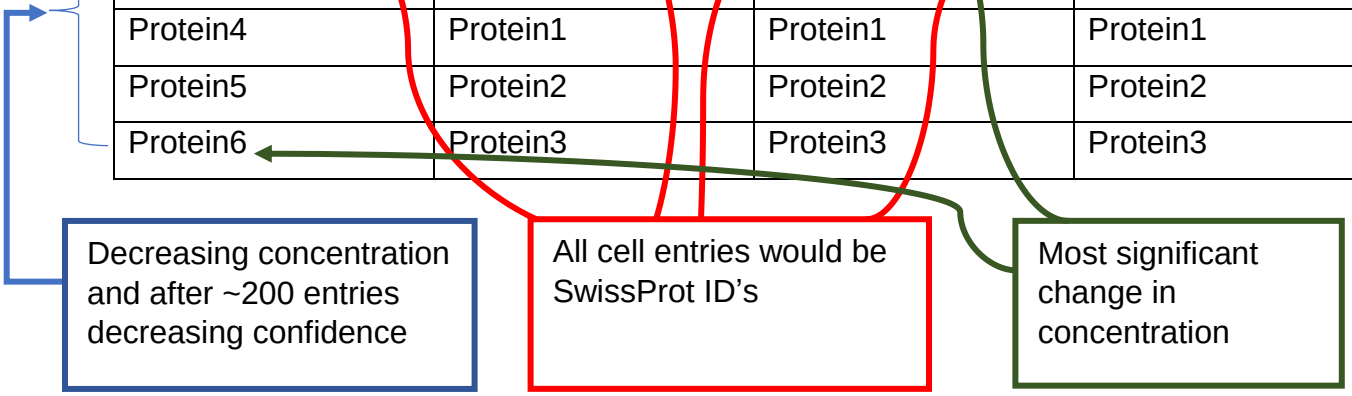
Protein Drug-Target Scoring Python Script

Work done at Merck & Co. is proprietary and for that reason not all information can be revealed.

This Python script was created during my internship in the summer of 2018 at Merck West Point in the Merck Research Laboratories (MRL), Pharmacodynamics, Pharmacokinetics and Drug Metabolism (PPDM) Discovery Bioanalytical (DBA) group. **This project was not asked of me but rather I undertook it out of interest for the problem since it was a tedious task that my superiors were having an issue with.**

The problem was trying to find out the most viable protein to target for a new drug given output from an LC-MS instrument. **Specifically, LC-MS instruments can test multiple unknown samples and create a table that contains columns, for each unknown sample, whereby the columns contain SwissProt identifiers for proteins. Each column is sorted by decreasing concentration/probability that the protein was found in the sample matrix**. An example is provided below:

No treatment	Treatment with Reagent 1	Treatment with Reagent 2	Treatment with Reagent 3
Protein1	Protein7	Protein6	Protein5
Protein2	Protein6	Protein7	Protein6
Protein3	Protein5	Protein5	Protein7
Protein4	Protein1	Protein1	Protein1
Protein5	Protein2	Protein2	Protein2
Protein6	Protein3	Protein3	Protein3



As illustrated above, for a given unknown sample the LC-MS instrument can output a list of more than 200 possible proteins (given by SwissProt ID's) in one column. The beginning elements in the column contain proteins that were found and are sorted in order of decreasing concentration down the column. Usually around the 200th entry and onwards, those proteins may or may not exist and so the column is sorted by decreasing confidence/probability of that protein existing in the unknown sample.

The goal was to create a script that can look at all possible sample conditions and rank all proteins into a new matrix in terms of the change in their concentrations/confidence over different conditions. Once we know those proteins that change the most from different conditions, we can explore those proteins for further drug discovery for eradicating a specific disease.

I created a Python script that took in this matrix with a header row describing each condition and then created a function to score each protein based on how much it has changed throughout all different conditions in relation to the "No treatment" or Wild-Type sample. A final tab-delimited text file with this new matrix was created and contained two columns. The first column had a name describing the two conditions that were compared. The second column contained a score for the two conditions compared. The final file was sorted in order of decreasing order.

The protein script was able to use this scoring function to find multiple novel protein drug-targets that are being further investigated at Merck for curing specific ailments.