ECRONICON OPEN ACCESS

EC MICROBIOLOGY
Research Article

Gene Co-Expressions Cannot Predict Protein-Protein Interactions in *Escherichia coli*

Marcus TE Chua, Andrei BG Dumanglas and Maurice HT Ling*

School of Applied Sciences, Temasek Polytechnic, Singapore

***Corresponding Author:** Maurice HT Ling, School of Applied Sciences, Temasek Polytechnic, Singapore.

Received: January 27, 2022; **Published:** February 28, 2022

Abstract

Gene co-expression is the correlation of gene expressions across multiple samples or conditions. Significant gene co-expressions have been used to construct gene co-expression networks and used to elucidate biological information. However, the suitability of gene co-expressions in predicting protein-protein interaction is not clear. In this study, ten gene co-expression measures were evaluated for its suitability in predicting PPIs in *Escherichia coli*. Our results show poor precision (precision ≤ 0.00188). This suggests that gene co-expression alone is not likely to be suitable to predict protein-protein interactions.

Keywords: *Gene Co-Expression; Protein-Protein Interaction; Bray and Curtis Coefficient; Cosine Coefficient; Canberra Distance; Euclidean Distance; Kendall's Tau; Manhattan Distance; Pearson's Correlation; Point Biserial Correlation; Spearman's Correlation; Tanimoto Coefficient*

Introduction

Gene co-expression network (GCN) refers to a network of genes as nodes and the presence of significant co-expression between pairs of nodes as edges [1], which has been shown be useful in elucidating important biological information [2]. For example, Reverter., *et al.* [3] used GCN to identify transcriptional regulation of bovine skeletal muscles. Ling [4] examined the overlap between GCN and gene co-occurrence in literature to identify potential hypotheses for future research. van Dam., *et al*. [5] used GCN to examine gene-disease association. Recently, Sharma., *et al*. [6] used GCN to identify key genes in active metabolite biosynthesis of a medicinal plant and Fajardo and Quecini [7] used GCN to examine expressional conservation between wild and cultivated grapes.

Pearson's correlation is one of the most common measures of gene co-expression, which has been used in many recent studies [8-11]. Other measures include Spearman's correlation [12] and weighted gene co-expression network analysis [13]. However, there are two main open questions. Firstly, it is not clear which measure is most suitable. Chay., *et al.* [14] demonstrated that different measures can impact on estimated genetic distance between organisms using DNA fingerprinting. Secondly, it is not clear whether GCN or which measure by which the resultant GCN is constructed is representative of protein-protein interaction (PPI) network derived from experimental data despite Piya., *et al*. [15] suggested substantial overlap between GCN and PPI network after auxin treatment in Arabidopsis.

Recently, Rajagopala., *et al.* [16] published a set of literature supported or experimentally verified PPIs in *Escherichia coli*. In this study, we examine several measures of gene co-expression on its ability in predicting PPIs in *E. coli*. Our results suggest that although Pearson's correlation is the best performing measure, its ability to predict PPIs is low. This suggests that PPIs cannot be predicted using gene co-expression.

Materials and Methods

Data Set: Gene expression data set from Faith., *et al.* [17], containing 266 samples, were downloaded from NCBI Gene Expression Omnibus [18] as GSE6836. The CEL files were normalized and the gene expression data were exported using affy package [19]. The probe set IDs were converted to Locus Tag IDs and only readings with Locus Tag IDs were used. Binary protein-protein interaction data set from Rajagopala., *et al.* [16] were used. Both microarray probe set IDs and protein IDs [Supplementary Table S5 of Rajagopala., *et al.* [16] were converted to Locus Tag IDs.

