Pacific Northwest
NATIONAL LABORATORY

PNNL-34207

Constructing Regulatory Networks to Compare Axenic and Interspecies Microbial Gene Transcription

April 2023

Johanna L. Owen
Pavlo Bohutskyi
Sarah K. Bogart
David A. Anderson

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-34207

Constructing Regulatory Networks to Compare Axenic and Interspecies Microbial Gene Transcription

April 2023

Johanna L. Owen
Pavlo Bohutskyi
Sarah K. Bogart
David A. Anderson

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Constructing regulatory networks to compare axenic and interspecies microbial gene transcription

J. Owen, P. Bohutskyi, S.K. Bogart, D. Anderson

Biological Science Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA

Abstract

In this preliminary study, we constructed gene regulatory networks (GRNs) from transcriptional expression data of axenic and interspecies microbial cultures with the goal of predicting how cocultivation affected greenhouse gas respiration by these species. The specific strains of *Methylotuvimicrobium alkaliphilum* 20Z, a methylotroph, and *Cyanobacterium stanieri* HL-69, a phototroph, were chosen for their viability in industrial bioprocessing. We ranked directed interactions between gene pairs based on the ability of the input gene to predict the expression of a target gene relative to their transcriptomes.

While we were able to identify topological differences between conditions, our initial findings require validation through experimental analysis and further modeling. We aimed to develop a systematic thresholding approach to optimize the accuracy of our networks. We filtered out trial networks separately from top gene interactions of the scored rankings. Parameters of unfiltered and

filtered networks were used to test and develop thresholding approaches. Knee point detection of edge weight distributions was explored as an approach for separating significant interactions from insignificant interactions in unfiltered networks. While knee detection failed to produce analogous networks for broad cross-condition comparisons, the results informed us about the proportions of significant edges present in unfiltered networks. We also calculated the average mean degree for nodes in a selection of trial networks to find a thresholding value characteristic to all groups. While we did not reach a definitive conclusion, we gained insight into the coregulatory structures of our groups and made critical evaluations of systematic methods for filtering networks. We recommend an iterative process for the inference of GRNs, where the most significant results from preliminary explorations are used to improve the efficiency with which regulatory motifs are chosen for experimental characterization. Experimental results can then inform the framework of adjusted models to improve broad interpretations of GRNs.

INTRODUCTION

Microbial consortia are known to exhibit linkages of genes between species in interactions such as cross feeding.[1] Conversion of wastewater products and experimental instances of methane and carbon dioxide biotransformation using interspecies interactions suggests that greenhouse gases could be captured for bioprocessing with microbes at an industrial scale.[2, 3, 4] Modeling the gene regulatory networks (GRNs) of species considered for industrial cultivation can provide insight into relevant metabolic processes.[5] Coculture expression data of a methylotroph, *Methylotuvimicrobium alkaliphilum* 20Z, and a phototroph,

Cyanobacterium stanieri HL-69, were compared with expression data from axenic cultures of both species through inference of GRNs from RNA sequencing data. Our study analyzed five groups of gene interactions from three types of microbial cultures, including *M. alkaliphilum* 20Z – *C. stanieri* HL-69 cocultures grown under various conditions. We isolated interspecies interactions from the coculture dataset and analyzed these separately. *M. alkaliphilum* 20Z genes and *C. stanieri* HL-69 genes from coculture were also analyzed separately. The coculture dataset was split into three categories so that expression differences between axenic and coculture conditions for single organisms could be compared and to observe interspecies interactions in isolation. Additionally, we sequenced and analyzed each axenic culture separately.

