Predicting gene-gene relationship with CNNC model

Jiajun Li (jiajun.li@epfl.ch), Ningwei Ma (ningwei.ma@epfl.ch) and Zhaorui Li (zhaorui.li@epfl.ch)
Supervisor: Prof. Yimon Aye
Hosting lab: Laboratory of electrophiles and genome operation

***Abstract*—In a complex cell system, it is now clear that biological function can not be attributed to an individual molecule in most of the cases. Network analysis of biological molecules has been proved informative. In gene network analysis, research shows that genes with related functions often have similar expression patterns. Pearson correlation coefficient is the most used representation in measuring similarity for gene-expression data. After getting the quantitative representation of the gene-gene relationship, further analysis such as clustering can be performed to further investigate the gene co-expression network. Due to the unsupervised property of the clustering method and the limitation of the Pearson correlation coefficient, some important gene-phenotype relationships may be dismissed while some false positive relationships may arise. Traditional gene sequencing experiments detect gene expression levels in a large number of cells thus only provide an average expression level which may not reflect the real interaction between genes which leads to false-negative results. Yuan and Bar-Joseph developed a convolutional neural networks-based method called CNNC analysing single-cell gene expression data and inferring gene-gene relationships. Under the hypothesis that genes with related function often have similar expression pattern, the CNNC model learns the expression pattern between two genes and use the learned parameter to predict whether an unknown gene has similar expression pattern with a known one. CNNC is claimed to be able to overcome the shortcomings of traditional gene relationship analysis method. In our report, 4-Hydroxynonenal(HNE) related genes from former experiment and literature are collected as gene labels and were supposed to have a stronger interaction with each other. CNNC model was applied to see if HNE related gene-gene expression pattern could be separated from HNE non-related gene-gene expression pattern. Some adjustments were done to the CNNC model. A possible explanation for the result and a discussion on CNNC model are given in the report.**

I. Introduction

A. Background

It has been over 60 years since Francis Crick stated the central dogma of molecular biology[1] and the central dogma of molecular biology still frames how we understand the information transfer between major biopolymers - DNA, RNA, and proteins. It remains a great challenge to understand the relationship between these information-carrying molecules and biological function.

With the advent of Next Generation Sequencing(NGS), the cost of gene sequencing is becoming more and more affordable, which allows scientists to investigate the relationship between genes and biological function or process at the genome level. In a complex cell system, it is now clear that biological function can not be attributed to an individual molecule in most of the cases[2]. Instead, biological functions are triggered by the interactions between small molecules, proteins, DNA, and RNA. If we abstract these biological constitutions to nodes and represent their interaction in edges, we will get a network (figure 1). The interaction type between biological constitutions is not limited and can be physical interaction, causal interaction, etc. This methodology is now widely used in biology and has been proved informative. Constructing a biology network can help us better understand the process of biological function.

In a gene network, each node represents a gene and the edges represent the interaction between two genes. Depending on the interests in the biological question, the corresponding quantitative representation should be defined. Research shows that genes with related functions often have similar expression patterns [3]. With this hypothesis, a gene co-expression network, where the interaction between two genes is quantified by the similarity of their expression pattern is developed. Pearson correlation coefficient is the most used representation in measuring similarity for gene-expression compared to other method[4]. However, the Pearson correlation coefficient has several drawbacks. First, it is valid only when there is a linear relationship. Second, it is sensitive to outliers. Third, it requires gene expression data to follow a normal distribution. Alternatively, mutual Information(MI) can measure non-linear relationship so is also used in gene relationship analyses[4].

After getting the quantitative representation of the gene-gene relationship, further analysis can be performed to further investigate the gene co-expression network. In traditional system biology methods, after calculating the Pearson correlation coefficient, the clustering algorithm is used to separating gene pairs into different groups called a module. The clustering method is an unsupervised method, which may lead to false-positive relationships.

Another problem lies in the acquisition of gene expression data. Traditional gene sequencing experiments detect gene expression levels in a large number of cells thus only provide an average expression level in a group. These average expression data may not reflect the real interaction between genes which leads to false-negative results. Recent single-cell gene sequencing technology[5] allows us to acquire gene expression data on a single cell and overcome the shortcoming of bulk gene sequencing.

