# Computational Statistics Final Project

**Note: there are two project options. Each group should select a project out of the two options.**

**Project 1: PrecisionFDA NCI-CPTAC Multi-omics Enabled Sample Mislabeling Correction Challenge**

**Aim:** Human sample labeling or annotation errors could arise during sample transferring, sample tracking, large-scale data generation, and data sharing/management. There is a pressing need to have computational programs to automatically screen for and to correct such incorrect sample labels or annotations. The goal of this challenge is to develop computational algorithms that can accurately detect and correct mislabeled samples and/or data in large-scale multi-omics studies.

**Background:** This is the US FDA challenge, which will be released on September 24, 2018 (https://precision.fda.gov/mislabeling). In this project, we will be focusing only on the "Stage one" part of the challenge. In this part, you will be given (1) clinical and (2) proteomic data for both training and test your model. The aim is to identify and correct mislabels in the test dataset. For details please the description paper here: https://www.nature.com/articles/s41591-018-0180-x.pdf.

**Data set details:**

TBA

**Instructions:**

Work with your group members and design a predictive model for identifying mislabelled samples using the above datasets (and any other data sources that you found to be useful).

*Summary of key aspects:*

- Required outputs**:**
  (1) Probability of each sample been mislabelled; and
  (2) A predicted list of mislabelled samples from training and test dataset, respectively.
- Evaluations:
  (1) Compare to the ground truth in training data, and test data (if such information is provided on the FDA website later on).
  (2) Benchmark the prediction accuracy (consider class imbalance etc.).
- Additional outputs:
  (1) All other visual analytic results.
  (2) All other tables of results that are useful.

*Detailed instructions:*

I. Feature creation, integration, and/or selection. Depending on what data FDA will release, we anticipate having both clinical data and proteomic data for both training and test. These data need to be converted into numeric or some other form that can be used by a classification model. It may also need to be properly normalised for the classification model to work. Consider how to integrate proteomic and clinical features and consider if a feature selection procedure is necessary. (8 points)

II. There are mislabelled samples in both training and test data. Consider how to deal with class label information with noise when you train your model. Consider how to deal with potential imbalance class distribution. (9 points)

III. Evaluate and benchmark your predictions. Compare your predictions to the ground truth in training data, and test data (if such information is provided on the FDA website later on). Think about how to benchmark the performance of your proposed classification procedure on the training dataset. Consider on to assess the model performance on the test data if the mislabelled samples are unknown. (8 points)

Write a detailed Rmd (R markdown) report to document the project. Include problem description, your implementation in R codes, detailed comment of your R codes, and discussion on each decision you have made on performing classification. Prepare a presentation based on your project report.

**Project 2: Kinase-substrate prediction using phosphoproteomics datasets**

**Aim:** Prediction is a central application in many real-world data analyses. In this project, we will aim to apply classification techniques for predicting novel kinase-substrates.

**Background:** Protein post-translational modifications (PTMs), which can activate or inhibit protein function/activity, have emerged as key regulators of various signalling pathways. Phosphorylation is a common type of PTM that is characterised by the addition of a phosphate group by a protein kinase to a serine, a threonine, or a tyrosine residue on a substrate protein. Recent advances in mass spectrometry (MS)-based technologies make it possible to profile proteome-wide phosphorylation events *in vivo* for investigating signal transduction cascades. A key goal is to identify the set of kinases and their corresponding substrates that underlie key signalling events over a course of time.

The time-course phosphoproteome profiling of insulin stimulated fat cells (3T3-L1) conducted by Humphrey et al. (2013) provides a unique opportunity to reveal previous unknown aspects of insulin pathways, a key for treating type II diabetes. Previous knowledge suggests that Akt and mTOR are central kinases involved in the insulin signalling in fat cells. It is known that kinases regulate their substrates by recognising substrate peptide sequence motif and substrates of the same kinase often have similar response profile. To this end, we aim to predict novel substrates of Akt and mTOR by using and/or extracting learning features from temporal phosphoproteomics data (and, if possible, combining these features with kinase-substrate recognition sequence motif).

