PERSPECTIVE
Molecular Biology and Physiology

Toward a Better Understanding of Species Interactions through Network Biology

Ryan S. McClure[a]

[a]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA

ABSTRACT Within the last decade, there has been an explosion of multi-omics data generated for several microbial systems. At the same time, new methods of analysis have emerged that are based on inferring networks that link features both within and between species based on correlation in abundance. These developments prompt two important questions. What can be done with network approaches to better understand microbial species interactions? What challenges remain in applying network approaches to query the more complex systems of natural settings? Here, I briefly describe what has been done and what questions still need to be answered. Over the next 5 to 10 years, we will be in a strong position to infer networks that contain multiple kinds of omic data and describe systems with multiple species. These applications will open the door for a better understanding and use of microbiomes across a variety of fields.

KEYWORDS microbiome, multi-omic, network

GENE COEXPRESSION NETWORKS

Over the last several years, there has been a rapid increase in the depth, cost efficiency, and use of high-throughput omics techniques that can query the abundance levels of a large number of biological molecules simultaneously. These include amplicon analysis, transcriptomics, proteomics, and metabolomics, among others. The collection of these data can answer many questions focused on responses of organisms to specific conditions. However, when experiments collect tens to hundreds of omics data sets, analysis can take the next step and use these data to infer feature coexpression networks. While the specific methods vary, all such networks are essentially based on linking pairs of features (with a feature being either a species, transcript, protein, metabolite, etc.) based on their correlation in abundance (i.e., expression) across a range of variable environmental conditions under which the omics data have been collected. Once inferred, such networks can show which features are coexpressed and can reveal a number of characteristics of the system under question, including how features are related to each other in their expression or abundance, which functional processes are coordinated, which features occupy central positions in the network and are thus important for the growth and bioactivity of the system, which transcripts/proteins act in regulatory pathways, and what the putative function of uncharacterized genes may be. Coexpression networks have been inferred for a number of different systems and have been an especially critical tool in examining human disease (1–3).

USING GENE COEXPRESSION NETWORKS TO EXAMINE CYANOBACTERIAL SYSTEMS

In our work, I have used network approaches with several different species with the greatest focus on bacterial systems. I have published a number of studies using network analysis to query a model cyanobacterial species, *Synechococcus* 7002. In one set of experiments, I collected transcriptomic data for *Synechococcus* 7002 under 42 separate short-term growth experiments and inferred a network using a mutual

Citation McClure RS. 2019. Toward a better understanding of species interactions through network biology. mSystems 4:e00114-19. https://doi.org/10.1128/mSystems.00114-19.

Copyright © 2019 McClure. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Ryan.McClure@pnnl.gov.

Conflict of Interest Disclosures: R.S.M. has nothing to disclose.

mSystems® *vol. 4, no. 3, is a special issue sponsored by Illumina.*

Network analysis of multi-species systems will soon lead to greater understanding of these communities and their use in improving human well-being

Received 15 February 2019
Accepted 13 April 2019
Published 28 May 2019

Downloaded from http://msystems.asm.org/ on September 21, 2019 by guest

information-based tool termed Context Likelihood of Relatedness (CLR). The resulting network of several hundred nodes (transcripts) and edges (instances of high coexpression between transcripts across conditions) was used to identify new targets of protein and RNA regulators in this organism and view how specific processes were coordinated during growth (4). I also expanded this analysis by identifying strong overlap between those genes that have high centrality (i.e., occupy well-connected positions) in the network and those that are critical for growth and metabolism (5). Centrality can be defined by a number of characteristics, with the two most common being degree (the number of edges a node has with other nodes) and betweenness (a measure of how important a node is in connecting two larger clusters of nodes). Nodes of high centrality can be thought of as keystone nodes, as their removal would have deleterious effects on network structure. The linking of network centrality to general importance has also been seen with networks of other systems, including human pathogens, with analysis identifying a strong association between central position in a network and involvement of the gene in infection and virulence, and human cancer data (2, 3, 6). Regarding networks involving more than one species, these have for the most part been limited to species coabundance networks. Such networks have been powerful tools used to identify important keystone species, including in systems related to human health (7, 8). However, taking the next step and linking specific genes across these species, analogous to the way that genes within a species can be linked as described above, is a field that is still relatively unexplored.