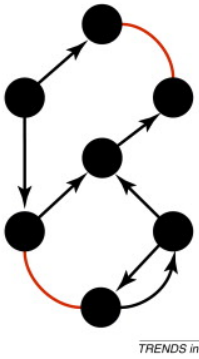


Fig. 1. An example of a network with nodes (black circles) and edges. [8]

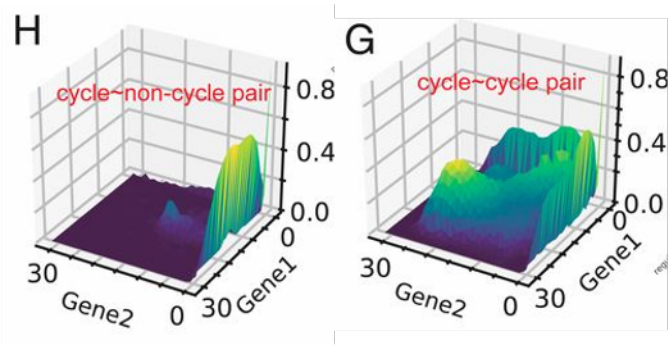


Fig. 2. Predicted expression pattern of a cell cycle cell cycle and a cell cycle non cell cycle gene pair[6]

B. Scientific Question

In gene expression analysis, a more common situation is that we know the phenotypes of the cells and want to know which genes contribute to these phenotypes, for example, diseases or biological processes. While in some other cases, like in ours, several genes related to disease or function are validated by former experiments. Then the question is if we can use these validated genes to find out novel biologically interesting genes that have been dismissed. The traditional gene-gene relationship inferring method using the Person correlation coefficient and clustering method has some drawbacks. Due to the unsupervised property of the clustering method and the limitation of the Pearson correlation coefficient, some important gene-phenotype relationships may be dismissed while some false positive relationships may arise.

Yuan and Bar-Joseph's convolutional neural networks-based method, CNNC[6], provides a solution to the problems and it is the reason why we are attracted by it and decide to apply their model to our biological question. With newly available single-cell expression data, Yuan and Bar-Joseph developed a to learn the relationship between two genes in single-cell expression data. In the article, the author collected genomics data of different mouse cell types from over 500 different scRNA-seq studies. The author also collected two sets of disease-related(asthma and chronic obstructive pulmonary disease, COPD, and on head and neck cancer, HNC) genes from "Malacards"[7]. By using the two sets of genes as labels, the author successfully predicted novel disease genes. In the top 10 predicted genes for asthma, one of them is recently determined to be a potential drug target for asthma therapy.

II. Models and Methods

A. Gene Labels and Data sources

We collected two set of labels. The first set is from Prof.Aye's lab. A total of 45 zebrafish genes differentially expressed between 4-Hydroxynonenal(HNE) treated group and control group are collected as HNE related genes. We collected 29 orthologous human and 25 orthologous mouse genes that present in each gene expression database. The second set is from Prof.Aye's published review[9] which summarised the HNE-sensitive protein. We collected 37 protein-coding human gene form the review. 31 of human genes are in our collected human gene expression database and 33 orthologous mouse genes are found in Yuan's mouse gene expression database .

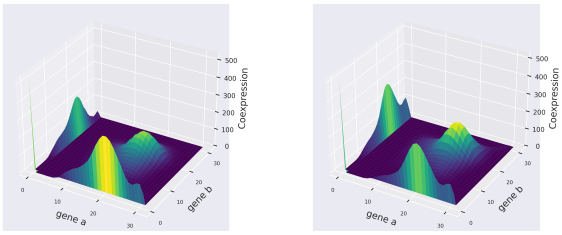
Our human single-cell expression data was obtained through R package "scRNAseq"[10] and mouse scRNA-seq dataset[11] used by Yuan.

B. CNNC model

Convolutional neural network for coexpression (CNNC) method is a supervised way to show the relationship between genes, using their coexpression data in a single cell. With tens of thousands of cell samples, each gene has a sequence of over 40,000 integers, indicating the counts of its occurrence in one observation of a single cell, and we normalized the data based on gene length and sequencing depth. For each pair of genes, we processed their expression as a 2D matrix (Normalized empirical probability function, NEPDF) which can be fed to convolutional neural networks (CNN) to reveal the relationship between them. The network was trained by different balanced training set to give a binary output, with 1 to show stronger correlation and 0 to show weaker.

