DNA/RNA/Virus Sequence Classification

Contact:**Muhammad_Nabeel.Asim@dfki.de**

Description

DNA/RNA/Virus sequence classification is widely considered an important task to better understand biological processes , their interaction with heterogeneous internal and external entities, association with a variety of convoluted diseases, and effective drug targets [1]. Although, experimental approaches used for biomedical data analysis have gradually matured and are achieving decent predictive performance across different task. Nevertheless, these approaches are extremely time consuming, expensive, less scalable, and provide less coverage. Besides, these approaches are far less efficient for large scale testing and less feasible for a large community of bio-informatics researchers [2]. Building on these shortcomings and availability of ever growing sequencing data, with each passing day, more and more computational predictive methodologies and frameworks are emerging [1], promising accuracies of which for different biomedical tasks have revolutionized the trust on healthcare applications based on Artificial Intelligence.The goal of very project is to implement and investigate multifarious machine and deep learning based methodologies which can achieve the optimal predictive performance for the tasks come under the umbrella of DNA, RNA or Virus sequence classification. **NOTE: Multiple students can work on this project, however each student will be assigned a different task which come under the umbreall of DNA, RNA, or Virus Sequence classification**

[1] BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches, B Liu - Briefings in bioinformatics, 2019
[2] RNA sequencing: advances, challenges and opportunities, F Ozsolak, PM Milos - Nature reviews genetics, 2011 - nature.com

Requirements

- Python
- Numpy
- Pandas
- Scikit-Learn
- Keras/Tensorflow/Pytorch

Test Task

You are suppose to report the performance of 2 machine learning and 2 deep learning classifier for short E. Coli DNA sequences datasets taken from the UCI Machine Learning Repository. Finding the optimal values of diverse hyper-parameter for all classifiers will be a plus. One such machine learning based implementation along with dataset is available at https://github.com/sajit9285/DNA-Classification-Project. For deep learning, you have to implement 2 distinct classification methodologies (tutorials are available on internet) for same dataset. You have to report accuracy, precision, recall, and F1-score along with experimental setup in form a git repository.