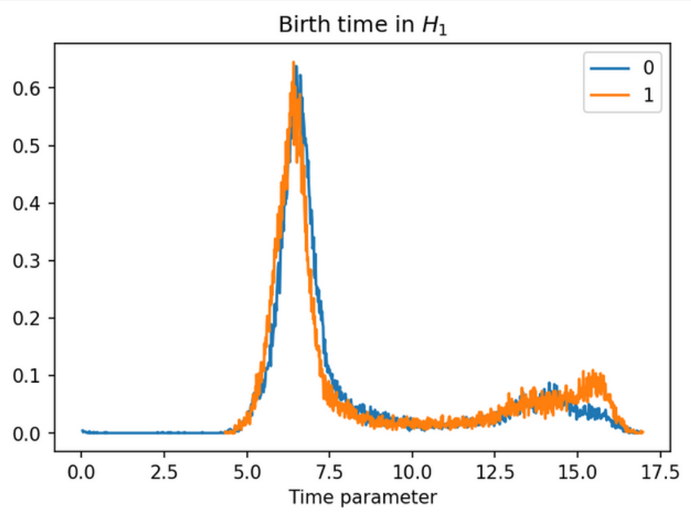
Topological Data Analysis



Persistence diagrams

On the x-axis we have the birth times and on the y-axis we have the death times. We compute one persistence diagram for each associated distance matrix and then we obtain some probability distributions for the onco and not onco fusions.





Kolmogorov-Smirnov

We compute the distance between the two distributions using the Kolmogorov-Smirnov test.

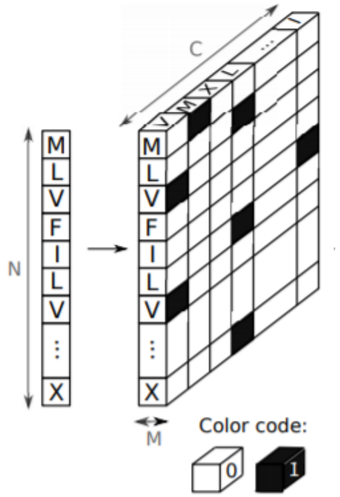
p-value persistence: 10^(-16)

p-value birth: 10^(-29)

We can reject the null hypothesis that the two distributions are the same.