Co-expression measures: Nosrmalized gene expressions from GSE6836 were used to generate absolute co-expression measures using coexp method in SeqProperties [20]. Ten co-expression measures will be used; namely, (i) Bray and Curtis coefficient [21,22], (ii) Cosine coefficient [21,23], (iii) Canberra distance [21,24], (iv) Euclidean distance [21,25], (v) Kendall's tau [26], (vi) Manhattan distance [21,27], (vii) Pearson's correlation (28,29), (vii) Point biserial correlation [30], (ix) Spearman's correlation [29], and (x) Tanimoto coefficient [21,31]. Each absolute co-expression was statistically tested for significance using randomization test [32]. The mean co-expression of 1000 random gene expression pairs were generated using coexp_rand method in SeqProperties [20]. Thirty replicates were performed to provide the grand mean and standard errors of randomized co-expressions, where standard deviation of randomized co-expressions can be estimated as the product of the mean standard errors and square root of 1000. Significant gene co-expressions were filtered from gene co-expressions in three ways. Firstly, absolute gene co-expressions higher than 1.645 times of standard deviation above the mean of randomized co-expressions were considered significant. Secondly, top 5 percentile of gene co-expressions were considered significant. Lastly, top 5 percentile of absolute gene co-expressions were considered significant.

Benchmarking GCN to PPI; Significant gene co-expressions were tabulated against protein-protein interaction from Rajagopala., *et al.* [16] as truth using Locus Tag IDs where true positive, false positive, and false negative would be tabulated; which would be used to calculate precision, recall, and F1-score.

Results and Discussion

Normalized gene expression data set from Faith., *et al*. [17]) consists of 4345 genes with Locus Tag IDs, which resulted in a total possibility of 9,437,340 pairwise gene co-expressions. There are 3923 PPIs in Rajagopala., *et al.* [16]'s data set, consisting of 2,044 unique Locus Tag IDs. Of which (Figure 1), 2012 Locus Tags were common in gene expression data set and PPIs, 2333 Locus Tags were found in gene expression data set but not PPIs, and 32 Locus Tags were found in PPIs but not gene expression data set. This suggests that PPIs are rare in *E. coli* but is consistent in terms of magnitude to an earlier study using pull-down assays [33]. Using all pairwise gene co-expressions against PPIs, the baseline precision and recall are 0.0359% and 87.2%, respectively; which gives a baseline F1-score of 0.072% (Table 2 to 4).

Randomization results shows that absolute Pearson's correlation above 0.749 (0.029 + 1.645 x 0.4378 = 0.749) to be significant (Table 1), which corresponds to the absolute Pearson's correlation above 0.75 by Reverter., *et al.* [3] that coincides with 1% false discovery rate. This suggests that randomization can be a suitable procedure to determine a statistically suitable correlation threshold. Using these

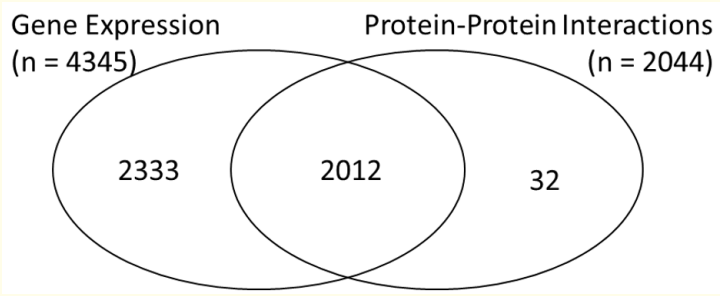

Figure 1: *Common Locus Tags in Data Sets.*

Co-expression Measure	Grand Mean	Standard Error	Standard Deviation
Bray and Curtis coefficient	0.900	0.0024	0.0766
Cosine coefficient	0.997	0.0001	0.0047
Canberra distance	26.546	0.5450	17.2355
Euclidean distance	29.620	0.6758	21.3623
Kendall's tau	0.029	0.0057	0.1810
Manhattan distance	456.742	12.0902	382.3240
Pearson's correlation	0.029	0.0138	0.4378
Point biserial correlation	0.026	0.0114	0.3608
Spearman's correlation	0.033	0.0115	0.3644
Tanimoto coefficient	0.942	0.0024	0.0768

Table 1: *Distributions for Null Hypotheses for Each Co-Expression Measure.*

Co-expression Measure	Number of Significant GCNs	Precision	Recall	F1-Score
Baseline	9,437,340	0.00036	0.87217	0.00072
Bray and Curtis coefficient	0	NA	NA	NA
Canberra distance	891,255	0.00031	0.07101	0.00062
Cosine coefficient	0	NA	NA	NA
Euclidean distance	730,231	0.00038	0.07025	0.00075
Kendall's tau	2,098,668	0.00038	0.20199	0.00075
Manhattan distance	637,694	0.00036	0.05905	0.00072
Pearson's correlation	131,083	0.00188	0.06266	0.00364
Point biserial correlation	576,371	0.00069	0.10168	0.00138
Spearman's correlation	752,842	0.00053	0.10094	0.00105
Tanimoto coefficient	0	NA	NA	NA

