Feature and sample selection over extremely large transcriptomics data sets for human phenotype prediction

Dr Daniel Buchan

daniel.buchan@ucl.ac.uk

Background

Prediction human disease phenotypes from genetic data is a challenging problem with a wide variety of applications in human health. One core application is in the development personalised medicine approaches in health and medicine. There now exists very large datasets of transcriptomics data in humans covering a wide variety of experimental conditions including disease an infections[1].

We are working on very large scale transcriptomics analyses which make use of hundreds of thousands of features. Efficient means for feature and sample selection are needed to enable downstream machine learning analysis. Both to improve the accuracy of classification methods and to reduce the computational costs in building such ML models.

Project Summary

The project will focus on feature selection over the ARCHS4 dataset (https://maayanlab.cloud/archs4/)[2] to find the optimal features for prediction of Human Phenotypes Ontology (https://hpo.jax.org/app/) terms [3]. The ARCHS4 dataset contains every transcriptomics experiment currently hosted at GEO (https://www.ncbi.nlm.nih.gov/geo/)[3] and contains more than 440,000 features. In this project we would like to investigate efficient methods for dimensionality reduction over very large dimensional datasets using filtering and wrapper methods. The project will also investigate whether dimensionality reduction is best accomplished over the whole dataset or on a per-HPO term basis. It may also be possible for motivated students to go on to investigate methods for negative class training set selection in protein bioinformatics.

The final output of the project will be a pipeline or methodology for performing fast, efficient dimensionality reduction over the ARCHS4 dataset.

Requirements

Python programming

Introductory Machine Learning or data sciecne (such as COMP088 or COMP081 or equivalent)

Good to know

Introductory Protein biochemistry

Introductory Bioinformatics (such as COMP082 or equivalent)

Knowledge of HPC programming

Refs and Sample Reading

1) Edgar R, Domrachev M, Lash AE.
Gene Expression Omnibus: NCBI gene expression and hybridization array data repository
Nucleic Acids Res. 2002 Jan 1;30(1):207-10
2) Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A.
Massive mining of publicly available RNA-seq data from human and mouse.
Nature Communications 9. Article number: 1366 (2018), doi:10.1038/s41467-018-03751-6
3) Köhler S, *et al*
The Human Phenotype Ontology in 2021, *Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D1207–D1217, https://doi.org/10.1093/nar/gkaa1043*
4) Stark, R., Grzelak, M. & Hadfield, J.
RNA sequencing: the teenage years.
Nat Rev Genet 20, 631–656 (2019). https://doi.org/10.1038/s41576-019-0150-2
5) Ibrahim Alsaggaf, Daniel Buchan, Cen Wan
Improving cell-type identification with Gaussian noise-augmented single-cell RNA-seq contrastive learning, *doi: https://doi.org/10.1101/2022.10.06.511191*
6) Vicente, A.M., Ballensiefen, W. & Jönsson, JI.
How personalised medicine will transform healthcare by 2030*: the ICPerMed vision. J Transl Med 18, 180 (2020). https://doi.org/10.1186/s12967-020-02316-w*