Human Gene Function Prediction Challenge

Genome-wide screens are experiments where each gene in an organism is systematically perturbed to establish its relationship to a phenotype of interest. In mammalian cell lines, these experiments typically use pooled lentiviral CRISPR-Cas9 libraries to knock out every gene in the organism in a single screen. For example, if the goal is to identify genes which are potential therapeutic targets in a given cancer, a CRISPR genome-wide screen could be conducted on a cell line derived from that particular cancer to identify genes that are essential in that specific genetic background (read [1] and [2] for more information). Or, if the goal is to identify human gene-gene interactions - where a double-mutant organism displays unexpectedly strong or weak phenotypes with 2 specific mutations- a genome-wide screen could be conducted on a single-knockout cell line to construct double mutants.

The genetic dependencies identified by such genome-wide screens have been shown to be very powerful for understanding gene function. Specifically, previous work in model organisms demonstrated that genes that exhibit highly similar dependency profiles tend to be involved in the same protein complex, pathway or biological process (e.g. see [3] for our previous results generating and interpreting this type of data in yeast). With the latest developments in CRISPR-Cas9 technology, genome-wide genetic screens can now be efficiently completed in human cells. This project focuses on developing machine learning approaches that use gene dependency data from genome-wide CRISPR-Cas9 screens to predict human gene function.

Provided data:

(1) **Human cancer cell line gene dependency profiles:** ~17,000 x ~600 matrix
We will provide you with genome-wide screening data from genome-wide CRISPR-Cas9 screens across several hundred cell lines derived from the Dependency Map project (https://depmap.org/portal/). Each row of this matrix corresponds to a single human gene, and each column of this matrix corresponds to a different human cancer cell line. Large negative values reflect a specific genetic dependency on that particular gene in the corresponding cell line (i.e. cases where a gene is specifically essential for growth in that cancer cell line).

(2) GO term annotation matrix: ~17,000 x ~200 matrix
To support a supervised machine learning approach, we will provide you with labels for ~200 different GO terms for which we would like you to build a supervised machine learning model. This matrix is binary and includes a row for each of the genes that appears in the gene dependency profile matrix described above: a 1 in a given position of this matrix means that the gene *is* annotated with the corresponding column's GO term, a -1 means that the gene is *not* annotated with the corresponding column's GO term, and a 0 means that we would like you to make predictions for that gene (a subset of gene annotations have been held back for us to evaluate the performance of your predictions).

Your challenge:

Your goal is to use a supervised machine learning approach to predict GO biological process annotations for each of the ~200 GO terms based on each gene's dependency profile. More specifically, given a single gene's dependency profile of length ~600 (i.e. a row in Matrix 1), you should provide predictions for each of the ~200 GO terms. You can train and evaluate your models on the genes that appear with labels in Matrix 2 (1s or -1s) (the "training" set), and you will submit your model's predictions on the genes that are unlabeled (0s) (the "validation" set). We will provide details on how to submit these predictions later. We will independently evaluate all teams' predictions with a variety of metrics we discussed in class. Note that for genes in the validation set, we have recoded their names such that no information other than the dependency profile matrix can be used to predict gene function.

If you are interested in participating in this challenge, please email Prof. Myers at chadm@umn.edu and we will provide you with the data files above.

References:

[1] Wang et al. Identification and characterization of essential genes in the human genome. Science. 2015 Nov 27;350(6264):1096-101. doi: 10.1126/science.aac7041.

[2] Meyers et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nature Genetics 2017 October 49:1779–1784. doi:10.1038/ng.3984.

[3] Costanzo et al. A global genetic interaction network maps a wiring diagram of cellular function. Science. 2016 Sep 23;353(6306). pii: aaf1420.

[4] DepMap project website: https://depmap.org/portal/