Our ambition was to determine if coculture conditions caused significant changes in the metabolic activity of these organisms as the result of interspecies interactions, and to understand underlying mechanisms. This report describes initial steps to this process. Our networks reveal the transcriptional structure of organisms as hierarchical communities organized from individual interactions between genes.[5, 6, 7, 8] In the results section, we compare neighborhood connectivity, clustering coefficients, and average node degree across networks. Average neighborhood connectivity is a measure of the average degree of nodes within the neighborhoods of a network, which characterizes the level of interconnectivity found in subcommunities of genes.[9] Clustering coefficients, which are calculated on a unit scale, describe the tendency of the network to organize into clusters of coregulated genes.[10] Node degree measures the number of connections a gene has with other genes. Through visual representation in Cytoscape, we were able to differentiate clusters of genes within networks and characterize the centrality of genes through measures of degree, where genes with higher degree are more central to networks.[11, 12, 13] Within the context of network size, these parameters characterize coregulatory hierarchies from various

perspectives. Decreased detection of coregulation was expected in interactions between the two organisms, since coexpression between genes of different organisms occurs at a higher level of biological organization than coexpression in single organisms.[6] In the beginning stages of network inference, a realistic goal is to identify a limited selection of coregulatory relationships from initial models for further experimental characterization. Coregulatory relationships validated through experimental data can then serve as a sparse ground truth network from which to remodel the broad characteristics of networks. For instance, networks may be extended from well-characterized regulatory motifs to their second-order network neighborhoods.[14]

One major challenge we faced was in determining an appropriate threshold to filter false positive interactions out of GRNs. Interactions with lower levels of significance are more likely to represent false positives in regulatory networks modeled with our methods.[5, 7, 15] A systematic approach for thresholding would provide a reliable foundation to make comparisons between GRNs that vary in basic qualities such as species composition.[5, 16] To address this challenge, we explored various techniques such as knee point detection of edge weight distributions and computation of average mean degree centrality across a selection of filtered groups.

Overall, our study contributes to ongoing efforts to improve network inference methods and provides a foundation for further research into regulatory mechanisms underlying interspecies interactions of microbes.

METHODS

Data sourcing. The raw RNA expression data used in our network analysis protocols were provided by Pacific Northwest National Laboratory. Raw counts were obtained for axenic *M. alkaliphilum* 20Z, axenic *C. stanieri* HL-69, and cocultures of the two organisms. Wet laboratory, sequencing, and alignment protocols were completed at disjointed times across

groups. Notably, the set of conditions tested for each group were not analogous, as the original experimental purposes for each group were not intended for cross-group analysis. Counts were acquired as text files. Text files were used to assemble a matrix for each group describing raw counts for individual genes across samples.

Cleaning and normalization of expression data. Raw data was cleaned and normalized before use for network inference. Ribosomal RNA counts were removed from all datasets to eliminate rRNA over-representation in sequencing data. The remaining raw counts were then normalized with the DESeq2 package in RStudio. The coculture dataset was split by the two organismal gene types prior to normalization to preserve expression patterns influenced by species identity. A variance stabilizing transformation (VST) was used to normalize the raw counts by dividing scaling factors calculated for samples into gene-wise dispersion estimates. The VST transforms count data to create constant variance along mean expression values.[17] Outlier samples were identified with principal component analysis and removed from datasets. Datasets were normalized and cleaned of outliers in repeating cycles until outliers were no longer identified.

Network inference. The GENIE3 algorithm was used to infer regulatory networks for the five distinct groups of genes. For each group, normalized expression data was processed by the GENIE3 algorithm in RStudio. Coculture data was remerged after normalization to infer interspecies interactions. The split coculture datasets were also used individually to infer single organism networks from coculture. GENIE3 generates a learning sample for each gene *j* containing the set of paired vectors between the expression of gene *j* across each *k* sample (x_k^j) with the expression for each other gene across samples (x_k^{-j}). Each learning sample was computed as a supervised, non-parametric regression subproblem within the general problem of

network recovery from all genes in the datasets. Using squared error loss and true expression values from the count data, a function f_j is found for each learning sample where error is minimized. The Random Forest method was selected, where regression trees were generated from 1,000 bootstrapped samples of vector pairs from each learning sample. Predictions made by input genes were averaged across each tree ensemble. A local ranking of inputs was generated for each learning sample based on the importance of an input gene in predicting target gene expression. Local rankings were aggregated into an overall interaction ranking of the network, where each interaction received a weighted value corresponding to rank.[7] An adjacency matrix with the weight of interactions between every pair of genes was obtained, with one entry for the first gene from a pair as the target and another entry for the second gene as target. Any interaction with a weight above zero corresponded to an 'edge' in the network.

