2017年华中师范大学数学建模竞赛校赛B题

基于基因表达数据推断基因调控网络是生物医疗大数据研究中的一个重要挑战。基因表达数据反映的是基因转录产物mRNA在细胞中的丰度。基因表达数据通常可用一个$n \times p$的矩阵X来表示，其中n是样本个数，p是基因个数，矩阵X的第(i,j)元x_{ij}表示第i个样本的第j个基因的表达情况。基因调控网络是指细胞内基因和基因之间的相互作用关系所形成的网络。基因调控网络通常可用一个$p \times p$的矩阵A来表示。矩阵A的第(j,k) 元a_{jk}表示第j个基因与第k个基因是否相互作用。若$a_{jk}=1$，则第j个基因和第k个基因之间有相互作用关系；若$a_{jk}=0$，则第j 个基因和第k个基因之间没有相互作用关系。由于生物技术的限制，基因表达数据可以高通量生物实验获取，但基因调控网络通常不容易通过高通量实验方法获取。因此研究计算方法从基因表达数据X推断基因调控网络A是一个重要的课题。

早期，科学家先通过相关系数（如Pearson相关系数、Spearman相关系数、互信息等）计算基因之间的相关性，然后通过阈值方法（保留大的相关性，过滤掉小的相关性）构建基因调控网络。基因之间的调控关系具有传递性。例如，若基因j调控基因k，基因k调控基因l，那么基因l的表达就会与基因i 的表达相关，若基因j没有直接调控基因l，而是通过调控基因k间接影响基因l的表达，我们则称基因j与基因k、基因k与基因l之间的关系为直接相互作用关系，称基因j与基因l之间的关系为间接相互作用关系。又如，若基因j同时调控基因k和基因l，而基因k和基因l之间没有相互调控关系，但由于基因k和基因l同时被基因j调控，它们的基因表达也有较强的相关性，此时我们称基因j与基因k、基因j与基因l之间的关系为直接关系，基因k与基因l之间的关系为间接关系。先前由相关系数方法得到的网络既包含直接关系又包含间接关系，因此不能准确地刻画基因之间的直接相互关系。基于该情况，解决以下问题：

1) 提出新的能够准确刻画基因之间直接相互作用关系的计算方法。并将其用于给定的基因表达数据（gene_expression.csv）构建基因调控网络，将结果以邻接矩阵的形式（文件名：Q1_result.csv）给出。
2) 通过先前生物医学实验，科学家已经收集了一部分基因之间的相互作用关系数据（注意该数据仅包含细胞内的一部分基因相互作用关系，而不是全部相互作用关系）。这些已知的相互作用数据能够对推断细胞内的整体基因相互作用网络提供重要信息。在问题1）的基础上，提出能够整合基因表达数据和先验相互作用关系数据的基因网络推断算法，使其能够更加准确地揭示基因之间的直接相互作用关系。将提出的方法用于给定的基因表达数据（gene_expression.csv）和先验网络数据（prior_networ.csv）用于构建基因调控网络，将结果以邻接矩阵的形式（文件名：Q2_result.csv）给出。

参考资料：

1) Barzel B, Barabasi A. Network link prediction by global silencing of indirect correlations[J]. Nature Biotechnology, 2013, 31(8): 720-725.
2) Feizi S, Marbach D, Medard M, et al. Network deconvolution as a general method to distinguish direct dependencies in networks.[J]. Nature Biotechnology, 2013, 31(8): 726-733
3) Friedman J H, Hastie T, Tibshirani R, et al. Sparse inverse covariance estimation with the graphical lasso[J]. Biostatistics, 2008, 9(3): 432-441.
4) Marbach D, Costello J C, Kuffner R, et al. Wisdom of crowds for robust gene network inference[J]. Nature Methods, 2012, 9(8): 796-804.
5) Zhao J, Zhou Y, Zhang X, et al. Part mutual information for quantifying direct associations in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2016, 113(18): 5130-5135.
6) http://blog.sciencenet.cn/blog-404304-770977.html