EXPANDING NETWORKS TO MULTIPLE SPECIES

One of the most powerful aspects of gene coexpression networks is the ability to link genes, based on their coexpression, that may be distantly separated on the physical genome. While other approaches to find related genes have focused on physical proximity, coexpression network analysis is location agnostic and can link related genes without this information, indeed in spite of the fact that related genes may be spread across the genome. As network approaches have been used to find gene pairs that are involved in related processes but are physically distant, our group reasoned that this approach could also be applied to systems of multiple species and would be able to link genes that are separated into distinct species but are involved in related processes such as division of labor or metabolic coordination.

APPLICATION OF GENE COEXPRESSION NETWORKS TO MULTISPECIES SYSTEMS

To apply gene coexpression analysis to multiple species, I collected a large amount of transcriptomic data for two species grown in coculture. These species consisted of an autotropic cyanobacterium, *Thermosynechococcus elongatus* BP1, and a heterotrophic species, *Meiothermus ruber* strain A. I cultured these species together under 25 separate short-term perturbations and collected metatranscriptomic data, covering the transcriptome of each species, after each perturbation. For nearly all the perturbations collected, no exogenous sources of organic carbon or nitrogen were added. This means that during growth *M. ruber* was completely dependent for these nutrients on *T. elongatus*. The dependency of *M. ruber* on *T. elongatus* leads to a large amount of interaction and sharing of resources between these organisms during growth. Using the collected transcriptomic data and the CLR program, I inferred a network that linked transcripts between each of these species based on their coexpression. Subsequent analysis of this network revealed that there was a large amount of transcript coordination between the two species that centered on amino acid production, nitrogen metabolism, terpenoid metabolism, and cellular motility (9). By focusing on centrality, I found that these processes occupied much more central (i.e., important) positions in a network comprised of transcripts from both species compared to networks inferred for each species alone. The increased position of these genes indicates greater importance for these processes during coculture and suggests that they are critical for the interactions between these species. This study was one of the first that used network analysis to look at multispecies bacterial systems, and this approach identified several

Downloaded from http://msystems.asm.org/ on September 21, 2019 by guest

processes, as well as specific genes, that may be crucial for phototroph-heterotroph interactions.

Networks that link genes across species based on coexpression are only beginning to be inferred. Two additional recent studies have used them to examine host-pathogen interactions. One study looked at multispecies networks to link processes in *Aspergillus flavus* and its host *Zea mays* and another inferred a network looking at *Candida albicans* during invasion of immune cells of mice (10, 11). By linking genes across species, these studies inferred where there are pathogenic responses of an infecting organism coordinated with immune responses of the host. Uncharacterized genes that show strong coordination across species may point to new aspects of pathogenicity or immune response, opening the door for the development of novel treatment regimens for several diseases.

INTERROGATING NETWORKS FOR BIOLOGICAL INSIGHT

One of the major challenges with network analysis, particularly when multiple species are involved, is drawing a distinction between edges in a network that point to actual instances of interaction, cross talk, or metabolite exchange between species and edges that merely reflect a similar response between two genes across conditions. Answering this question will allow for a much greater amount of relevant biological information to be gleaned from a network analysis that examines multiple species. I have applied several different analytical techniques to such networks to try to highlight which edges may point to true interactions. First, as metatranscriptomics usually requires a metagenome (though *de novo* assembly is possible), some information about the genomic potential of each constituent of a multispecies system can be gained. Edges that link processes that are involved in biosynthesis of certain metabolites and vitamins in one species (metabolic pathways) with processes that are involved with obtaining these molecules in another species (transport and uptake systems) likely point to true interactions. In this scenario, one species makes a molecule and the other takes it up. Networks that can identify these instances of correlation are powerful ways to generate strong hypotheses for downstream experiments, possibly using labeled carbon sources or tagged metabolites to track the flow of carbon and nitrogen through a community, ideally from one species to another as the network predicts.

COMBINING MULTIPLE COEXPRESSION MEASUREMENTS IN A SINGLE NETWORK

I have also looked at the direction of correlation of the gene pair linked by an edge. While mutual information does not provide this information, other analyses such as Pearson or Spearman correlation coefficient can. By looking at similar processes whose genes are negatively correlated between two species, it is possible to identify which processes might be primarily carried out in one organism with the products being shared with other species in the system. The downregulation of certain metabolic processes in one species while another species upregulates these same processes points to this sharing. It is energetically more favorable to obtain nutrients from a partner species than synthesize them yourself. Another approach that other researchers have explored is the collection of time-series data to infer the sequence of events within interactions. The biosynthesis of a needed metabolite by one species preceding the increased expression of the uptake processes for this metabolite by another species is stronger evidence of interaction and sharing than the observation of both of these events lacking a time element (12). Other studies have also looked at metabolic pathways in microbial systems by linking species to metabolites to identify which metabolites may be related to certain microbial community structures (13).