Table 2: *Performance of Co-Expression Measures I. Absolute gene co-expressions higher than 1.645 times of standard deviation above the mean of randomized co-expressions are significant.*

Co-expression Measure	Precision	Recall	F1-Score
Baseline	0.00036	0.87217	0.00072
Bray and Curtis coefficient	0.00054	0.06519	0.00108
Canberra distance	0.00023	0.01018	0.00045
Cosine coefficient	0.00063	0.07522	0.00124
Euclidean distance	0.00030	0.02239	0.00060
Kendall's tau	0.00074	0.08820	0.00146
Manhattan distance	0.00041	0.11609	0.00082
Pearson's correlation	0.00083	0.09913	0.00164
Point biserial correlation	0.00083	0.09913	0.00164
Spearman's correlation	0.00070	0.08463	0.00140
Tanimoto coefficient	0.00056	0.06701	0.00111

Table 3: *Performance of Co-Expression Measures II. Top 5 percentile of gene co-expressions are significant.*

Co-expression Measure	Precision	Recall	F1-Score
Baseline	0.00036	0.87217	0.00072
Bray and Curtis coefficient	0.00054	0.06519	0.00108
Canberra distance	0.00023	0.01018	0.00045
Cosine coefficient	0.00063	0.07522	0.00124
Euclidean distance	0.00030	0.02239	0.00060
Kendall's tau	0.00060	0.09381	0.00119
Manhattan distance	0.00041	0.11609	0.00082
Pearson's correlation	0.00063	0.10882	0.00126
Point biserial correlation	0.00063	0.10882	0.00126
Spearman's correlation	0.00057	0.09151	0.00113
Tanimoto coefficient	0.00056	0.06701	0.00111

Table 4: *Performance of Co-Expression Measures III. Top 5 percentile of absolute gene co-expressions are significant.*

thresholds, our results suggest that Pearson's correlation is optimal in terms of precision and F1-score (Table 2).

Using top 5 percentile for co-expression (Table 3) or absolution co-expression (Table 4), our results suggests that both Pearson's correlation and point biserial correlation are equally suitable as there is perfect correlation (r = 1) between co-expression values calculated using Pearson's correlation or point biserial correlation. Interestingly, Canberra distance which fair worse than baseline in both statistically determined threshold (Table 2) or co-expression (Table 3) in terms of precision, is a suitable when absolution co-expression (Table 4) was used. This suggests that the type of threshold and co-expression measure should be considered together. In addition, our results suggests that a more stringent threshold is likely to give better precision and F1-score compared to a less stringent threshold (Table 5).

However, our results suggest that all evaluated gene co-expression measures have low precision in predicting PPIs (Tables 2 to 4, precision ≤ 0.00188). This suggests that gene co-expression is a poor predictor of PPIs. Several studies have suggested that interacting

Percentile	Number of Significant GCNs	Precision	Recall	F1-Score
Baseline	9,437,340	0.00036	0.87217	0.00072
> 90	1,518,628	0.00044	0.17155	0.00088
> 91	1,343,990	0.00047	0.15953	0.00093
> 92	1,172,692	0.00049	0.14653	0.00098
> 93	1,003,382	0.00053	0.13507	0.00105
> 94	837,776	0.00058	0.12283	0.00115
> 95	676,166	0.00063	0.10882	0.00126
> 96	519,878	0.00073	0.09684	0.00145
> 97	370,858	0.00090	0.08461	0.00177
> 98	230,883	0.00121	0.07108	0.00238
> 99	105,341	0.00224	0.06011	0.00432

Table 5: *Effects of Percentile Threshold using Pearson's Correlation on Performance. Absolute gene co-expressions were used.*

proteins are more likely to be expressionally correlated [34,35], especially for permanent interacting proteins such as within ribosomes and proteosomes [36]; which is consistent with our results in terms of precision but not recall. This suggests that a large proportion of PPIs may not be identified using gene co-expressions. Moreover, the practicality of using gene co-expression to predict PPIs is low due to low precision. In spite of this, gene co-expression may be a screening useful tool for other purposes; such as, gene-disease associations [5,37,38], and disease subtype classification [39-41].

