Building bacteria-phage interaction networks using the CRISPR locus

Madeleine Bonsma[1] and Sidhartha Goyal[1]

Bacteria-phage interaction networks provide an important window into the functioning and ecology of microbiomes. Here we utilize the CRISPR locus to build and analyze the structure of such networks. We find that a CRISPR-derived network in Human Microbiome Project data is both nested and modular, and a network constructed from the CRISPRdb database is modular.

Prokaryotes and their phage predators are abundant in many environments and can strongly impact their environments. Importantly, recent evidence links bacteria in the human microbiome to such phenomena as obesity[1], cancer[2], and immune disorders[3]. Phages influence their hosts in turn, producing population-level effects such as gene transfer[4] and mediation of pathogenic bacteria outbreaks[5], an effect exploited in phage therapy[6]. The interpretation of phage-bacteria interaction networks has wide-ranging implications for understanding the role of microbiomes in their environments.

Recently discovered prokaryotic adaptive immune system CRISPR-Cas[7–10] provides another window into bacteria-phage networks. Bacteria and archaea that possess a CRISPR-Cas system can develop a memory of past phage infections by incorporating small samples of phage DNA, called *spacers*, into a specific CRISPR locus. Many spacers can be stored in a CRISPR locus: up to 587 spacers have been documented in *Haliangium ochraceum*, but most CRISPR loci contain fewer than 50 spacers. Unlike previous which largely consisted of direct experiments with cultured bacteria and phages (Ref [[11]] compiles 38 such studies). This method, while yielding detailed and accurate results, is time-consuming to the point of being unfeasible for large networks. As well, only a small fraction of the microorganisms in natural environments can be cultured in a laboratory at all[12], meaning that a significant portion of microbial ecosystems remains inaccessible by this technique.

In this work, we propose and demonstrate both large-scale and small-scale phage-bacteria interaction networks constructed using the information contained in the CRISPR locus. Many species of bacteria and archaea possess the CRISPR system: according to the database CRISPRdb[13], 84% of archaeal genomes analyzed (126/150) and 45% of bacterial genomes analyzed (1176/2612) possess at least one CRISPR region. To the extent that CRISPR-Cas is utilized in a bacterial strain, the CRISPR locus provides a detailed snapshot of phage interaction history, which can be used to construct an interaction map. Displaying bacteria-phage relationships in this way facilitates comparison to previous experimental infection studies and opens the door to ecological analysis of microbiomes using existing network analysis metrics such as modularity (how well a network can be divided into subgroups) and nestedness (to what extent the interaction ranges of members are subsets of other interaction ranges)[11,14–16].

We construct CRISPR-based networks by aligning spacers using BLAST to a compilation of virus and phage databases and recording high-scoring matches. The results are subjected to the same analysis metrics for nestedness and modularity as the traditional experimental infection matrices. The CRISPR networks constructed here exhibit modularity on large scales, consistent with previous work[11]. Clustering between sub-groups of bacteria and phage is potentially indicative of ecologically distinct groups of interacting bacteria and phages.

CRISPR-based networks require much less experimental effort to construct than experimental infection studies. Additionally, CRISPR data can be extracted from metagenomic data with existing approaches[13,17–19] and used to build networks that more accurately capture interactions between bacteria and phages that cannot be lab-cultured. We explore this approach with samples from the Human Microbiome Project[20] using Crass[17] to extract candidate repeats and spacers.

Our analysis shows that a bacteria-phage interaction network in Human Microbiome Project data is nested and modular to a greater degree than two null model datasets, one with random interactions sampled from a Gaussian distribution, and one with the same number of interactions shuffled into random positions. We also find modularity in a large network constructed using CRISPR locus data from the database CRISPRdb[13]. This work shows promise as a method of investigating bacteria-phage interaction networks.

Acknowledgements: much of the code used to analyze data was contributed through open-source collaborations – see collaborators list. https://github.com/goyalsid/phageParser/graphs/contributors

[1]Department of Physics, University of Toronto, Toronto, Canada.

References

1. Turnbaugh, P. J. *et al. Nature* **444,** 1027–1031 (2006).
2. Schwabe, R. F. & Jobin, C. *Nat. Rev. Cancer* **13,** 800–12 (2013).
3. Mazmanian, S. K., Cui, H. L., Tzianabos, A. O. & Kasper, D. L. *Cell* **122,** 107–118 (2005).
4. Sano, E., Carlson, S., Wegley, L. & Rohwer, F. *Appl. Environ. Microbiol.* **70,** 5842–5846 (2004).
5. Faruque, S. M. *et al. Proc. Natl. Acad. Sci. U. S. A.* **102,** 6119–6124 (2005).
6. Levin, B. R. & Bull, J. J. *Nat. Rev. Microbiol.* **2,** 166–173 (2004).
7. Bolotin, A., Quinquis, B., Sorokin, A. & Dusko Ehrlich, S. *Microbiology* **151,** 2551–2561 (2005).
8. Pourcel, C., Salvignol, G. & Vergnaud, G. *Microbiology* **151,** 653–663 (2005).
9. Makarova, K. S., Grishin, N. V, Shabalina, S. a, Wolf, Y. I. & Koonin, E. V. *Biol. Direct* **1,** 7 (2006).
10. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. *J. Mol. Evol.* **60,** 174–182 (2005).
11. Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. *Proc. Natl. Acad. Sci. U. S. A.* **108,** E288–E297 (2011).
12. Mardis, E. R. *Trends Genet.* **24,** 133–141 (2008).
13. Grissa, I., Vergnaud, G. & Pourcel, C. *BMC Bioinformatics* **8,** 172 (2007).
14. Barber, M. J. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **76,** 066102 (2007).
15. Newman, M. E. J. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 8577–8582 (2006).
16. Weitz, J. S. *et al. Trends Microbiol.* **21,** 82–91 (2013).
17. Skennerton, C. T., Imelfort, M. & Tyson, G. W. *Nucleic Acids Res.* **41,** e105 (2013).
18. Edgar, R. C. *BMC Bioinformatics* **8,** 18 (2007).
19. Bland, C. *et al. BMC Bioinformatics* **8,** 209 (2007).
20. Peterson, J. *et al. Genome Res.* **19,** 2317–2323 (2009).