St. Jude Children's Research Hospital

St. Jude BioHackathon

Title
Machine learning pipeline to predict locations of mutations in different cancer types

Category
Processing Pipelines And Methods

Challenge
Can we predict which missense mutations will likely appear (at the residue level) for a given cancer type? The challenge would consist of building an ML pipeline to predict the locations and substitution of mutations in genes in samples of different cancer types. Current ML solutions either aim at identifying cancer driver genes (Malebary et al, Sci.Rep2021) or used older technologies (SVM, 10.1109/iarwisoci.2014.7034632). The aim of the challenge would be to assess with state-of-the art ML methods how accurately we are currently able to estimate the appeareance of cancer mutations.

The input features for each sample could consist of amino-acid sequences of a target gene and a cancer type label, and the participants would be asked to predict the location and amino-acid substitution observed in that gene and cancer type. This data could be obtained from available public databases (I.e. COSMIC, TCGA, etc).

Teams would be free to pull in additional genomic data (I.e. homologous sequences, natural variant data), structural data, network data etc, with the exception of explicitly using cancer mutational data to build a predictor. The teams pipelines would be benchmarked on a hold out set of mutated cancer genes.

Benefit
This challenge could allow assessing what the efficacy of current ML state-of-the art methods (DL, transforms, protein language models etc) are on predicting the likelihood of mutations in cancers. Such ML pipelines could be beneficial both for understanding what genomic and biological features impact the likely location of mutations in cancer gens, as well as helping analyze and process the large amount of data available in the St.Jude PECAN database related to mutational landscae in paediatric cancers.

Helpful Tools, Packages, or Software
ML: Free choice of tools (Tensorflow, Keras, Pytorch, JAX etc) Pretrained models: MSA transformer (FAIR resource), UniRep

Test Data
Cancer genomic datasets: TCGA, COSMIC, SJ. PECAN?