Network filtration. Prior to filtration, the processed coculture data contained a merged adjacency matrix of interspecies and single organism interactions. Single organism interactions were eliminated to produce an exclusively interspecies network. The resulting networks of all five groups, consisting of millions of edges, required extreme reduction for accuracy and analysis. Networks were filtered using several trial thresholds to define the maximum number of edges allowed in each trimmed network, ranging from 500 to 2,500 edges in increments of 250 edges. Information for the resulting trial networks were then exported as SIF files for use in Cytoscape and as GraphML files for Fast Greedy community detection.[8, 12]

Community detection. Automated community detection was performed using the Fast Greedy algorithm to identify densely interconnected gene communities in each trial network. This algorithm employs a greedy optimization to maximize network modularity. The algorithm groups subcommunities together to maximize a specific modularity score.[8] Only communities

with at least twenty genes were selected. A CSV file listing the module identity of each gene was exported for use in Cytoscape.

Graphing trial networks. Cytoscape visualized and analyzed trial networks for all groups. SIF files containing pairwise interactions between genes were imported to build network graphs. Community detection results were imported to the node table, where nodes represented genes. The NetworkAnalyzer tool computed topological parameters for each network as a whole and for each gene. Cytoscape style functions were used to color-code each community and to scale the representative size of each gene by centrality values.[12, 13] Topological parameters were saved from Cytoscape as CSV files. Networks for each group were compared within and across groups in a table organized by the shifting magnitude of trial thresholds. Trial networks were evaluated for the number of edges that their primary components contained. A single network was selected from each group to create a selection of networks with the closest possible range in the size of their primary components. Topological parameters for the resulting selection of networks were compared between groups. The mean value of average node degrees across groups was calculated. Cumulative distributions of node degree for all groups were plotted together in RStudio.[18]

Knee point detection. The Kneedle algorithm was used with Python in Jupyter to detect knee points for edge weight distributions of the unfiltered networks obtained from GENIE3. This was done as a potential technique for finding appropriate thresholds to trim the networks. This data-driven approach differentiates significant interactions from insignificant interactions, which can lead to more accurate network analyses by minimizing false positive edges.[15] Networks for axenic *M. alkaliphilum* 20Z, axenic *C. stanieri* HL-69, and interspecies interactions were created using knee point thresholds to test this technique as a filtering approach.

RESULTS

Table 1. Parameters describing axenic and coculture networks of *M. alkaliphilum* 20Z and *C. stanieri* HL-69 [a]

Gene group	Total nodes in network	Total edges in primary component	Percentage of edges in secondary components	Clustering coefficient	Average neighborhood connectivity (edges per node per neighborhood)
Axenic *M. alkalaphilum* 20Z	3,148	2,310	27%	0.240	6.609
M. alkaliphilum 20Z from coculture	2,605	2,065	21%	0.132	4.483
Axenic *C. stanieri* HL-69	2,698	2,376	12%	0.110	4.802
C. stanieri HL-69 from coculture	2,733	2,523	08%	0.109	5.133
Interspecies interactions	2,852	2,269	20%	< 0.001	5.010

[a] Parameters were obtained from Cytoscape[12, 13] to compare a selection made from one trial network of each group, chosen based on similar numbers of edges in primary components.