To implement this model, we first generated gene pairs as our raw data. For example, we used 31 human genes from the Prof.Aye's lab which is known to be relevant as our "Known gene set", then randomly chose 31 unknown genes from the library as our "Unknown gene set". From "Known gene set" and "Unknown gene set", we chose 1/4 of them to be the "Test gene set", which can be divided into "Test-known gene set" (c) and "Test-unknown gene set" (d) containing 7 genes each. Meanwhile, we generated "Train-known gene set" (a) and "Train-unknown gene set" (b) with 24 elements each. To get NEPDFs as training set, we calculated 2D histogram for all gene pairs where gene1 is in (a) and gene2 is in $(a)+(b)$, thus getting $24 \times 24 \times 2$ gene pairs with label 1. To generate test gene pairs, we calculated NEPDFs for gene pairs where gene1 is from (a) and gene2 is from $(c)+(d)$ with label 0 by the assumption that the probability of finding similar genes is very low in such a huge library, thus got $24 \times 7 \times 2$ gene pairs. Then We take a closer look at the typical NEPDF of two genes.

Based on the Yuan's original model, our CNN includes an input layer, 6 convolutional layers, 3 maxpooling layers and a flatten layer, as well as the final "sigmoid" output layer. For



(a) Unrelated human gene NEPDF (b) Related human gene NEPDF

Fig. 3. Typical NEPDF

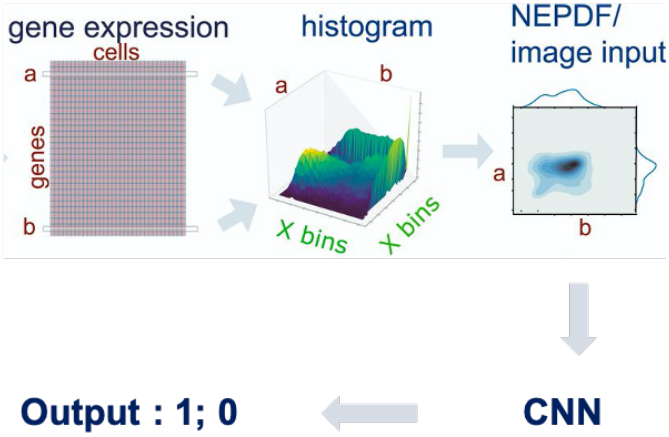


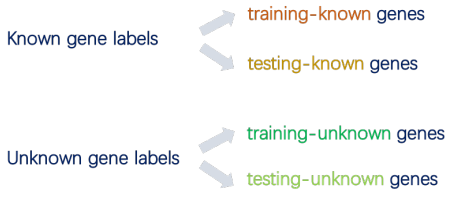
Fig. 4. Work flow of the model, figure editted from Yuan's article[6]

human and mouse, we have three batches of different labels, we used these data to test the performance of the model and the methods.

III. Results

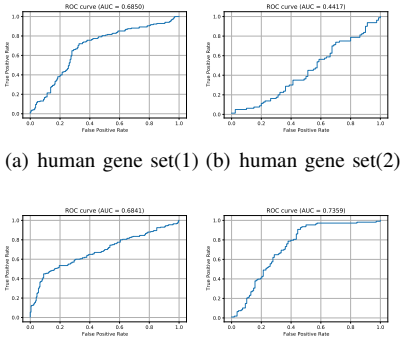
A. Baseline

Simply applying the model, the results are not satisfying. After considering the data size and the initial training performance, we decided to use the first group of human scRNA expression as the our baseline, and adjust our model.



Train sets	Gene A	training-known
	Gene B	training-known + training-unknown
Test sets	Gene A	training-known
	Gene B	testing-known + testing-unknown

Fig. 5. Gene pairs of train sets and test sets



(a) human gene set(1) (b) human gene set(2)

(c) mouse gene set(1) (d) mouse gene set(2)