**Data set details:**

- InsulinPhospho.txt

The main phosphoproteomics dataset. Rows are phosphorylation sites and columns are specific characteristics of each phosphorylation site including:

| Column names | Details |
|---|---|
| Identifier | Identifier for each phosphorylation sites. It is a combination of *gene symbol* (character) and *phosphorylation site* (numeric) on that gene. |
| Seq Window | This column contains the sequence of the phosphorylation site. It is a 13-amino acid sequence in which the 7$^{th}$ position corresponds to the amino acid that is phosphorylated. |
| 15s, 30s, 1m, 2m, 5m, 10m, 20m, 60m | Time-course insulin stimulated phosphorylation profiles in log2 fold change compared to time point 0. Time point 0 corresponds to the fat cells prior to insulin stimulation. |
| Avg Fold | Average Log2 fold change of the entire time-course compared to time point 0. |
| AUC | The area under the curve of a phosphorylation sites compared to time point 0. This is scaled to [0, 1]. |
| Ins 1, Ins 2 | Fat cells with insulin treatment (20 min). Two replicates. These are matched to LY and MK treatment. |
| LY | A generic inhibitor of large part of insulin pathway. It is treated to inhibit Akt and mTOR from phosphorylating their substrates. |
| MK | A specific inhibitor of Akt. It is treated to inhibit Akt from phosphorylating their substrates. |

- Akt_substrates.txt

This file contains identifiers of known Akt substrates.

- mTOR_substrates.txt

This file contains identifiers of known mTOR substrates.

**Instructions:**

Work with your group members and design a predictive model for identifying novel Akt and mTOR substrates using the above datasets (and any other data sources that you found to be useful).

*Summary of key aspects:*

- Required outputs**:**
  (3) Probability of each phosphorylation sites been a substrate of Akt or mTOR; and
  (4) A predicted list of substrates for Akt and mTOR, respectively.
- Evaluations:
  (3) Compare to previous predictions (Prediction_2016.xlsx).
  (4) Benchmark the prediction accuracy (consider class imbalance etc.).
- Additional outputs:
  (3) All other visual analytic results.
  (4) All other tables of results that are useful.

*Detailed instructions:*

IV. Feature creation, integration, and/or selection. The raw features are the time-course log2 fold changes after insulin stimulation and log2 fold change of inhibitor treatments (LY and MK). We have extracted average log2 fold change and AUC from the time-course data and created two additional features to summarise the time-course features. Consider extracting additional features if possible, consider how to integrate these with motif information, and also consider if a feature selection procedure is necessary if you would come up with a large number of potential features. (8 points)

V. Only a subset of substrates of Akt and mTOR are known, and there may be known substrates of Akt and mTOR that are not activated by their respective kinases in fat cells and in our experimental conditions. In addition, there is a large number of phosphorylation sites that have no known kinases associated with them (you could look into PhosphoSitePlus database which contains a large number of "known" kinase-substrate relationships, but with higher false positives). Consider how to deal with partial class label information with noise. Consider how to deal with potential imbalance class distribution. (9 points)

VI. Evaluate and benchmark your predictions. Compare your predictions to the prediction in our previous report (Prediction_2016.xlsx). Think about how to benchmark the performance of your proposed classification procedure when the "ground truth" is (partially) unknown. Consider using simulation and/or create a conservative lower and/or upper bound for performance. (8 points)

Write a detailed Rmd (R markdown) report to document the project. Include problem description, your implementation in R codes, detailed comment of your R codes, and discussion on each decision you have made on performing classification. Prepare a presentation based on your project report.

**Presentation Rubric:**

Mark for presentation will be determined by how well the above key points are addressed. Each group will have a maximum of 7 mins for presentation. Please explicitly note each member's contribution to the project on your final slide. Each group will receive a single mark and all members in a group will have the same mark unless group members are noted to contribute unequally.

| **Key points** | **Exceptional** | **Proficient** | **Fair** | **Developing** | **Inadequate** |
|---|---|---|---|---|---|
| (1) Feature creation, integration (and selection if necessary) | 8 | 6.5 | 5 | 3.5 | 1 |
| (2) Proposed a classification procedure to integrate datasets for prediction | 9 | 7 | 5 | 3 | 2 |

| | | | | | |
|---|---|---|---|---|---|
| (3) Validation and benchmark of prediction results | 8 | 6.5 | 5 | 3.5 | 1 |
| (4) Presentation quality | 5 | 4 | 3 | 2 | 1 |