Conclusion

Ten gene co-expression measures were examined for their applicability to predict PPIs in *E. coli*. Although Pearson's correlation consistently outperforms other measures, it has low precision. Hence, gene co-expression measures alone are not suitable to predict PPIs.

Data files for this study can be downloaded at https://bit.ly/GCN_vs_PPI.

Acknowledgement

This work was conducted as part of the Temasek Polytechnic School of Applied Science Differential Research Project under the Student Project fund (TP_PR1052).

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

1. Stuart JM., *et al*. "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules". *Science* 302.5643 (2003): 249-255.

2. Serin EAR., *et al*. "Learning from Co-expression Networks: Possibilities and Challenges". *Frontiers in Plant Science* 7 (2016): 444.

3. Reverter A., *et al*. "Construction of Gene Interaction and Regulatory Networks in Bovine Skeletal Muscle from Expression Data". *Australian Journal of Experimental Agriculture* 45 (2005): 821-829.

4. Ling MH. "Understanding Mouse Lactogenesis by Transcriptomics and Literature Analysis [Doctor of Philosophy]. [Parkville, Victoria]: The University of Melbourne (2009).

5. Van Dam S., *et al*. "Gene Co-Expression Analysis for Functional Classification and Gene-Disease Predictions". *Briefings in Bioinformatics* 19.4 (2018): 575-592.

6. Sharma A., *et al*. "Transcriptome Profiling Reveal Key Hub Genes in Co-Expression Networks Involved in Iridoid Glycosides Biosynthetic Machinery in Picrorhiza kurroa". *Genomics* 113.5 (2021): 3381-3394.

7. Fajardo TVM and Quecini V. "Comparative Transcriptome Analyses between Cultivated and Wild Grapes Reveal Conservation of Expressed Genes but Extensive Rewiring of Co-Expression Networks". *Plant Molecular Biology* 106.1-2 (2021): 1-20.

8. Zheng Q., *et al*. "Comparative Transcriptome Analysis Reveals Regulatory Network and Regulators Associated with Proanthocyanidin Accumulation in Persimmon". *BMC Plant Biology* 21.1 (2021): 356.

9. Wang Y., *et al*. "Integrated Profiling Identifies CCNA2 as a Potential Biomarker of Immunotherapy in Breast Cancer". *OncoTargets and Therapy* 14 (2021): 2433-2448.

10. Li X., *et al*. "A Five Immune-Related lncRNA Signature as a Prognostic Target for Glioblastoma". *Frontiers in Molecular Biosciences* 8 (2021): 632837.

11. Li S., *et al*. "hsa_circ_0000729, A Potential Prognostic Biomarker in Lung Adenocarcinoma". *Thoracic Cancer* 9.8 (2018): 924-930.

12. Husain B., *et al*. "NetExtractor: Extracting a Cerebellar Tissue Gene Regulatory Network Using Differentially Expressed High Mutual Information Binary RNA Profiles". *G3 Bethesda* 10.9 (2020): 2953-2963.

13. Niu X., *et al*. "Weighted Gene Co-Expression Network Analysis Identifies Critical Genes in the Development of Heart Failure After Acute Myocardial Infarction". *Frontiers in Genetics* 10 (2019): 1214.

14. Chay ZE., *et al*. "Russel and Rao Coefficient is a Suitable Substitute for Dice Coefficient in Studying Restriction Mapped Genetic Distances of Escherichia coli". *iConcept Journal of Computational and Mathematical Biology* 1 (2010): 1.

15. Piya S., *et al*. "Protein-Protein Interaction and Gene Co-Expression Maps of ARFs and Aux/IAAs in Arabidopsis". *Frontiers in Plant Science* 5 (2014): 744.

16. Rajagopala SV., *et al*. "The Binary Protein-Protein Interaction Landscape of Escherichia coli". *Nature Biotechnology* 32.3 (2014): 285-290.

17. Faith JJ., *et al*. "Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles". *PLOS Biology* 5.1 (2007): e8.

18. Barrett T and Edgar R. "Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis". *Methods in Enzymology* 411 (2006): 352-369.

