**CCor**: a whole genome network-based similarity measure between two genes

Yiming Hu and Hongyu Zhao

Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, U.S.A.

**Motivation**

- **Purpose**: understand the interactions and functional relationships between genes and infer network modules.
- **Data**: gene expression profiles – microarray data or RNA-seq.
- **How to choose a similarity measure for clustering?**
- **Two types of similarity measure**: Pairwise: e.g. correlation coefficients, mutual information; Network-based, e.g. topological overlap measure (TOM).
- **Pairwise-ignore biological structure between two genes and other genes in the network; Network-based: lack theoretical analysis of the measure**
- **Objective**: develop a network-based measure and compare the distribution of within-module similarity and between-module similarity

**Methods**

- **Data**: expression level of \( p \) genes of \( n \) individuals, \( X_{n \times p} \), diagonal elements of \( W \) are \( 1 \)
  
  \[
  r_{ij} = \frac{1}{n} \sum X_{ij} X_{ij}
  \]

  \[
  v_i = (v_{ij}, ..., v_{i(j-1)}, v_{i(j+1)}, ..., v_{ijp})
  \]

  \[
  CCor_{ij} = \hat{c}(v_i, v_j)
  \]

  \[
  mCCor_{ij} = \hat{c}(v_i, v_j, w, \lambda)
  \]

- **Theoretical Results**

  Following the assumptions in Methods

  \[
  \begin{align*}
  \Sigma_i \Sigma_j & \equiv \text{cor}(v_i, v_j) \\
  \mu & \equiv \frac{1}{n} \sum v_i \\
  A & \equiv I_K - \frac{1}{n} \mathbf{1}_K \mathbf{1}_K^T \\
  CCores & \equiv \frac{1}{Cov(v_i - \mu, v_j - \mu)} \\
  \end{align*}
  \]

  Denote

  \[
  \alpha \equiv 1 - 2c_{min}(\mathbf{1}_K, \mathbf{1}_K) - 2c_{min}(\mathbf{1}_K, \mathbf{1}_K) - 2c_{min}(\mathbf{1}_K, \mathbf{1}_K)
  \]

  with probability no less than

  \[
  p_1 = p_2 = 100, n = 100
  \]

  **Compare the difference between within-module (CCor1) and between-module (CCor2, p=1) similarities**

**Real Data Analysis**

- **Hierarchical clustering with 1-similarity as the distance measure; a dynamic tree cutting method**
- **Gene Ontology (GO) enrichment analysis of resulting modules**
- **Record 5 most significant GO enrichment p-values for each detected module and pool the p-values together.**

**Simulation**

- Use similarity measure to classify gene pairs into within-module and between-module and calculate AUC

\[
W = \left[ \begin{array}{ccc}
-0.5 & 0 & 0.2 \\
0.5 & 0 & 0.2 \\
0 & 0.5 & 0.2 \\
0 & 0 & -0.5 \\
0 & 0 & 0.5 \\
0 & 0 & 0.5 \\
\end{array} \right] \]

**Table 1**: Comparison of five methods using AUC when sample size and noise level vary

**Conclusion**

Compared to Pearson correlation, we have found that CCor could amplify the difference between gene pairs from the same module and those from different modules, and therefore, better infer gene modules. In comparison with other methods, CCor yields the best performance in detecting gene pairs in the same module in simulation. And in real data applications where module sizes are usually unbalanced, mCCor shows its power and performs the best across three datasets in terms of finding more biologically meaningful results. Through this study, we show that network-based similarity measure could potentially bring more biological insights.

**References**


**Acknowledgement**

The authors would like to thank Jina Li, Yu Lu, Qiongshi Lu, Tianqi Liu and Tao Wang for their insightful suggestions.