Project Title: KNNOR= A KNN minority Oversampling for predicting imbalanced Rare and Common Disease-Associated Non-Coding Variants

Project Description:

Disease and trait-associated variants represent a tiny minority of all known genetic variation, and therefore there is necessarily an imbalance between the small set of available disease-associated and the much larger set of non-deleterious genomic variation, especially in non-coding regulatory regions of human genome.

Machine Learning (ML) methods for predicting disease-associated non-coding variants are faced with the causality dilemma - such variants cannot be easily found without ML, but ML cannot begin to be effective until a sufficient number of cases have been found. Most of state-of the- art ML-based methods do not adopt specific imbalance-aware learning techniques to deal with imbalanced data that naturally arise in several genome-wide variant scoring problems, thus resulting in a significant reduction of sensitivity and precision. Most approaches dealt with this dilemma when calculating the performance measure and uses methods such as balanced accuracy etc…

We present a novel method that adopts imbalance aware learning strategies based on nearest neighbor resampling techniques and apply it in two different contexts; the prediction of non-coding variants associated with Mendelian and with complex diseases.

Goal is to show that imbalance-aware ML is a key issue for the design of robust and accurate prediction algorithms and propose a method and an easy-to use software tool that can be effectively applied to this challenging prediction task.

Available to the student are:

-pseudo code

-One manuscript to use as an example

-Data to use which requires the student to download

Duties/Activities:

1- Get familiar with the pseudo code that we have written in Matlab and python and used on simulated data
2- try it on several simulated datasets
3- Apply it to the genetic variant datasets to generate new data
4- Combine the code with RF algorithm
5- Check the accuracy of prediction when combining KNNOR with RF classifier.

6- Available data are:

Neg. selection	imb.ratio
Mendelian data	
±100 Kb	1:302
±500 Kb	1:1432
±1000 Kb	1:2765
TAD	1:1406
GWAS data	
±100 Kb	1:80
±500 Kb	1:277
±1000 Kb	1:409
TAD	1:269

Required Skills: fluent in programming either in Matlab, or Python or R

Preferred Intern Academic Level: None

Learning Opportunities:

Student will learn about bioinformatics and computational biology, Mendelian disease, writing a paper to publish ….

Expected Team Size: *it is preferable to have team projects*

Mentors

Name: Halima Bensmail, S. Brahim and A. Lattab

Email: hbensmail@hbku.edu.qa, alattab@hbku.edu.qa