How To Use GeneNT to Reconstruct Coexpression Network with Controlled Biological Significance and Statistical Significance

Dongxiao Zhu

April 2, 2007

Many biological functions are executed as a module of coexpressed genes which can be conveniently viewed as a coexpression network. Genes are network vertices and significant pairwise co-expressions are network edges. The core of coexpression network recontruction problem is how to determine if the observed co-expressions are biological and/or statsitically significant. The earliest approach is to use a conservative correlation cut-off, such as .6. Coexpression are biologically significant if the observed (absolute value) correlation is above this cutoff. Later approach attempted to jointly consider biological and statsitical significance. For example, one approach first tests whether the population correlation parameter is different from 0 or not at some significance level. Assuming ρ is the true correlation, the following pair of hypotheses were tested:

$$|\rho| = 0 \ \text{ versus } \ |\rho|! = 0. \tag{1}$$

If rejected an correlation cutoff was applied to those significance correlated gene pairs. This approach does NOT "controls" either biological or statistical significance. The P-value obtained is tied to the hypothesis whether true correlation is 0, which is often not of our interest. Indeed, we need to test the following pair of hypothses:

$$|\rho| \leq C \ \text{ versus } \ |\rho| > C, \tag{2}$$

where C is the correlation cutoff. The P-value obtained from this pair of hypothesis tests is more related to our interest, i.e. whether the true correlation greater than the correlation cutoff or not. We therefore, claim that the hypothesis test 2 controls biological significance and statsitical significance simultaneously.

The detailed test procedure was reported in [Zhu *et al.*, 2005a]. Very briefly, it achieves goal using a two-stage procedure. The stage I tests whether the correlation is different from 0 or not 1, the stage II test is for those correlation parameters that are significantly different from 0 at certain level in stage I test, construct simultanes confidence intervals for these parameters, and use the upper or lower bounds of those parameters to screen significant gene pairs. The reason why we can not test 1 is because the null distribution of the ρ has been intractable so far.

Now let's assume the gene coexpression network has been properly reconstrcuted with controlled biological and statistical significance. Since the network is typically very sparse, it imposes so-called network constraint for many types

of multivariate data analysis. We discuss some impact of network constraint on the popular hierarchical clustering methods. There are two factors that will significantly affect the performance of hierarchical gene clustering: how to measure the pairwise distance matrix between genes and how to measure distance between clusters. For detailed description of avaialble choices for the two factors, please refer to my recent book chapter [Zhu *et al.*, 2007]. The clustering method implemented in this package calculate a "hybrid" distance matrix consisting of both direct distance and shortest path distance. According to the network constraint, if a pair of genes are directly connected their distance is defined as 1-correlation; if a pair of genes are not directly connected their distance is calculated as the shortest-path connecting them [Zhu *et al.*, 2005b]. In the following example, we will do a series of data analysis starting from network reconstruction all the way to network constrained clustering.

```
> library(GeneNT)
> data(dat)
> g1 <- corfdrci(0.2, 0.5, "BY")

The screened pairs are now in your working directory.

> pG1 <- g1$pG1
> pG2 <- g1$pG2
```

The above code implements the two-stage screening procedure based on Pearson correlation coefficient. Warning: time complexity of this function is n^2, therefore, if you have a few hundreds genes or more, be prepared that the function runs a few hours to a few days.

```
> g2 <- kendallfdrci(0.2, 0.5, "BY")

The screened pairs are now in your working directory.

> kG1 <- g2$kG1
> kG2 <- g2$kG2
```

The above code implements the two-stage screening procedure based on Kendall correlation coefficient. Warning: time complexity of this function is n^2, therefore, if you have a few hundreds genes or more, be prepared that the function runs a few hours to a few days.

```
> getBM(pG2, kG2)
```

The above code generate a compatiable format for the sfotware Pajek to visualize data

```
> g <- ncclust(6, pG2, kG2)

The Network constrained distance matrix and dendrogram labels have been written to the def
```

The above code implements network constrained clustering with the scaling parameter set to 6.

If you prefer to read the above descriptions in a more rigirous way.

References

[Zhu *et al.*, 2007] Zhu, D., Dequeant, M.L. and Hua, L. (2007) Comparative Analysis of Clustering Methods for Microarray Data. To appear.

[Zhu *et al.*, 2005a] Zhu, D., Hero, A.O., Qin, Z.S and Swaroop, A. (2005) High throughput screening of co-expressed gene pairs with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS). *J. Comput. Biol.*, **12**(7), 1027-1043.

[Zhu *et al.*, 2005b] Zhu, D., Hero, A.O., Cheng, H., Khanna, R. and Swaroop, A. (2005) Network constrined clustering for gene microarray data. *Bioinformatics*, **21**(21), 4014-4021.