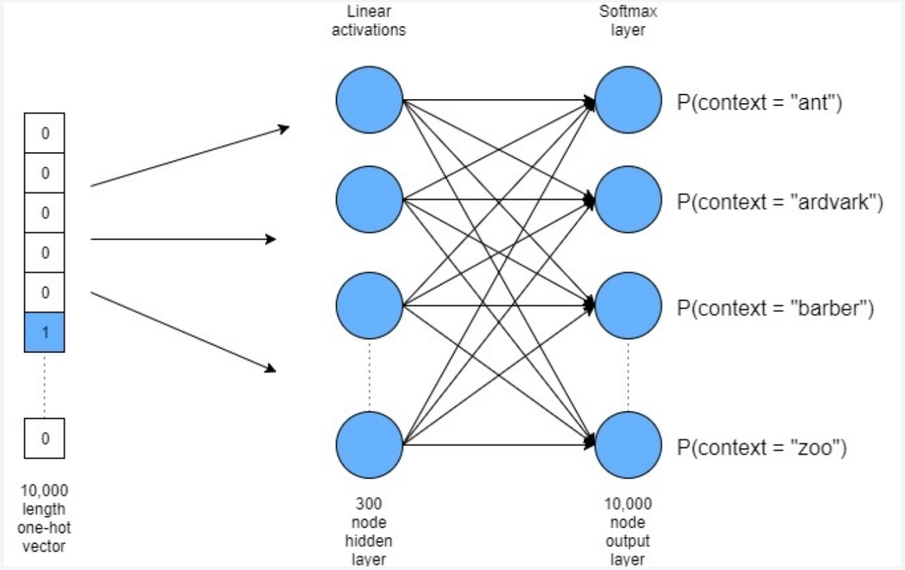
Encoding

One Hot Encoding



Seq length x 20

Word2Vec

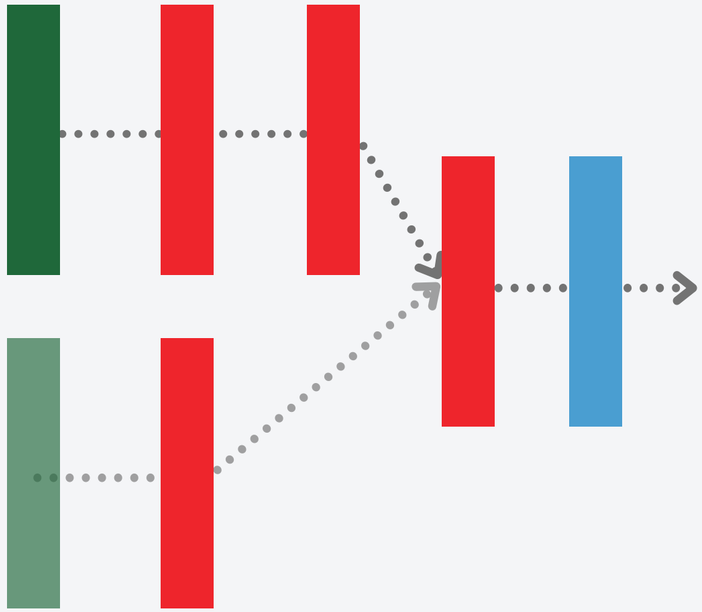


100

Our networks

Matrix
1 Channel

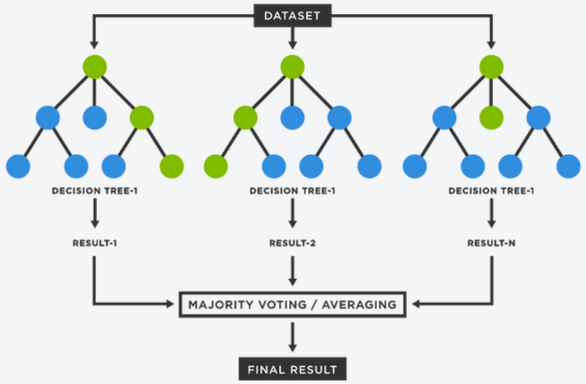
Sequence



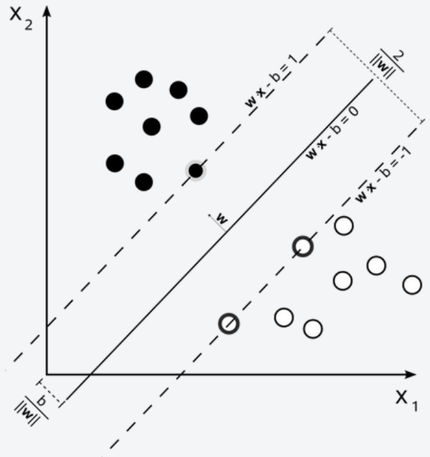
Classification
Priorization

Conv
Linear
Classifier

RANDOM FOREST..



.. and SVM



Matrix → 1 channel Vectorized

Sequence → Vectorized

Results

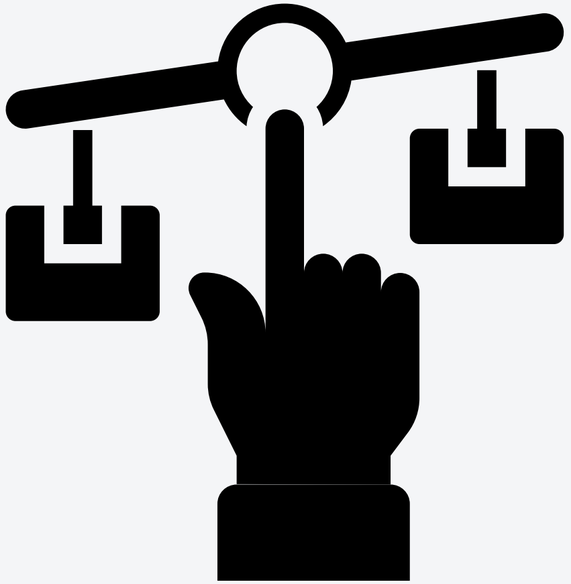
	CNN	Random Forest	SVM
MATRIX	?	87%	
HT SEQ	/	90%	90%
W2V		91%	89%

CONCLUSION

Computationa complexity
Error propagation
Not available repository

Cannot exploit 3D structure at its best

Information are already stored into the sequence!



Thank you!