Gene Expression Data Analysis to Identify Drug-Response Specific Gene Networks Using Matrix Factorization

Team: 멜론빙수 (Melon Iceflake)
Lee Sun Ho, Nam Hyo Rim, Moon Chan Gyung
Professor: Park Sang Hyun
Asistant: Ha Ji Hwan

Background

Phenotype: an observable property of an organism such as rash, determined by gene response.
Gene Expression: Gene Expression A gene is called expressed when it produces (usually) proteins
GEO: Gene Expression Omnibus, a public gene data repository
Matrix Factorization: a machine learning technique that uses factorizing of a matrix. A factor is called as "latent space."

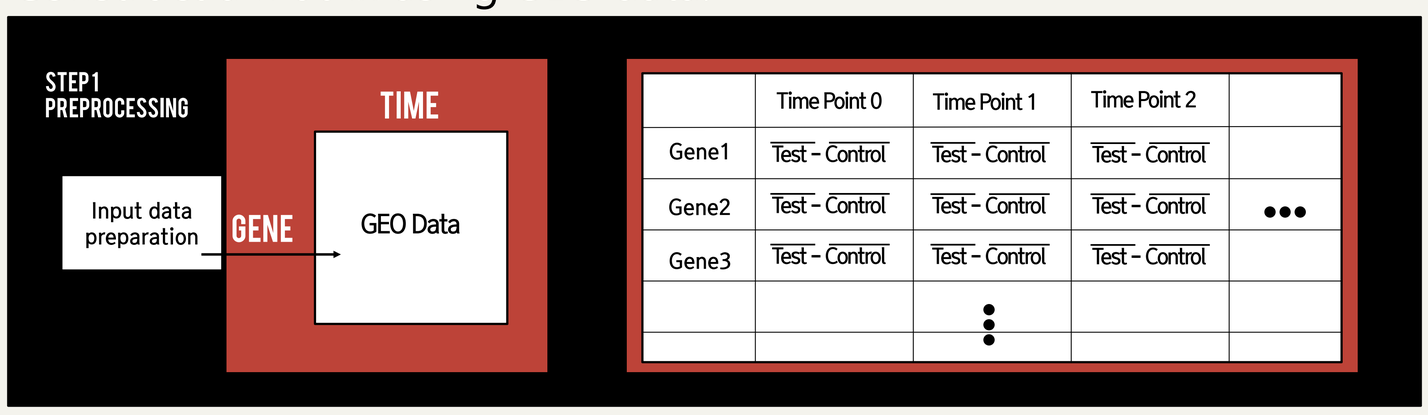
Intro

Identifying gene networks facilitates research on drug mechanism and transcription factor. Existing studies rather focus on individual genes. We take network approach due to interaction among genes. We use Matrix Factorization to extract properties of gene. It also facilitates analyzing datasets of different platforms and data with outliers.

Implementations

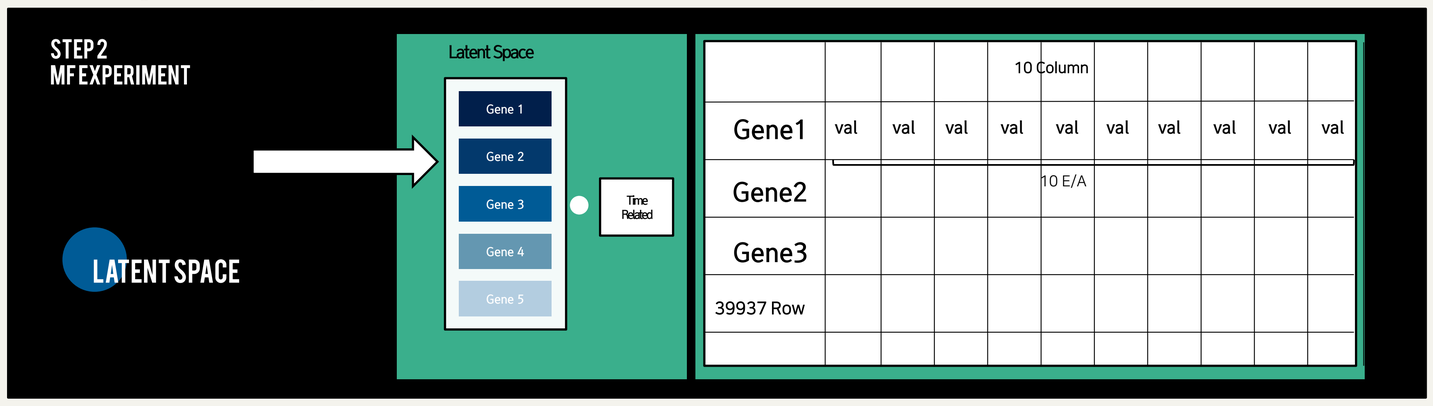
STEP 1(PREPROCESSING)