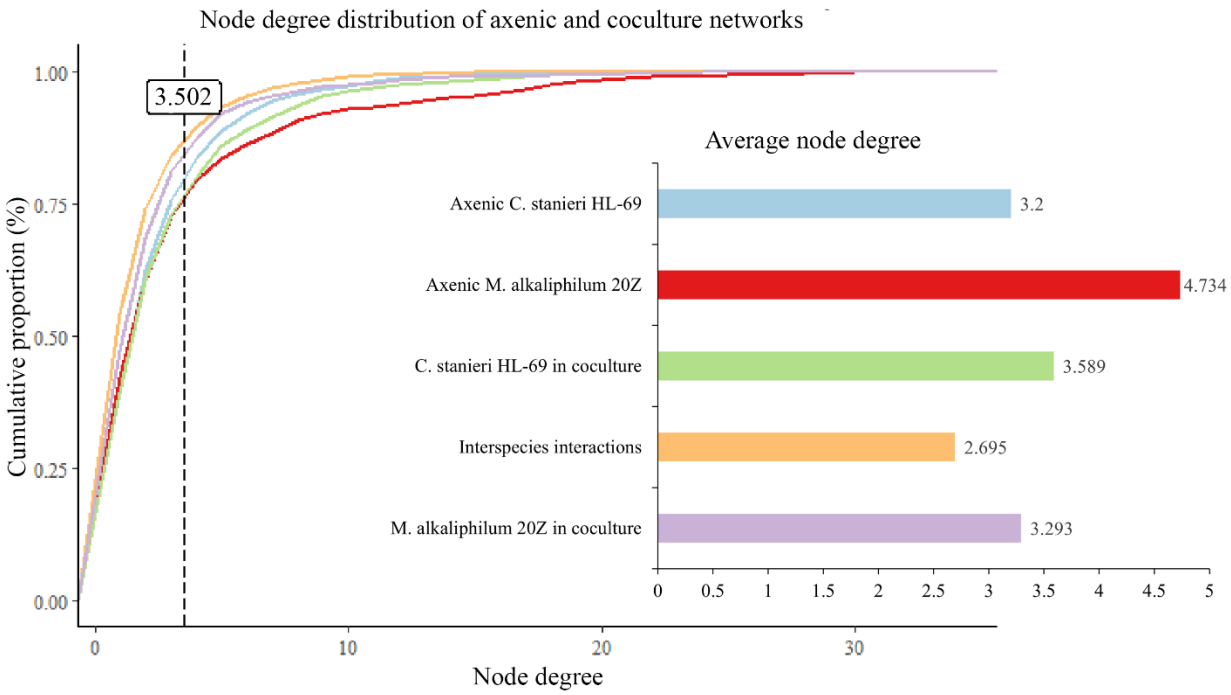

FIG. 1. Cumulative distributions for the degree parameter of individual nodes plotted for each of the five groups and the average node degree parameter of each group. The average mean degree across all five groups (3.502) is indicated with a vertical dotted line on the plot of distributions.

