Gene Networks Based on the Graphical Gaussian Model

Shisong Ma*

Department of Plant Biology & Genome Center, University of California, Davis, California, USA
*For correspondence: sma@ucdavis.edu

[Abstract] This protocol describes how to build a gene network based on the graphical Gaussian model (GGM) from large scale microarray data. GGM uses partial correlation coefficient (pcor) to infer co-expression relationship between genes. Compared to the traditional Pearson' correlation coefficient, partial correlation is a better measurement of direct dependency between genes. However, to calculate pcor requires a large number of observations (microarray slides) greatly exceeding the number of variables (genes). This protocol uses a regularized method to circumvent this obstacle, and is capable of building a network for ~20,000 genes from ~2,000 microarray slides. For more details, see Ma *et al.* (2007). For help regarding the script, please contact the author.

Data and Software

1. Data

   Large-scale microarray data:

   The microarray data should be derived from the same platform, preferably from Affymetrix slides. Some good examples are: Affymetrix *Arabidopsis* ATH1 Genome Array, Affymetrix Human Genome U133 Plus 2.0 Array, and Affymetrix Mouse Genome 430 2.0 Array. A recommended place to search for this type of data is at the gene expression omnibus from NCBI (http://www.ncbi.nlm.nih.gov/geo/). The number of slides should be larger than 1,000.

2. Software
   a. R (http://www.r-project.org/)
   b. The GeneNet package for R: (http://www.uni-leipzig.de/~strimmer/lab/software/genenet/index.html)
   c. Cytoscape (http://www.cytoscape.org/)
   d. Perl and C++ software environment

Equipment

1. Personal computer: Intel Core2 E6420 processor (or similar processing capability)

Procedure

A. Preparation of the microarray data

1. Download the microarray data from your favorite database, and format it into a single table of expression intensities, with every row representing a gene and every column representing a microarray experiment. A good example can be found here for *Arabidopsis* transcriptomes: http://affy.*arabidopsis*.info/narrays/help/usefulfiles.html. You can use the file titled super bulk gene download.
2. Remove any columns (experiments) containing large number of 'null' measurements, and then do the same for any genes containing 'null' measurements.
3. Normalize the expression intensities between experiments using the quantile normalization method.

B. Random sampling and partial correlation calculation

1. Randomly pick 2,000 genes from the large expression table and make a small expression table for these 2,000 genes. A Perl script can be written to do this step.
2. Using the GeneNet package to calculate partial correlation between these 2,000 randomly selected genes. The GeneNet package should be lauched within the R environment, and the specific function to be used is 'ggm.estimate.pcor' with the default settings.
3. Save the resulting partial correlation matrix, together with the gene ids for the 2,000 genes.
4. Repeat the step from 1 to 3 at least 1,999 times. The more the better. After these calculations, most of the gene pairs should be sampled >10 times, each time with a calculated pcor.
5. Determine the final pcor values for every gene pair, so that pcor value with the smallest absolute values will be kept. This should be done via consolidating the resulted pcor matrix. This should be done with a C++ script.

C. Network building and analysis

1. To test the significance of the resulted pcors, the function 'ggm.test.edges' in GeneNet can be used. From all the pcors, ~2,000,000 can be randomly selected and fed into the function, so that a pValue for significance can be calculated.

2. Depending on the pValue, a cutoff for the pcors can be set. A good estimation would be 0.1, 0.08, and 0.05. Any pcor with absolute value larger than the cutoffs can be retained.
3. A Pearson' correlation coefficient filter should be applied. Gene pairs with Pearson' correlation coefficient value between -0.3 and 0.3 should be removed.
4. After the pcor selection and Pearson correlation coefficient filters, the remaining gene pairs are said to have interaction between each other, and can be used to build a gene network using Cytoscape software. The network analysis can be done with the Cytoscape software itself.

Acknowledgments

This protocol was developed by the author in Hans Bohnert's lab, Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. The work was supported by grants from the National Science Foundation Plant Genome Project (DBI-0223905) and University of Illinois at Urbana-Champaign institutional grants.

References

1. Ma, S., Gong, Q. and Bohnert, H. J. (2007). An Arabidopsis gene network based on the graphical Gaussian model. *Genome Res* 17(11): 1614-1625.
2. Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol* 4: Article32.