A deep learning model in predicting gene-expression with histone modification

1. Background and objective

Whether a gene can be expressed at a higher level or a lower level is regulated through diverse mechanisms. Histone modification is one of these mechanisms. Histones interact with DNA strings to from bead'-like structures called nucleosomes. Histone modifications such as methylation could change the interactions with DNA resulting in changes in the spatial arrangement (e.g. exposure of promoter region) and sensitivity of DNA strand to gene regulation factors. Different from DNA changes, the histone modification does not alter DNA sequence and can be reversible, providing a potential way to treat cancer patients with 'epigenetic drugs'. Thus, have a deeper understanding of the gene regulation mechanism through histone modification is essential to our human health and pharmacology.

With the advancement of sequencing techniques, histone modification and gene expression profiles can be acquired relatively easier. However, the code of histone modifications in controlling gene expression has not been fully deciphered yet. Many computational models such as linear regression (Karlić et al., 2010), Support Vector Machines (Cheng et al., 2011), Bayesian networks (Cheng et al., 2011), Random Forests (Dong et al., 2012), rule-based learning models (Ho et al., 2015), as well as binning' approaches (Cheng et al., 2011) were used to address this problem. However, these methods either rely on multiple models to separate prediction and combinatorial analysis, or ignore the subtle differences among signal distributions of histone modification, or fail to analyze the neighbouring regions as a whole.

The recent development of the deep learning technique has gained its success in many prediction tasks such as image classification (Krizhevsky et al., 2012). When processing images, deep learning (such as CNNs) not only take local features into consideration, but also perceives global information as a whole to make a prediction. Such a feature of deep learning can also be very helpful in predicting gene expression with the information of histone modification through viewing the modification sites as a whole feature map rather than isolated information islands.

This practice will apply a CNN framework that could combine histone modification information from different regions and make a prediction on gene expression status.

2. Materials and methods

Gene expression status and histone modification sequences were acquired from Dr. Ahmed Ashraf. The dataset totally consists of 15845 genes and their corresponding histone modifications information. For each gene, the dataset consists of one gene expression status (0 or 1) and 100 neighbouring modification sites for each of the five histone proteins (H3K4me3, H3K4me1, H3K36me3, H3K9me3 and H3K27me3 respectively). In other words, the model input is a 15845*100*5 matrix, while the output is 0 or 1. This dataset was used to train a deep learning model.

The deep learning network was constructed on the Pytorch framework and consists of five modules. The models are data loading and separation module, CNN network module, training workflow module, accuracy evaluation module, as well as executive module respectively.

CNN network used in this practice is roughly the same as the one used in the provided reference paper (Singh et al., 2016). Specifically, the first layer is a CNN layer. In this

layer, twenty 5*5 kernels were used. After the CNN layer, a batch normalization layer was added to accelerate network converging, followed by a ReLu activation layer. As the output shape of this layer is not intuitive for the following max pooling, a reshape layer was added. Then the output is processed further by a max-pooling layer to learn invariant features. The rest layers include a flatten layer, a dropout layer to increase generalization, a deep multilayer perceptron architecture that consists of two hidden layers (with a sigmoid layer at the end of each hidden layer). The output of this network is a probability prediction matrix (1*2) associated with a histone sample.

The network was trained for 50 epochs. At the beginning of each epoch, the input dimension is 80 (batch size)*1 (number of channel)*100 (number of modification sites) * 5 (number of histone proteins). After the convolution layer [with 20 kernels each of which has a kernel size of (5, 5) and a stride of (1, 1)], the dimension was changed to 80 (batch size)*20 (number of kernels)* 96 [downward slides with the kernel of (5, 5)] * 1 [horizontal slide with the kernel of (5, 5)]. To be less abstract when doing max pooling, the matrix was reshaped to 80*1*20*96. After the max pooling operation [with a kernel size of (1, 3) and a stride of (1, 3)], a matrix with a dimension of 80* 1*20*32 was produced. The matrix was flattened and fed into two consecutive linear layers with an output shape of 80*120 and 80*2 respectively. The softmax function outputs the possibilities corresponding to the two classes. The loss between the prediction output and the ground truth was calculated with CrossEntropyLoss function, while the Adam optimizer was used to adjust the network parameters. The parameter-updated network will process the next batch until all the ~11091 (0.7 * 15845) training samples are processed. At the end of each epoch, the network will print the recorded accuracy on the training dataset and the evaluated accuracy on the testing dataset. Then the next epoch begins.

3. Results and discussion

The final accuracy of this network is ~ 87% for the training dataset, while it is 86% for the testing dataset. This network converged effectively and reached ~85% accuracy at the beginning epochs. When training the network, different settings such as batch size, learning rate, dropout probability, filter kernel size, with and without of batch normalization layer, were used. The results did not show an apparent difference among these settings. It is worth noting that the accuracy of this network is not as high as DeepChrome, although the basic network architectures are very similar. The DeepChrome reached a higher accuracy of 94% on some of the 56 cell types. The difference in network performance can be explained below. First, DeepChrome used a different dataset, and it is hard to directly compare the practiced network and the DeepChrome. Second, the accuracy of DeepChrome does not have a consistent performance when predicting 56 cell types, and the lowest accuracy was 66%.

To improve the network performance, there are three possible directions. First, the training dataset needs to be generated and validated with extra attention. For sequencing experiments, the final results can be affected by many factors such as protein quality, reagent brand, instrument condition, operator skills, as well as downstream analyzing software and algorithms. Thus, protein quality, experiment control, as well as experiment replication are critical factors when preparing the histone modification dataset. Second, extra CNN layers can be added to the network to improve the capacity of the network.

However, it needs to be very careful when applying a complex network on a relatively simple dataset, as overfitting can be a major problem by doing so. Third, the hard-negative training dataset can be collected and used to re-train the network. This method was reported by other researchers (Fuentes et al., 2018) and still waiting for further validation in this practice.

4. Conclusion

The trained CNN network could perform the classification task with a good accuracy while further improvement can be made concerning the dataset quality, network structure, and re-training with hard-negative data.

Reference

Cheng, C., Yan, K.-K., Yip, K.Y., Rozowsky, J., Alexander, R., Shou, C., and Gerstein, M. (2011). A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome biology* 12**,** 1-18.

Dong, X., Greven, M.C., Kundaje, A., Djebali, S., Brown, J.B., Cheng, C., Gingeras, T.R., Gerstein, M., Guigó, R., and Birney, E. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome biology* 13**,** 1-17.

Fuentes, A.F., Yoon, S., Lee, J., and Park, D.S. (2018). High-performance deep neural network-based tomato plant diseases and pests diagnosis system with refinement filter bank. *Frontiers in plant science* 9**,** 1162.

Ho, B.H., Hassen, R.M.K., and Le, N.T. (2015). Some Current Advanced Researches on Information and Computer Science in Vietnam.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences* 107**,** 2926-2931.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Advances in neural information processing systems.

Singh, R., Lanchantin, J., Robins, G., and Qi, Y. (2016). DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 32**,** i639-i648.