Human Gene Function Prediction Challenge

CSCI 5461 Spring 2023

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. Such double mutant screens have been used extensively in model organisms (see [3] for our previous results in yeast) and the resulting genetic interaction profiles (i.e. the pattern of interactions for a specific gene of interest) have been shown to be highly informative of gene function. With the latest developments in CRISPR-Cas9 technology, double mutants can now be constructed efficiently in human cells. This project focuses on developing machine learning approaches that use recently generated human genetic interaction data from genome-wide CRISPR-Cas9 screens to predict human gene function.

Provided data:

(1) **Human genetic interaction profiles:** ~17k x ~200 matrix

We will provide you with genome-wide screening data from genome-wide CRISPR-Cas9 screens in a human cell line. Specifically, you will be given the results of ~200 genome-wide screens across single-knockout human cell lines, where each value in the ~17k genes x ~200 cell lines matrix serves as a proxy for the fitness of the double-mutant organism with both the row gene and column gene knocked out. Large positive values indicate double-mutant cells that were more fit than expected, and large negative values indicate double-mutant cells that were less fit than expected (i.e. positive or negative genetic interactions).

(2) GO term annotation matrix: ~17k x ~1,000 matrix

To support a supervised machine learning approach, we will provide you with labels for ~1,000 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 genetic interaction 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 challenges (you can tackle either one or both):

Challenge A:

Your goal is to use a supervised machine learning approach to predict GO biological process annotations for each of the ~1,000 GO terms based on each gene's interaction profile. More specifically, given a single gene's interaction profile of length ~200 (i.e. a row in Matrix 1), you should provide predictions for each of the ~1,000 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 genetic interaction matrix can be used to predict function.

Challenge B:

Generating genetic interactions through genome-wide CRISPR screens is expensive and time-consuming. You can think of this process as filling in one column at a time of a tall, thin matrix. The current dimensions of this matrix are 17k x 200 genes, and each new query gene mutant screened using a genome-wide CRISPR screen will add one column to the 200 dimension of this matrix. Because these are genetic interactions, both the rows and the columns of the matrix correspond to genes. Your challenge is to build a machine learning model that, given the row for a given gene (the 1 x 200 library genetic interaction profile), predict the 17k x 1 interaction profile that will result when that gene is screened as a query (see Fig. 1 and Gene Y in Fig. 2a). If we could build an accurate model to predict query profiles, we could avoid screening the remaining ~16.8k genes and could fill in the remaining genetic interactions computationally.

As a training set for this problem, you will be able to use the ~200 genes for which we have already generated genome-wide profiles. We have both the 1 x 200 interaction profile for those genes as library genes as well as their complete 17k x 1 interaction profile as query genes (see Gene X in Fig. 2b). We will withhold an additional ~30 query genes for which we've already collected the data that will be used as a validation set.

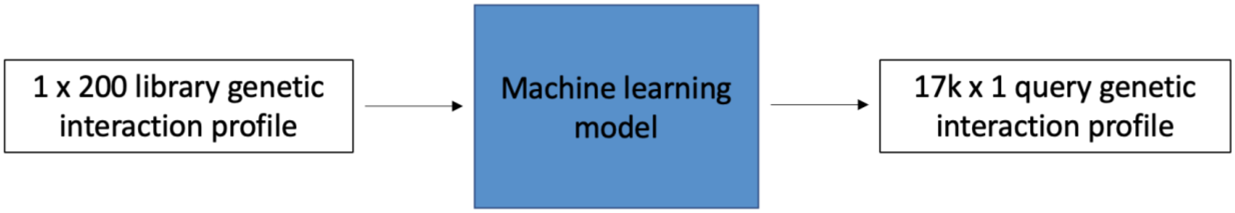
1 x 200 library genetic interaction profile
Machine learning model
17k x 1 query genetic interaction profile

Figure 1. Machine learning prediction task example for Challenge B

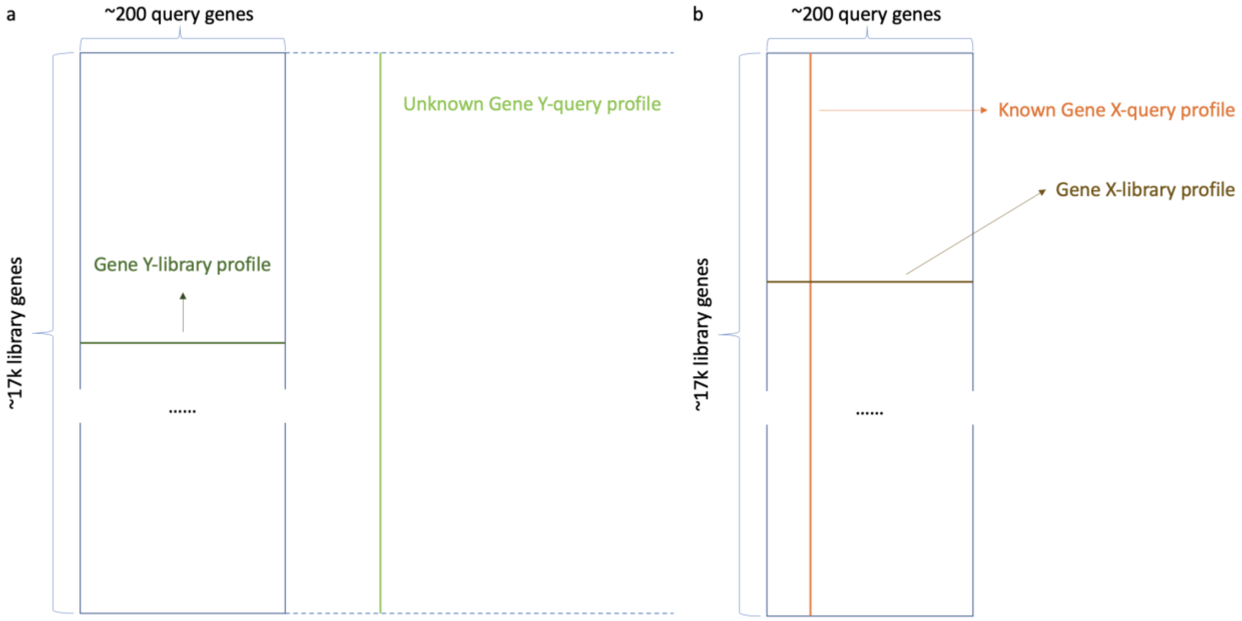

Figure 2. (a) Gene Y only has its library profile available in our dataset and has not been screened as a query yet. The goal is to predict its query genetic interaction profile (17k x 1) based on its known library genetic interaction profile (1 x 200). (b) Gene X is present as both a query gene and library gene in our dataset.

If you are interested in participating in one or both of these challenges, 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.