Construct a matrix using GEO data.



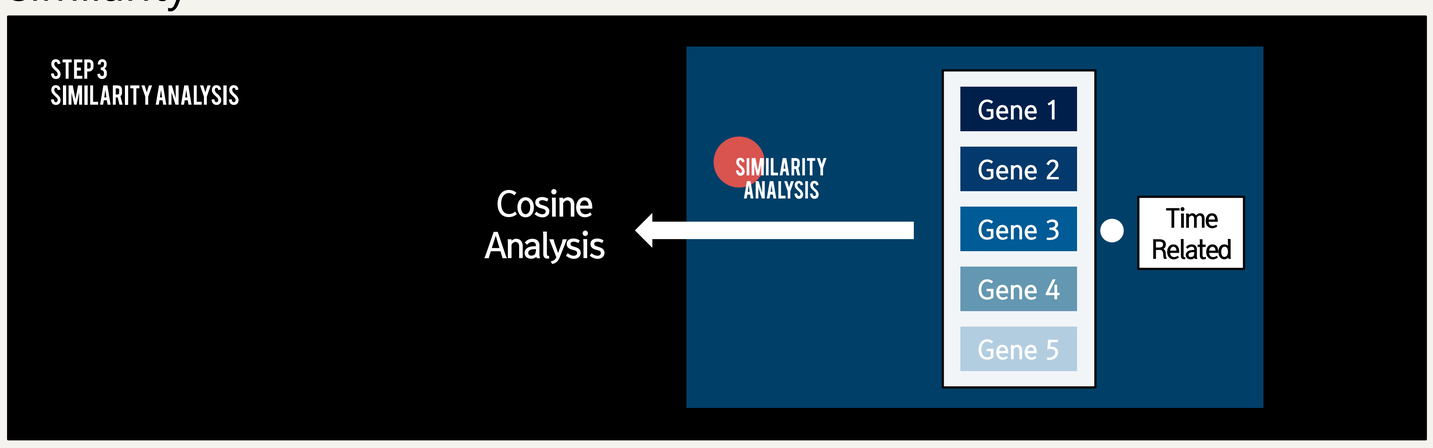
STEP 2(MF EXPERIMENT)

Obtain a latent space of 10 columns.
We adjusted open-source code (LibRec) for Matrix Factorization.

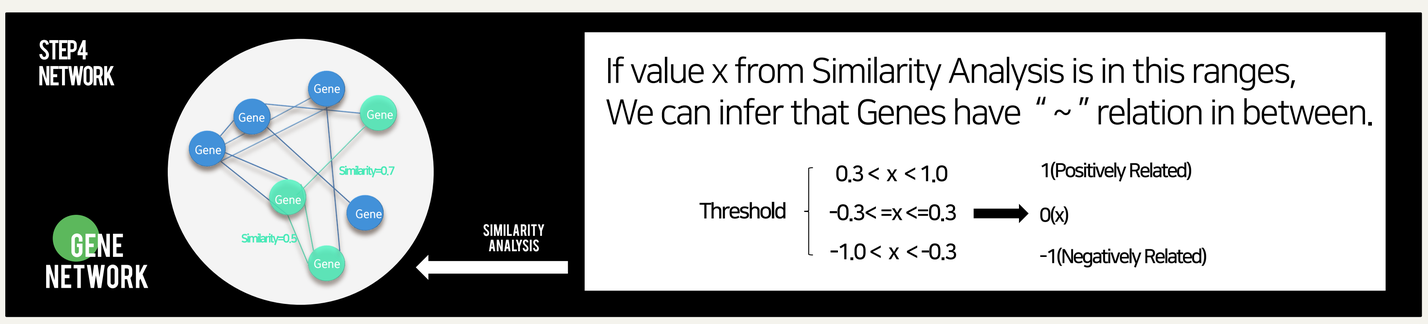


STEP 3(SIMILARITY ANALYSIS)