Fig. 6. Baseline of CNNC applied on different dataset

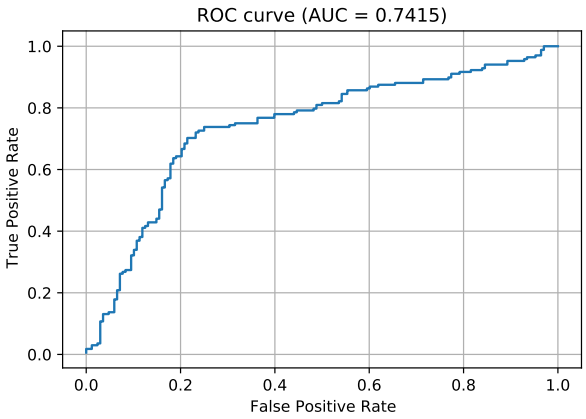


Fig. 7. ROC curve of the fine-tuned model

B. Fine-tune

One of our problems is the size of our data. Because our tasks are similar to some extent, and the dataset of Yuan's is relatively big, we use fine-tune, one kind of transfer learning to utilize Yuan's trained model. Since the task of the original model is a three classification task, we delete the last layer of the original model, add a Dense layer to change the output to our task and retrain the model. We also try to treat the original model as a feature extraction function and apply a two task classifier on the extracted features. This two methods achieve a similar result, which is a more satisfying result as shown in Fig. 7 and Fig. 8.

When the overfitting is severe, the performance on the training set can be amazingly good but the test accuracy could be embarrassing, especially for our small size. To avoid overfitting and distortion, we set the learning rate relatively low. We also explored the effect of the number of epoch, and the results showed epoch set to 200 is the best choice to achieve the trade-off between bias and variance.

C. Simplification

Since overfitting is quite severe due to the small dataset, and some hyperparameters need to be optimized, we applied different methods to find a better model.

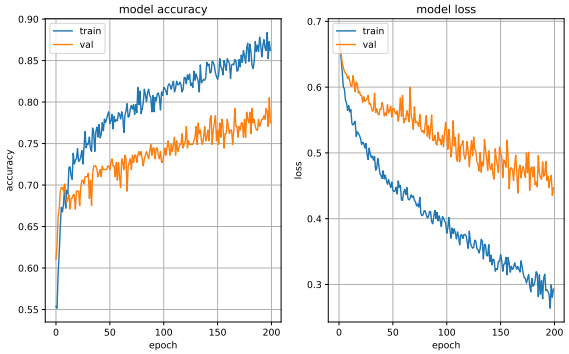


Fig. 8. Accuracy and loss of fine-tuned model

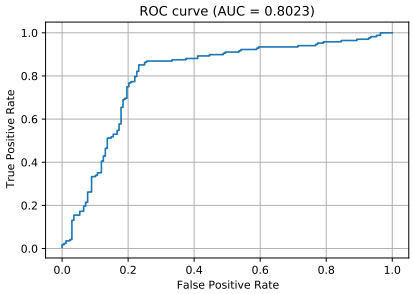


Fig. 9. ROC curve of modified method: 4 layer convolutional net with filter size 3x3, padding 0, ReLU activations, 2x2 max pooling. Learning rate 1e-3, batch size 32, 200 epochs

We first tried support vector machine (SVM) based on our known gene and unknown gene NEPDFs, the results in Fig. 11 showed high accuracy on training set and test set.

To achieve a better accuracy, we tried to simplify the cnnc model. We deleted 2 convolutional layers and 1 maxpooling layer, reduced the dropout probability, and changed batch size and epochs, to get a better performance. As shown in Fig. 9 and Fig. 10, the model reached a 0.79 accuracy and a 0.8 AUC score.

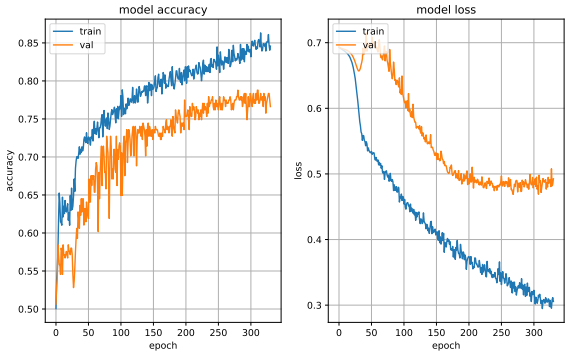


Fig. 10. Accuracy and loss of simplified model

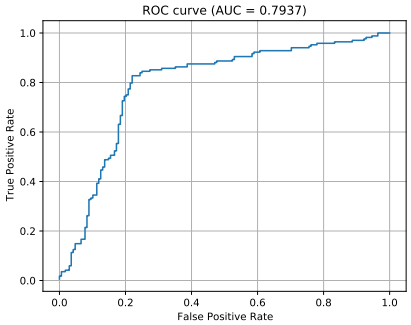


Fig. 11. ROC curve of SVM method [8]

D. Conclusion

We have shown that for two known genes, the model can tell whether they are related or not with an acceptable accuracy. Whereas for some unknown genes, the accuracy of the model is not significantly better than guessing. For further gene study in practice, we want to investigate whether a gene's NEPDF from unknown set with each gene in the known set can be used to judge whether the gene has the same effect as the gene we know.

