LIFEMap

FSC Fondo per lo Sviluppo e la Coesione

Ministero della Salute

Progetto LIFEMap

Seminario scientifico

28 febbraio 2025
12:00 - 13:00

Enhancing Pathogenicity Prediction of BRCA Variants Using BERT-Based Genomic Language Models

Relatore: Leonardo Masci – Università de L'Aquila

Abstract

Accurately identifying a genetic mutation, especially VUS, is crucial in determining its possible effect on human health and influencing medical choices. Recent developments in deep learning, specifically transformer models that use attention mechanisms, have made great progress in NLP and offer promising possibilities for the interpretation of genomic data. This study examines how DNABERT2, a Genomic Language Model based on BERT, can improve pathogenicity predictions for BRCA variants. We focus on BRCA genes because of their significant connection to breast cancer and the large number of expert annotations available. The DNABERT2 model has been fine-tuned using both benign and pathogenic labeled sequences, constructed starting from ClinVar variants annotations by modifying an initial wild-type sequence. The transformer-based model shows high accuracy in classification tasks, surpassing traditional tools that can vary depending on the genomic regions and types, achieving around 90% accuracy. The model successfully detects hidden patterns in the genetic code, offering potential help to biological and medical professionals in their annotation work.

Link - https://meet.google.com/vaa-oczm-yjy