19. Gautier L., *et al*. "Affy - Analysis of Affymetrix GeneChip Data at the Probe Level". *Bioinformatics* 20.3 (2004): 307-315.

20. Ling MHT. "SeqProperties: A Python Command-Line Tool for Basic Sequence Analysis". *Acta Scientific Microbiology* 3.6 (2020): 103-106.

21. Ling MH. "COPADS, I: Distances Measures Between Two Lists or Sets". *The Python Papers Source Codes* 2 (2010): 2.

22. Bray JR and Curtis JT. "An Ordination of the Upland Forest Communities of Southern Wisconsin". *Ecological Monographs* 27.4 (1957): 325-349.

23. Singhal A. "Modern Information Retrieval: A Brief Overview". *The Bulletin of the Technical Committee on Data Engineering* 24.4 (2001): 35-43.

24. Lance GN and Williams WT. "Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses")". *The Computer Journal* 9.1 (1966): 60-64.

25. Pidò S., *et al*. "Computational Analysis of Fused Co-Expression Networks for the Identification of Candidate Cancer Gene Biomarkers". *NPJ Systems Biology and Applications* 7.1 (2021): 17.

26. Kendall MG. "A New Measure of Rank Correlation". *Biometrika* 30.1-2 (1938): 81-93.

27. Krause EF. "Taxicab Geometry: An Adventure in Non-Euclidean Geometry". New York: Dover Publications (1987): 88.

28. Student. "Probable Error of a Correlation Coefficient". *Biometrika* 6.2-3 (1908): 302-310.

29. Liesecke F., *et al*. "Ranking Genome-Wide Correlation Measurements Improves Microarray and RNA-Seq Based Global and Targeted Co-Expression Networks". *Scientific Reports* 8.1 (2018): 10885.

30. Lev J. "The Point Biserial Coefficient of Correlation". *Annals of Mathematical Statistics* 20.1 (1949): 125-126.

31. Rácz A., *et al*. "Life Beyond the Tanimoto Coefficient: Similarity Measures for Interaction Fingerprints". *Journal of Cheminformatics* 10.1 (2018): 48.

32. Hooton JWL. "Randomization tests: statistics for experimenters". *Computer Methods and Programs in Biomedicine* 35.1 (1991): 43-51.

33. Arifuzzaman M., *et al*. "Large-Scale Identification of Protein-Protein Interaction of Escherichia coli K-12". *Genome Research* 16.5 (2006): 686-691.

34. Grigoriev A. "A Relationship Between Gene Expression and Protein Interactions on the Proteome Scale: Analysis of the Bacteriophage T7 and the Yeast Saccharomyces cerevisiae". *Nucleic Acids Research* 29.17 (2001): 3513-3519.

35. Bhardwaj N and Lu H. "Correlation Between Gene Expression Profiles and Protein-Protein Interactions Within and Across Genomes". *Bioinformatics* 21.11 (2005): 2730-2738.

36. Jansen., *et al*. "Relating Whole-Genome Expression Data with Protein-Protein Interactions". *Genome Research* 12.1 (2002): 37-46.

37. Paci P., *et al*. "Gene Co-Expression in the Interactome: Moving from Correlation Toward Causation via an Integrated Approach to Disease Module Discovery". *NPJ Systems Biology and Applications* 7.1 (2021): 3.

38. Hartman RJG., *et al*. "Sex-Dependent Gene Co-Expression in the Human Body". *Scientific Reports* 11.1 (2021): 18758.

39. Bar H and Bang S. "A Mixture Model to Detect Edges in Sparse Co-Expression Graphs with an Application for Comparing Breast Cancer Subtypes. Oliva G, editor". *PLOS ONE* 16.2 (2021): e0246945.

40. Mallik S and Zhao Z. "ConGEMs: Condensed Gene Co-Expression Module Discovery Through Rule-Based Clustering and Its Application to Carcinogenesis". *Genes* 9.1 (2017): 7.

41. Qin L., *et al*. "Application of Weighted Gene Co-Expression Network Analysis to Explore the Potential Diagnostic Biomarkers for Colorectal Cancer". *Molecular Medicine Reports* 21.6 (2020): 2533-2543.

Volume 18 Issue 3 March 2022