Measure similarity between genes by using normalized cosine similarity

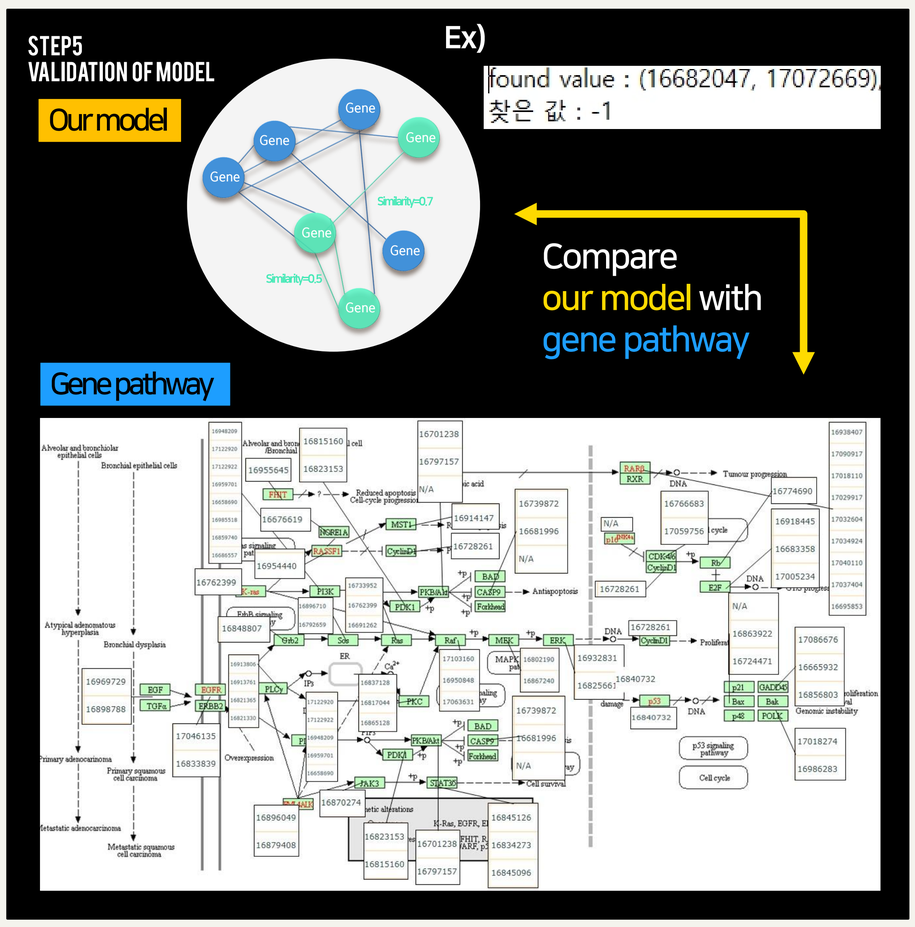


STEP 4(NETWORK CONSTRUCION)



STEP 5(VALIDATION)

Compare the network to existing gene pathway by calculating false negative (type I error).



Our approach

1. Apply Matrix Factorization(Biased MF) to dataset
2. Utilize GEO data(GSE 84094, GSE84095)
3. Use data of Non-Small Cell Lung Cancer of human for validation.

Results

As a result of validation test, type I error occurred. Compared to other models that predict the pathways with moderate precision, our model could predict only about 34% of existing pathways, which requires improvement on precision. Adding extra columns in original matrix is being considered to improve the precision of the MF method.