Documentation for gene Co-expression coefficient analysis

# Background

The gene co-expression analysis was assigned to Atom Wang for Class Project BCB420\_2015 Systemikon. The purpose of this assignment was to generate a matrix containing maximal information coefficients (MIC) of gene pairs given from a gene list in order to find genes that are co-regulated. Those co-regulated genes may aid us identifying an interaction network within a system.

Maximal information coefficient belongs to Maximal information-based nonparametric exploration statistics that identifies and classifies relationships. It captures a broader range of associations compared to the classic Pearson Correlation.

# Method

## Software: R

MIC

The calculation of MIC is based on David *et al.* published in 2011. By using the minevra package in R, the command “mine” easily gives a list of correlation coefficient including MIC

## Dataset

Since the primary interest of Systemikon is to identify underlying biological systems in normal non-diseased tissue in human, gene expression data across multiple tissues of 79 human samples were extracted from Dataset GSE1133 (GPL96 Affimetrix platform). The expression of each probes were then annotated with corresponding Entrez Gene ID.

## Script logic

The script generates a table containing MICs of each gene pair which was asked for correlation. Two “for” loops were used to construct this table (see script for details).

# Result

## Benchmarks

Various numbers of genes were tested for the efficiency of the code. Comparing with Pearson Correlation, it takes a significant more amount of time to finish computing gene expression correlation using MIC (Figure 1).

Figure 1 Computation time comparing MIC and Pearson Correlation. Various gene numbers were chosen and the corresponding time of computation was plotted. Pearson correlation is significantly faster compared to that of MIC.

## Test gene list and Test random list

A test group of genes and a random group of genes were chosen. The test group contains 50 genes and the random group contains 28 genes. 1225 MIC and Pearson Correlation coefficients (PCC) were generated for the test group and 378 were generated for the random group.

## MIC vs Pearson Correlation

To understand MIC’s relationship with PCC, MICs form the test group and random group was plotted against PCCs. The random group showed a more linear like correlation between the MIC to PCC than that in the test group (Figure 2). Interestingly, there are gene pairs that have higher MICs than PCCs.

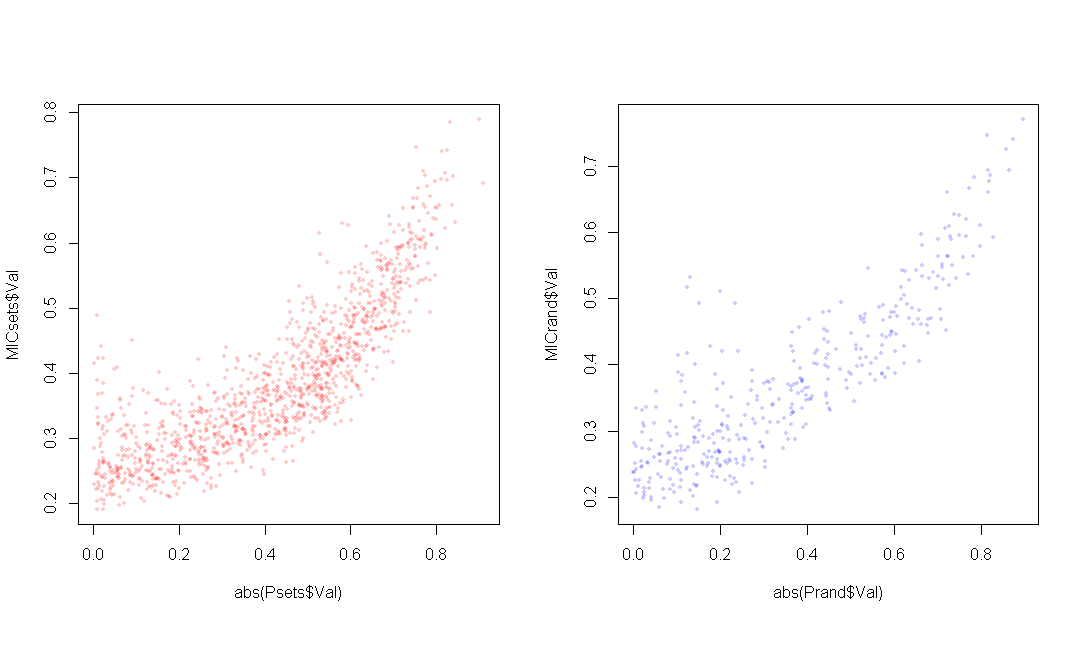


Figure 2 Scatter plot showing MIC against Pearson Correlation Coefficients in test group and random group.

To understand gene co-expression and MIC a bit more, 960 gene pairs were identified to have higher MIC than PCC and 643 gene pairs were identified to have higher PCC than MIC. Gene expressions of the top 4 gene pairs with the greatest difference in MIC vs PCC was plotted and documented in “Gene expression analysis for low MIC high Pearson.ppt” and “Gene expression analysis for high MIC low Pearson.ppt”.

Interestingly, the top 4 gene pairs with MICs higher than PCCs involve the gene ATP2A3, a calcium transporting ATPase. It is ubiquitously expressed yet with diverse levels in different tissues. The genes paired with ATP2A3 are generally all highly expressed in all different tissues. While Pearson Correlation shows there is nearly no correlation in these gene pairs, they have MICs that are above 0.4.

The Top 4 gene pairs with PCCs greater than MICs all showed a near linear correlation between their gene expressions. It is unexpected that MICs gave rise to lower values than PCCs, even though both of them are suggesting a near linear correlation.

# Limitations

There are two major limitation of this script. The first is that there are multiple probes for certain house-keeping genes. The script has yet to omit those out. The second is that the efficiency of the code is challenged when a greater quantity of genes is being analyzed. The calculation of MIC is takes a greater amount of time to calculate compared to PCC. It takes approximately 45 hours to calculate MICs for 5000 genes, yet it only takes 30 minutes to calculate PCCs.