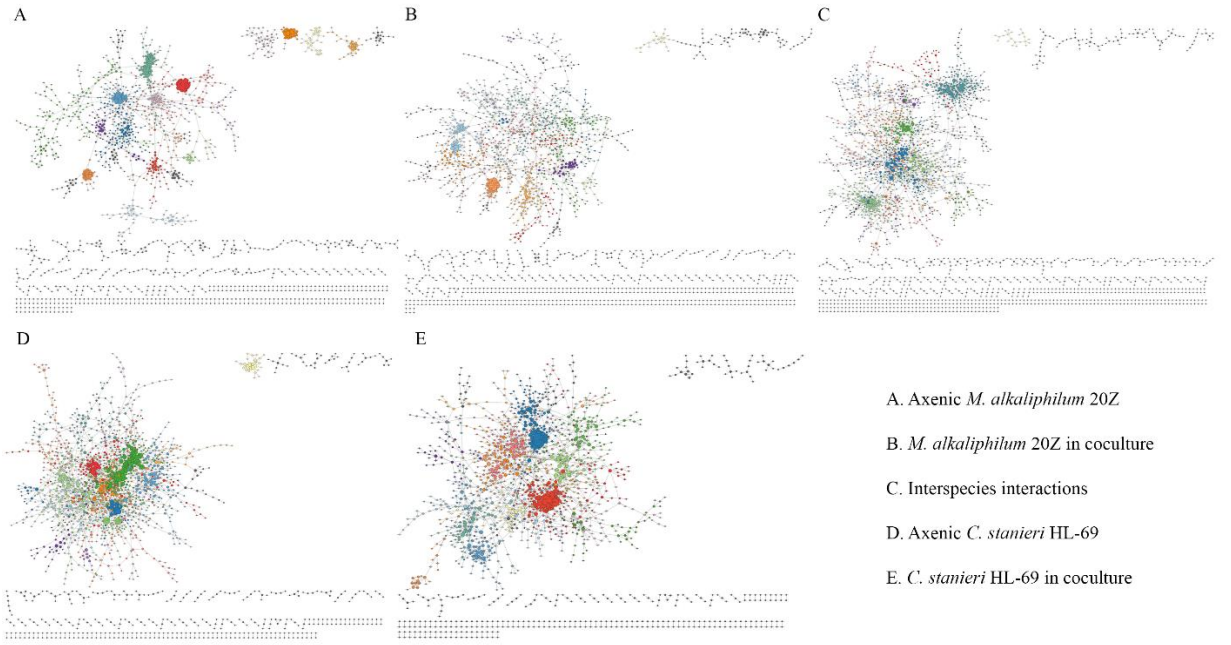

FIG. 2. A selection of trial networks from each of the five groups studied. Secondary components appear as segments that are disconnected from primary components. Colors indicate gene modules, and the scale of a node represents its degree centrality.