IV. Discussion

Our team think one problem with our model is in the database we use. In authors' article, they collected genomics expression datasets of different mouse cell type from over 500 different scRNA-seq studies instead of genomics data set from a single study that focuses on one biological question, for example, asthma. For genes such as cell cycle genes that are crucial for maintaining basic cell survival, they may be more easily to respond to disturbances from the external environment than genes that are related by very specific biological activity. Thus in a combined database that consists of more than 500 different studies, different gene expression pattern can be shown between cell cycle related genes and non-related genes(figure 2). While in our case, from the NEPDF graph, we do not see an intuitive gene expression difference(figure 3), which could explain why our model can not reach a satisfying accuracy.

Choosing a gene expression database that is directly relevant to our investigated biological question may be a solution. But biological datasets are usually expensive and time-consuming. Single molecular sequencing technique is developed in recent years but it is still not a routine experimental procedure[12]. The expenses for single-cell library generation and sequencing are over 3500 dollars per experiment[13]. For diseases that are widely investigated around the world, open-source gene expression databases can be collected. While for cutting edge biological function/process that has not been investigated by many people, a gene expression database needs to be built from scratch.

Then there is another question: how many data is sufficient for CNNC model. Neural network model benefits from a large number of data. We are not sure if the CNNC model can perform well in single-cell gene expression datasets from only

several experiments. Of course one could collect gene expression datasets that are not directly relevant to the investigated biological process. But there is a trade-off. If these datasets actually do not contain much information of labelled genes, addition noises are added.

REFERENCES

[1] M. Cobb, "60 years ago, francis crick changed the logic of biology," *PLoS biology*, vol. 15, no. 9, p. e2003243, 2017.

[2] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature reviews genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[3] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *science*, vol. 302, no. 5643, pp. 249–255, 2003.

[4] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: mutual information, correlation, and model based indices," *BMC bioinformatics*, vol. 13, no. 1, p. 328, 2012.

[5] B. Hwang, J. H. Lee, and D. Bang, "Single-cell rna sequencing technologies and bioinformatics pipelines," *Experimental & molecular medicine*, vol. 50, no. 8, pp. 1–14, 2018.

[6] Y. Yuan and Z. Bar-Joseph, "Deep learning for inferring gene relationships from single-cell expression data," *Proceedings of the National Academy of Sciences*, vol. 116, no. 52, pp. 27 151–27 158, 2019.

[7] N. Rappaport, M. Twik, I. Plaschkes, R. Nudel, T. Iny Stein, J. Levitt, M. Gershoni, C. P. Morrey, M. Safran, and D. Lancet, "Malacards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search," *Nucleic acids research*, vol. 45, no. D1, pp. D877–D887, 2017.

[8] A. de la Fuente, "From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases," *Trends in genetics*, vol. 26, no. 7, pp. 326–333, 2010.

[9] S. Parvez, M. J. Long, J. R. Poganik, and Y. Aye, "Redox signaling by reactive electrophiles and oxidants," *Chemical reviews*, vol. 118, no. 18, pp. 8798–8888, 2018.

[10] D. Risso and M. Cole, *scRNAseq: Collection of Public Single-Cell RNA-Seq Datasets*, 2020, r package version 2.4.0.

[11] A. Alavi, M. Ruffalo, A. Parvangada, Z. Huang, and Z. Bar-Joseph, "A web server for comparative analysis of single-cell rna-seq data," *Nature communications*, vol. 9, no. 1, pp. 1–11, 2018.

[12] J. Eberwine, J.-Y. Sul, T. Bartfai, and J. Kim, "The promise of single-cell sequencing," *Nature methods*, vol. 11, no. 1, pp. 25–27, 2014.

[13] Y. J. Wang, J. Schug, J. Lin, Z. Wang, A. Kossenkov, K. H. Kaestner, H. Consortium *et al.*, "Comparative analysis of commercially available single-cell rna sequencing platforms for their performance in complex human tissues," *bioRxiv*, p. 541433, 2019.