Centers of transcription factor correlation networks

Bioinformatics lab for module on graphs and biological networks*
Room 109, Härtelstraße 16-18, D-04107 Leipzig, Germany
July 2-13, 2012

I. SYNOPSIS

This practical course aims at the construction and comparative analysis of networks of transcription factors and regulated genes. Construction of networks is based on the correlation between observed expression levels. Comparisons of networks from different species (human vs. chimpanzee) is made a) in terms of central nodes according to various centrality measures and b) direct subtraction of networks.

II. DATA FILES

In each of the four data files, a row consists of the following pieces of information, in this order, separated by TAB characters.

1. line number
2. node i (identified by name)
3. node j (identified by name)
4. P-value p_{ij}
5. correlation ρ_{ij}

Thus each row defines a potential edge with a weight ρ_{ij}. Note that there may be two distinct rows for node pairs (i, j) and (j, i) while the weights are actually symmetric, $\rho_{ij} = \rho_{ji}$.

The first two files, `TFcvscCORTab.txt` and `TFhvshCORTab.txt`, provide information for each pair in a set $n = |V| = 704$ nodes. The nodes are transcription factors in chimpanzee and human (one file for each species). In the other two files, `ChimpDiffTFvsG.txt` and `HumDiffTFvsG.txt`, each node $i \in V$ is a transcription factor (node naming identical to the previous files) and $j \in W$ is a gene. The set of genes W is disjoint from V.

III. SOME MORE BACKGROUND ON THE DATA

The provided correlation data are based on observed expression levels of transcription factors (TFs) and genes in B cells of human and chimpanzee [1, 2]. They reflect the results of $\mu = 12$ experiments in each of the species. In an experiment, the expression levels of all nodes (=TFs and genes) are recorded simultaneously by a microarray [3, 4]. By $x_i^{(k)}$ we denote the expression level of node i in the k-th experiment. The corresponding rank vector is

$$y_i^{(s)} = |\{r \in \{1, \dots \mu\} : x_i^{(s)} \geq x_i^{(r)}\}| \tag{1}$$

The rank order correlation (*Spearman's rho*) between nodes i and j is then defined as the Pearson correlation coefficient

$$\rho_{ij} = \mathrm{corr}(y_i, y_j) \tag{2}$$

between the rank vectors. The P-value is the probability of obtaining a correlation at least as large in magnitude when randomly reshuffling expression levels across experiments.

---

*Electronic address: `klemm@bioinf.uni-leipzig.de`

IV. TASKS

This list of tasks is a suggestion for extracting potentially interesting information from the data and the resulting networks. Tasks 1-7 should be completed during the first week. In the second week, we shall elaborate on alternative approaches for comparing the graphs for different species.

1. Reading and filtering the data. Read the data from the file into memory and convert them into a graph structure. Requires some planning ahead: choose your internal graph representation such that the upcoming tasks can be performed efficiently.

   We only want to keep significantly correlated pairs of nodes as edges. Thus it makes sense to define the edge set by

   $$\{i, j\} \in E \Leftrightarrow p_{ij} < \theta \tag{3}$$

   with a threshold $\theta \in]0, 1[$. Typically, one uses $\theta = 0.05$ for significance discrimination. How does the total number of edges $|E|$ depend on θ? What is the distribution of edge weights ρ_{ij} for $\{i, j\} \in E$ depending on θ? Decide if you are going to work with weighted or unweighted graphs.

2. Degree and strength distributions. Find the distributions of degree and – if you work with weighted graphs – of strength. The strength of a node is the sum of weights of its incident edges.

3. Cores and shell index. A k-core of a graph is a maximal induced connected subgraph of minimal degree k [5, 6]. It can be defined analogously for graphs with non-negative weights – how?. Find all k-cores of your graph. In particular, you obtain the connected components of the graph itself as the k-cores with $k = 0$.

4. Eigenvector centrality. The adjacency matrix (or weight matrix) of each connected component of the graph has a non-degenerate maximum eigenvalue. Find the unique corresponding eigenvector with non-negative entries [7, 8].

5. Distance matrix. Calculate the distance d_{ij} between all pairs of nodes i, j where there is a path between i and j. As a side product, you also identify the connected components of the graph. (Again, if you work with weights, spend some thought on how to include them in the distances in a sensible manner).

6. Distance-based centralities. Based on the distance matrix, compute the excentricity, status and centroid value of each node [9].

7. Comparison of centers and species. Prepare a comprehensive overview of the most central nodes in the graphs of the two species for the various centrality measures

8. Direct graph comparison. Develop a method to directly compare the two graphs. E.g. prepare a difference graph between the two.

---

[1] E. Choy, R. Yelensky, S. Bonakdar, R. M. Plenge, R. Saxena, P. L. De Jager, S. Y. Shaw, C. S. Wolfish, J. M. Slavik, C. Cotsapas, et al., PLoS Genet **4**, e1000287 (2008).
[2] K. Nowick, T. Gernat, E. Almaas, and L. Stubbs, Proc. Natl. Acad. Sci. USA (2009).
[3] A. Schulze and J. Downward, Nat Cell Biol **3**, E190 (2001).
[4] D. D. Dalma-Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada, in *DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols*, edited by A. Kimmel and B. Oliver (Academic Press, 2006), vol. 410 of *Methods in Enzymology*, pp. 3 – 28.
[5] I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, cs.NI/0504107; cs.NI/0511007. (2005).
[6] M. Kitsak, L. K. Gallos, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, Nature Physics **6**, 888 (2010).
[7] P. Bonacich, Social Networks **29**, 555 (2007).
[8] K. Klemm, M. A. Serrano, V. M. Eguiluz, and M. San Miguel, Sci. Rep. **2**, 292 (2012).
[9] S. Wuchty and P. F. Stadler, Journal of Theoretical Biology **223**, 45 (2003).