Gene Expression Distance

Initial Data Acquisition

The files “RankMatrix.txt” and “cmap_instances_02.xls” are downloaded from the following link:

https://portals.broadinstitute.org/cmap/index.jsp#

Prefiltering

The dataset consists of samples taken from four different cell lines. At the first step, each cell line is selected separately. These are the MCF7, the HL60, the PC3 and the SKMEL5. Out of the 6100 instances of the total dataset, 3113 belong to the MCF7 cell line, 1229 to the HL60 cell line, 1741 to the PC3 cell line and only 17 to the SKMEL5. This fact leads to the exclusion of the SKMEL5 cell line from the dataset, since it is poorly represented when compared to the rest cell lines.

At the next filtering step, each cell line is treated as a separate dataset.

MCF7

In this cell line, three arrays are present, HG-U133A, HT_HG-U133A_EA and HT_HG-U133A. In a total of 3113 instances, 2740 belong to the HT_HG-U133A array and only 373 in both of the other two arrays. To enhance the homogeneity of the data, only the 2740 instances are selected, which belong to the HT_HG-U133A array, leaving the rest 373 instances out of the MCF7 cell line dataset. The next filter regards the experiments’ duration. The desired duration is 6 hours, since almost all the samples have this experiment duration. At the MCF7 array HT_HG-U1333A dataset all of the experiments have a 6-hour duration, so they remain 2740. Afterwards, the 2740 instances are further filtered according to their concentration. From each compound, only the instances with the most representative concentration through its duplicates are kept. At the final step, 2616 instances are present. These contain only complete replicates from the 1211 unique drugs that exist in the final MCF7 dataset after the filters. This is the dataset that will be inserted in the GSEA algorithm, after the replicates of each drug that it contains have been merged through the suitable rank merging algorithm.

HL60

The same process is followed also for the HL60 cell line. The total instances of this cell line are 1229. The arrays present at the HL60 cell line are two, HG-U133A and HT_HG-U133A. They have 344 and 885 representatives respectively. Due to both of the arrays’ high amount of instances, they are both kept. Moreover, all of the instances have experiment duration of 6 hours, so nothing is lost in terms of the 6-hour filtering. At the next step, the 1229 instances are further filtered according to their concentration. From each compound, only the instances with the most representative concentration through its duplicates are kept. At the final step, 1168 instances are present. Here, because of the fact that two arrays are present, an extra filter is applied. This filter aims to locate and leave out compounds that exist in both arrays, and keep the ones that are in the HT_HG-U133A array, since this is the one with the most instances in the cell line’s dataset. From this filter, nine instances are also removed. So, the dataset now counts to 1159 instances. These contain only complete replicates from the 1078 unique drugs that exist in the final HL60 dataset after the filters. This is the dataset that will be inserted in the GSEA algorithm,

after the replicates of each drug that it contains have been merged through the suitable rank merging algorithm.

PC3

At this cell line, there is a total of 1617 instances. The total instances of this cell line are 1741. The arrays present at the PC3 cell line are two, HG-U133A and HT_HG-U133A. They have 124 and 1617 representatives respectively. To enhance the homogeneity of the data, only the 1617 instances are selected, which belong to the HT_HG-U133A array, leaving the rest 124 instances out of the PC3 cell line dataset. Moreover, all of the instances have experiment duration of 6 hours, so nothing is lost in terms of the 6-hour filtering. At the next step, the 1617 instances are further filtered according to their concentration. From each compound, only the instances with the most representative concentration through its duplicates are kept. At the final step, 1543 instances are present. These contain only complete replicates from the 1161 unique drugs that exist in the final PC3 dataset after the filters. This is the dataset that will be inserted in the GSEA algorithm, after the replicates of each drug that it contains have been merged through the suitable rank merging algorithm.

After all the cell lines are filtered according to the above mentioned methods, three distance matrices are computed (one for each cell line). They contain the pair-wise distances between all the compounds that are present in each dataset. So, the final distance matrix for the MCF7 cell line will have a dimension of (1211x1211), for the HL60 will be (1078x1078) and for the PC3 will be (1161x1161).