**Project Proposal**

Anna Romanov 321340580 [annarom@campus.technion.ac.il](mailto:annarom@campus.technion.ac.il)

Maxim Kolchinsky 320983216 [kolchinsky@campus.technion.ac.il](mailto:kolchinsky@campus.technion.ac.il)

Project 9 - Comparing between supervised and unsupervised approaches to classify gene expression profiles of cancer patients

**Dataset:** We chose the following dataset for our project:

<https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5801&fbclid=IwAR2YoCBZlS1p9Ef7H0dMCqSAj0rsk8ysOuIEyde_3mqnsnb51GnjKHaHNeU>

Dataset title: Protein kinase C δ deficiency effect on breast cancer cells.

Dataset description: The dataset contains gene expression results for cells with down-regulated PKCδ, as well as non-treated cells, from 2 breast cancer cell line types. Data from over 47,000 probes is present. In addition, many of the probes are associated with certain GO terms.

This dataset is linked to the paper “Down Regulation of CLDND1 Induces Apoptosis in Breast Cancer Cells” by Chandrani Achari, Sofia Winslow and Christer Larsson.

**Biological background:**

Identification of targets for apoptosis induction is important to provide novel therapeutic approaches in breast cancer. Earlier studies by the authors showed that down regulation of protein kinase C δ (PKCδ) induces death in breast cancer cells. With the goal of identifying previously unrecognized apoptosis regulators in breast cancer cells, global expression analysis with microarray was performed after down regulation of PKCδ in the basal breast cancer cell lines. The cell lines used were BT-549 and MDA-mB-468 , both are breast cancer cell lines from Invasive ductal carcinoma and adenocarcinoma respectively.

Results show that some differential expression levels such as CALM3 siRNA effectively increased cell death in some of the cell lines while others showed inconsistent results.

**Main question**: Our goal is to classify genes according to GO terms associated with cell apoptosis which is induced by the treatment.  
We will compare the results and insights obtained from supervised and unsupervised methods.

The purpose of our research is to be able to identify whether a certain gene is affected by PKC. By knowing this information, we will be able to know ahead of time whether PKC treatment will be effective on other types of cancer based on the genes associated with those types.

Clustering of genes will give us indication on how high the expression of a gene is as a consequence of the treatment which will also help us judge not only the amount of genes which will induce apoptosis following the treatment but also how effective will the treatment be on those genes.

**Tools:**

In order to generate labels for each gene, we will first extract GO terms that are related to cell apoptosis. For that, we will use the site “amigGO” http://amigo.geneontology.org . We’ll use python scripts to assign a GO term for each gene in the dataset, based on the set of terms already assigned to genes in the dataset.  
We will use the python library scikit-learn in order to apply machine learning algorithms on the data.