From_genes_to_TF explanation

weighted_mat function

This function is used in order to get an aggregated gene expression column for the compounds that have complete replicates in the dataset. The function takes as input a (NxM) matrix, where N is the number of probes (22283 for array 1 and 22277 for array 2) and M is the number of complete replicates of a compound that exist in each dataset. The function calculates the Spearman correlation between the samples and attributes a weight to each one of them according to it. Finally, the replicated samples are aggregated to a single one, containing information from all of the replicates thanks to this function. The resulting matrix will be of dimensions (Nx1).

Algorithm

At the first step, the instances for the compounds of each cell line are imported. Then, the perturbations that were constructed from the raw data and are in the form of log(Fold Change) will be imported. At this part, only the array 1 (HG-U133A) and array 2 (HT_ HG-U133A) perturbations are needed, since the third array was excluded at the prefiltering process.

In the next step, a perturbation matrix is built for each one of the three cell lines in order to be used at VIPER. The perturbations of each cell line are selected and given as an input to the weighted_mat function (see above). This process provides a final gene expression signature matrix with unique compound perturbations, according to the ones that are present in each cell line.

This method is followed at the MCF7 and the PC3 cell line as mentioned above, as they both contain samples from a single array. In the case of HL60 cell line, the fact that two arrays are present (HG-U133A and HT_HG-U133A) creates the need of two separate sub-datasets, because the probes from an HG-U133A sample are 22283, while from an HT_HG-U133A sample are 22277. Also, two different arrays can't be put as input to get a transcription factor expression from VIPER. So, the dataset regarding HL60 is divided in two datasets that will be treated independently until the transcription factor calculation from VIPER. On this basis, the above procedure is applied to each one of them separately.

At this point, a perturbation matrix with unique gene expression signatures that contain information from their replicates is available for all of the datasets.

These matrices are handled according to the VIPER documentation and formation needs and used as inputs to it. Through VIPER, the transcription factor activity is formed. This comes in a form of matrix, with 279 rows, representing 279 transcription factors and a single column for each one of the compounds inserted to the algorithm. For the two sub-datasets of HL60, the transcription factor activity is computed separately and then bound together to a single dataset, since both of the sub-datasets have the same number of transcription factors in the rows and no information is lost or damaged through this binding.