CHALLENGES WITH INFERRING NETWORKS OF MULTI-OMIC DATA

Delineating true interactions from network noise is one challenge facing multispecies networks. Another is the incorporation of omics data from multiple species into a single network. The response of a community is comprised of contributions from each constituent member of the community with the interactions between them being

Downloaded from http://msystems.asm.org/ on September 21, 2019 by guest

critical to the community response. Because of this, networks that model community behavior should include as many interactions as possible from different species. However, our group has found that inferring networks that include, for example, transcriptomic data from multiple species often leads to very segregated networks with clusters of transcripts from each species separated from each other and very few edges linking transcripts across species. I have explored a number of alternative network inference tools aside from CLR and Pearson correlation coefficient and have found that a random forest approach, GENIE3, is particularly well suited to inferring networks that link transcripts across species with no loss in overall network robustness or accuracy. Other researchers have also used this approach for interrogating networks, and studies that rank network inference tools have found that GENIE3 is the most accurate choice when linking *Escherichia coli* target and regulator pairs (14). Whether GENIE3 is the proper tool for multispecies network inference generally remains to be seen, but inferring integrated networks that better link species within a microbial community will be crucial to creating the best models possible of these systems. The observation that GENIE3 is well suited to creating integrated networks suggests that researchers may want to focus on random forest inference methods when building networks of multispecies systems or, at the very least, explore several inference methods before deciding on an approach, as there appears to be great variability in how inference methods fit characteristics of a certain experiment (multispecies, multi-omic, proteomics, transcriptomics, etc.). The strong segregation of edges in multispecies systems also suggests that when looking for correlation between species, these instances of cross-species correlation may be "drowned out" in effect by much higher levels of correlation between genes within a species. Whole-scale ranking of all instances of correlation (both those within and those across species) may miss certain correlations that are high compared to other correlations across species but low compared to correlations within species. Ranking each of these types of correlations separately may be a better strategy for predicting interspecies interactions.

CONCLUSIONS AND FUTURE DIRECTIONS

Interspecies interactions within microbiomes are extremely complex, and identifying interactions and understanding their contribution to the overall community can be very difficult. However, as community response is a culmination of specific interactions, their delineation is critical. With the large increase in the availability of omics analysis, we now have the tools to use network inference to predict and understand interspecies interactions in natural microbial communities. Some of my own work and other recent studies have shown that such networks are powerful ways to view how processes are related between species. They can predict where metabolites might be shared between species or where there may be instances of division of labor and metabolic coordination by looking for instances of coordinated production and uptake. Through the application of centrality, they can also be used to identify genes, and thus processes, that are particularly important to a system. The success of networks that have been seen with simple systems (phototroph/heterotroph, host/pathogen) shows the utility of this approach, and by understanding the challenges and where the pitfalls are, we are well positioned as a community to apply network techniques to more complex systems comprised of many species. Moving forward, it will be critical to ensure that the right network method is being applied to the right experiments. We can see here that many inference approaches have particular strengths over others, and the correct application of these approaches will be needed to gain the most robust insight into the systems we are interrogating. As we continue to apply network approaches to more complex systems, the hurdles of these more species-rich networks will likely require a greater incorporation of metagenomics and metabolomic data, either knowledge of certain pathways (KEGG databases) or the collection of metabolomics, to query what molecules are present. Over the next decade, I expect to see several new network studies that specifically attempt to link not only species but specific transcripts and proteins across species in complex systems such as soil, marine environments, or human anatomical

Downloaded from http://msystems.asm.org/ on September 21, 2019 by guest

sites. The increase in metagenomic data as well as critical work being done on how to use network inference tools for these complex multispecies systems will make multispecies networks easier to infer, interrogate, and interpret. The very detailed and specific knowledge that multispecies networks can offer opens up the possibility for increased understanding of a number of microbiomes, allowing us to modify and harness them to improve human health and wellbeing.

