Notes for network meeting - First week in Sep

Jiaxin Hu

1 Sep 2 Reading group

In the last semester, our primary goal is applying some statistical methods to see whether there are significant differences in the gene co-expression network under different environments or other covariates.

In this semester, our goal is to study the association between the natural selection speed and the gene expression or co-expression. The selection speed of genes may be a function of the network architecture; e.g., the “hub” gene in the bow-tie structure is supposed to have a faster selection speed. Current data includes: tissue level and single cell level RNA sequence data, the selection speed of the genes that are generated by allele frequency data.

One thing need to notice is that: the hypothesis of the selection speed and gene network architecture is based on the gene regulatory network. It is unknown whether we can use correlation network (e.g. WCGNA package) to estimate the regulatory network properly.

2 Sep 7 RNA group

Dan wants to involve the results about the selection strength and the gene centrality in their paper. There are previous papers discuss the biological relation between gene or cell functions to the selection, and thus we also want to find the association between the cell types and the selections, where cell types are obtained by expression data clustering.

The key question here is how to choose a proper network character to describe the centrality of the genes. We may then do a regression where the gene network/gene centrality serve as the response and the selection strength serves a the predictor.

Wan (from Northeastern) said they have the curvature results for all the genes, which is calculated by the tissue level RNA sequence data. A smaller curvature indicates a larger centrality of the corresponding gene. Dan’s students would also like to share the selection results they have.

3 Sep 9 Reading group

This reading group discusses the Bayesian method to select important covariates to gene expression and the gene co-expression networks simultaneously. The simultaneous estimation is achieved by the prior distribution for the latent variable: $Z_i \sim \mathcal{N}(\beta_0 + \boldsymbol{M}_i\beta, \Omega^{-1})$, where $Z_i, \boldsymbol{M}_i$ is the latent variable and covariates for the i-th sample.

Three interesting questions arise during the meeting: (1) is there any power analysis for the estimation procedure? No uncertainty measure associated with the network we get, and more discussion for the experimental design (e.g. how many sample size we need) is helpful. (2) Can we use other machine learning methods like random forest to select the important covariates? Random forest can capture more non-linear associations than linear regression. (3) The critical problem for the proposed method is that we do not directly involve the covariates in the network estimation. As the prior formula shows, the covariates are not involved in the likelihood estimate of Ω. Our previous questions is how Ω changes along with different M. Therefore, this method may not be the best fit for our goal.

References