Plan for the Graph CNN

# Construct 4 networks

## Download RNA-seq from TCGA and refer to the paper for preprocessing.

## The expression level of 56716 genes is expressed in log2 (FPKM+1), where FPKM is the number of fragments per million mapping readings per kilobase. In order to reduce the complexity of the model, a total of 7091 most informative genes were selected, with an average expression level> 0.5 and standard deviation> 0.8. In this study, we normalized gene expression between 0 and 1 to ensure the convergence of the model.

## Co-expressions

## Two different PPI (protein-protein interaction) graphs were generated from the string database (<https://string-db>)

## , co-expression graph and PPI graph. To create a co-expression map, Spearman correlation is calculated to generate a correlation matrix between each gene in the data set. If the correlation between two genes is greater than 0.6 and p <0.05, the weight in the adjacency matrix is ​​1, otherwise it is 0. If the correlation with a given gene is not greater than 0.6, the gene is removed from the gene list, resulting in a total of 3,866 genes in the co-expression map. The graph structure consists of 3,866×3,866 adjacency matrix represented by

## PPI Database

## All 7091 genes were entered into the BioMart database to find the corresponding unique Ensembl protein id. All human protein interactions are downloaded from the STRING website. Since there are non-coding genes in the TCGA data set and the number of proteins in the string database is limited, a total of 4444 genes were selected to construct the chart. Consider the links between genes with medium confidence in the STRING database. If you consider the connection between two genes, put a weight of 1 in the adjacency matrix. The PPI graph is represented by 4444 times 4444 adjacency matrix as . Due to the quantity and quality of data coverage, convenient visualization support, and user-friendly file exchange format, the string database was chosen for PPI interaction.

## Singleton Nodes

## All 7091 genes are used in PPI and a single node graph, and all 2647 genes not included in the PPI graph are treated as single nodes. The 7091×7091 adjacency matrix includes the 4444×4444 adjacency matrix PPI in the PPI diagram in the upper left corner, and it is zero elsewhere. The same is true in joint expressions and singleton diagrams. The other 3,225 genes not included in the co-expression map are included as singleton nodes, the upper left corner and zero are expressed in other places as W co to generate a 7091 by 7091 adjacency matrix.

## Using ChebNet, use (W co-expr), (W PPI), (W co) as input to train the classification network. To classify different cancers or normal. Architecture

## 



# Data

All research data are public data sets

Which can be find here <https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html>

# R eference

C [lassification of Cancer Types Using Graph Convolutional Neural Networks](https://www.frontiersin.org/articles/10.3389/fphy.2020.00203/full)