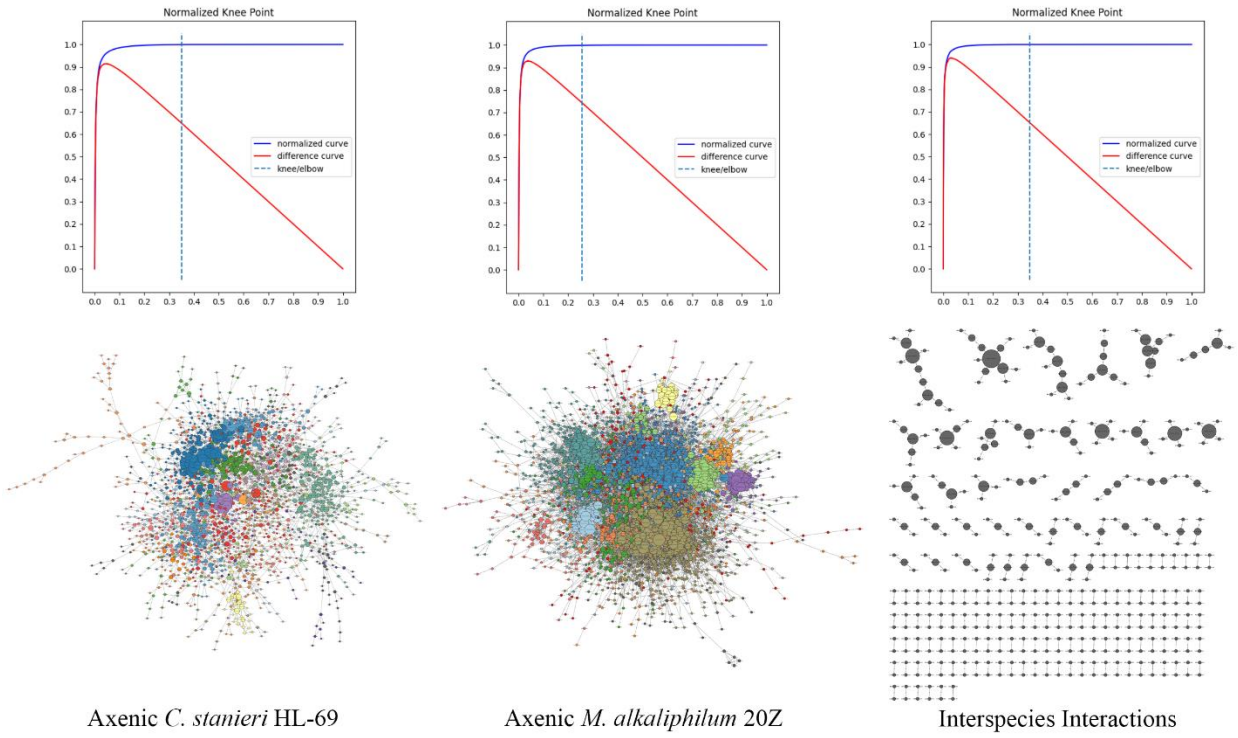

FIG. 3. Knee point detection plots and resulting networks for axenic *C. stanieri* HL-69, axenic *M. alkaliphilum* 20Z, and interspecies interactions. Colors indicate gene modules, and the scale of a node represents its degree centrality.

Notable differences were found between the parameters of the five groups studied. Table 1 presents a comparison of parameters between a selection of trial networks. The primary components of all trial networks across groups were found to converge most closely in size within a range of 2,065 to 2,523 edges. The axenic *M. alkaliphilum* 20Z group had the highest average neighborhood connectivity (6.609 edges per node per neighborhood) out of all networks. The clustering coefficient of the axenic *M. alkaliphilum* 20Z group was larger (0.240) than the

average clustering coefficient for all other groups (0.088). Connectivity (4.483 edges per node per neighborhood) and clustering (0.132) decreased for *M. alkaliphilum* 20Z in coculture.

The interspecies network included only those interactions between a *M. alkaliphilum* 20Z gene and a *C. stanieri* HL-69 gene. The clustering coefficient of the interspecies network was less than 0.001, and the average degree of this group was the smallest (2.695 edges per node) out of all groups [Figure 1]. The average neighborhood connectivity of the interspecies network was moderate (5.010 edges per node per neighborhood) in relation to the range across groups (4.483 to 6.609 edges per node per neighborhood). Distinct communities of genes with varying centrality are evident in the visual representation of the interspecies network shown in Figure 2.

The average mean degree across all groups was 3.502 edges per node. Figure 1 shows that 3.502 corresponds to an approximate knee of the degree distributions for all groups. Exact knee point detection of edge weight distributions highlighted the proportion of significant interactions in the networks analyzed. Networks created with knee thresholds were constructed from edges above the knee point of their edge weight distribution. This resulted in a relative low number of significant edges found for the interspecies network (281 edges), a relative high number of significant edges found for the axenic *M. alkaliphilum* 20Z network (15,203 edges), and a relative intermediate number of significant edges found for *C. stanieri* HL-69 (3,560 edges) [Figure 3]. The pre-filtered interspecies network from GENIE3 had over 53 million edges, while the axenic *M. alkaliphilum* 20Z network from GENIE3 had over 17 million edges, and the GENIE3 axenic *C. stanieri* HL-69 network had over 9 million edges.

DISCUSSION

The higher clustering coefficient and average neighborhood connectivity along with a greater number of significant interactions found in the axenic *M. alkaliphilum* 20Z network

suggests that coexpression of genes for *C. stanieri* HL-69 is either less coordinated or is less organized around strategies of coregulation. Decreased connectivity and clustering in the single organism *M. alkaliphilum* 20Z network from coculture compared to axenic conditions implies that cocultivation with *C. stanieri* HL-69 disrupted *M. alkaliphilum* transcription patterns.[6]

The considerably lower number of significant edges and decreased clustering coefficient of the interspecies network was expected in comparison to networks involving single organism interactions, as coregulation between the genes of two separate organisms occurs at a higher level of biological organization.[6] Hierarchical community organization and genes with elevated degree centrality are evident in the visual representation of the interspecies network. A lower proportion of significant interactions found with knee point detection indicates that the interspecies network may have a greater concentration of false positives than other groups studied.[5, 7, 15] Broad analysis of this network may be prone to error, although evidence of interspecies coregulation may be supported through specific investigation of genes with high degree centrality and high local clustering coefficients.[14]