ACKNOWLEDGMENTS

This research was supported by the Department of Energy Office of Biological and Environmental Research (BER) and is a contribution of the Scientific Focus Area "Phenotypic response of the soil microbiome to environmental perturbations." PNNL is operated for the DOE by the Battelle Memorial Institute under contract DE-AC05-76RLO1830.

REFERENCES

1. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, Lasorella A, Aldape K, Califano A, Iavarone A. 2010. The transcriptional network for mesenchymal transformation of brain tumours. Nature 463:318–325. https://doi.org/10.1038/nature08712.
2. McDermott JE, Diamond DL, Corley C, Rasmussen AL, Katze MG, Waters KM. 2012. Topological analysis of protein co-abundance networks identifies novel host targets important for HCV infection and pathogenesis. BMC Syst Biol 6:28. https://doi.org/10.1186/1752-0509-6-28.
3. McDermott JE, Taylor RC, Yoon H, Heffron F. 2009. Bottlenecks and hubs in inferred networks are important for virulence in Salmonella typhimurium. J Comput Biol 16:169–180. https://doi.org/10.1089/cmb.2008.04TT.
4. McClure RS, Overall CC, McDermott JE, Hill EA, Markillie LM, McCue LA, Taylor RC, Ludwig M, Bryant DA, Beliaev AS. 2016. Network analysis of transcriptomics expands regulatory landscapes in Synechococcus sp. PCC 7002. Nucleic Acids Res 44:8810–8825. https://doi.org/10.1093/nar/gkw737.
5. Song HS, McClure RS, Bernstein HC, Overall CC, Hill EA, Beliaev AS. 2015. Integrated in silico analyses of regulatory and metabolic networks of Synechococcus sp. PCC 7002 reveal relationships between gene centrality and essentiality. Life (Basel) 5:1127–1140. https://doi.org/10.3390/life5021127.
6. Chou WC, Cheng AL, Brotto M, Chuang CY. 2014. Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer. BMC Genomics 15:300. https://doi.org/10.1186/1471-2164-15-300.
7. Layeghifard M, Li H, Wang PW, Donaldson SL, Coburn B, Clark ST, Caballero JD, Zhang Y, Tullis DE, Yau YCW, Waters V, Hwang DM, Guttman DS. 2019. Microbiome networks and change-point analysis reveal key community changes associated with cystic fibrosis pulmonary exacerbations. NPJ Biofilms Microbiomes 5:4. https://doi.org/10.1038/s41522-018-0077-y.
8. Quinn RA, Whiteson K, Lim YW, Zhao J, Conrad D, LiPuma JJ, Rohwer F, Widder S. 2016. Ecological networking of cystic fibrosis lung infections. NPJ Biofilms Microbiomes 2:4. https://doi.org/10.1038/s41522-016-0002-1.
9. McClure RS, Overall CC, Hill EA, Song HS, Charania M, Bernstein HC, McDermott JE, Beliaev AS. 2018. Species-specific transcriptomic network inference of interspecies interactions. ISME J 12:2011–2023. https://doi.org/10.1038/s41396-018-0145-6.
10. Musungu BM, Bhatnagar D, Brown RL, Payne GA, OBrian G, Fakhoury AM, Geisler M. 2016. A network approach of gene co-expression in the Zea mays/Aspergillus flavus pathosystem to map host/pathogen interaction pathways. Front Genet 7:206. https://doi.org/10.3389/fgene.2016.00206.
11. Tierney L, Linde J, Muller S, Brunke S, Molina JC, Hube B, Schock U, Guthke R, Kuchler K. 2012. An interspecies regulatory network inferred from simultaneous RNA-seq of Candida albicans invading innate immune cells. Front Microbiol 3:85. https://doi.org/10.3389/fmicb.2012.00085.
12. Barman S, Kwon YK. 2018. A Boolean network inference from time-series gene expression data using a genetic algorithm. Bioinformatics 34:i927–i933. https://doi.org/10.1093/bioinformatics/bty584.
13. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, Graeber TG, Sonnenburg JL, Horvath S, Huttenhower C, McGovern DP, Fornace AJ, Jr, Borneman J, Braun J. 2013. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome 1:17. https://doi.org/10.1186/2049-2618-1-17.
14. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, DREAM5 Consortium, Kellis M, Collins JJ, Stolovitzky G. 2012. Wisdom of crowds for robust gene network inference. Nat Methods 9:796–804. https://doi.org/10.1038/nmeth.2016.

Downloaded from http://msystems.asm.org/ on September 21, 2019 by guest