Knee point thresholds were not useful for generating a selection of analogous networks for broad comparisons, as these thresholds resulted in networks that were too variable in size. However, knee point detection provided valuable information about the degree to which significant instances of coregulation occurred across groups, which also helps to characterize the proportion of false positives that a network may contain.[5, 7, 15]

Average mean degree centrality may be an effective method for further reduction of our networks as a value that is reflective of all five groups.[11] However, the construction of accurate regulatory networks is fundamentally challenged if ground truth networks with which to compare results are not available. It is recommended that a limited selection of genes participating in top-

ranked interactions are selected from our networks for experimental characterization. Networks can then be re-modelled around confirmed regulatory relationships before broad inferences are made.[14] There have been ongoing efforts by our team to characterize *M. alkaliphilum* 20Z and *C. stanieri* HL-69 genes through standard databases.

Our methodology reflects a common issue in network inference: the lack of a standard approach to threshold selection.[5, 16] While we attempted to overcome this challenge by exploring multiple techniques and carefully selecting parameter values, the lack of a clear consensus on an optimal approach remains a major limitation of our study. Despite this limitation, we were able to compare a set of filtered networks for each of the five groups at an analogous level. Our analysis of topological parameters revealed significant differences between the networks of different groups, and potential gene regulatory mechanisms can be explored through a focus on the most significant interactions that were found.

ACKNOWLEDGEMENTS

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI). Thank you to Pacific Northwest National Laboratory (PNNL) for support of this research, especially the Biological Sciences Division. Thank you to Pavlo Bohutskyi and Zachary Johnson for guidance in design, implementation, and peer review feedback.

REFERENCES

1. S. M. Krause, T. Johnson, Y. Samadhi Karunaratne, Y. Fu, D. A. Beck, L. Chistoserdova and M. E. Lidstrom, Proc Natl Acad Sci U S A 114 (2), 358-363 (2017).

2. P. Bohutskyi, L. A. Kucek, E. Hill, G. E. Pinchuk, S. G. Mundree and A. S.

Beliaev, Bioresource Technology 260, 68-75 (2018).

3. R. Khanongnuch, R. Mangayil, V. Santala, A. G. Hestnes, M. M. Svenning and A. J. Rissanen, Front Microbiol 13, 874627 (2022).

4. N. Pattharaprachayakul, J.-i. Choi, A. Incharoensakdi and H. M. Woo, Biotechnology and Bioprocess Engineering 25 (6), 829-847 (2020).

5. F. M. Delgado and F. Gómez-Vela, Artificial Intelligence in Medicine 95, 133-145 (2019).

6. O. Q. H. Zinani, K. Keseroğlu and E. M. Özbudak, Trends in Genetics 38 (1), 73-81 (2022).

7. V.A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, PLOS ONE 5, e12776 (2010).

8. A. Clauset, M. E. J. Newman and C. Moore, Physical Review E 70 (6), 066111 (2004).

9. J. M. González, Heliyon 7 (1), e05867 (2021).

10. J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertész, Physical Review E 75, (2007).

11. G. Scardoni and C. Laudanna, (2012).

12. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, Genome Res **13** (11), 2498-2504 (2003).

13. Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer and M. Albrecht, Bioinformatics 24 (2), 282-284 (2007).

14. R. S. McClure, C. C. Overall, E. A. Hill, H.-S. Song, M. Charania, H. C. Bernstein, J. E. McDermott and A. S. Beliaev, The ISME Journal 12 (8),

2011-2023 (2018).

15. V. Satopaa, J. Albrecht, D. Irwin and B. Raghavan, presented at the 2011 31st International Conference on Distributed Computing Systems Workshops, 2011 (unpublished).

16. L. L. Elo, H. Järvenpää, M. Orešič, R. Lahesmaa and T. Aittokallio, Bioinformatics 23 (16), 2096-2103 (2007).

17. M.I. Love, W. Huber, and S. Anders, Genome Biology 15, (2014).

18. W. H, ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag, New York, 2016).

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov