Modeling Phage Species Abundance

David Bangor[1], Beltran Rodriguez Brito[2], Peter Salamon[3], James Nulton[4], Ben Felts[5], Joe Mahaffy[6], Mya Breitbart[7], Forest Rohwer[8]

Keywords: Lander/Waterman, viruses, phage, biodiversity, metagenomes

Bacteriophage (phage) are viruses that infect bacteria. Once inside their host bacteria, phage rapidly multiply and eventually kill the host by causing it to explode and release the new phage particles. These free phage particles are the most abundant biological entities in the biosphere, with and estimated 10^{31} phage particles on the planet. The total number of phage species, however, is essentially unknown. Recently we have started to shotgun sequence the DNA from uncultured phage communities. From this data, we obtained overlapping fragments, which means that the same phage genome has been resampled. This observation allows us to mathematically model the phage community.

Previous studies of phage diversity have been performed using culture-based methods [1]. This involves first culturing a bacterial host from the environment, and then isolating phage that can infect that bacterial host. Unfortunately, most bacterial hosts (>99%) cannot be cultured and not all phage produce identifiable plaques. To circumvent these limitations, we developed a method to shotgun sequence DNA from uncultured phage communities. To do this, total genomic DNA was isolated from natural phage communities containing approximately 10^{12} particles. The genomic DNA was physically sheared into 1-2 kilobase long fragments. Then the fragmented DNA sequences were cloned and sequenced. Finally, the DNA sequences were analyzed using Sequencher [2] to identify sequences that overlapped with 98% identity over 20 bp. An overlap between sequences means that the same genome has been re-sampled. A contig spectrum was created from the Sequencher analysis, where a 1-contig means the sequence had no overlaps, a 2-contig means two sequences overlapped, a 3-contig means three sequences overlapped, etc... The contig spectra from 4 environmental samples were then used to predict the population structure of the phage communities using a modified Lander/Waterman algorithm [3, 4, 5, 6].

In the current analysis a number of different distributions were compared. The first two models are the Broken Stick and the Niche Preemption [7], which are ecological models based upon the division of resources into niches. The other models were the common empirical functional forms - Power Law, Exponential Law, Logarithmic, and Lognormal. To make the comparisons, the contig spectrua obtained from the samples were used to model the populations assuming one of the 6 functional forms. The error between the predicted and the actual contig spectra were determined.

[1]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: heimdalle@yahoo.com

[2]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: brodrigu@rohan.sdsu.edu

[3]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: salamon@saturn.sdsu.edu

[4]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: jnulton@mail.sdsu.edu

[5]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: bfelts@myth.sdsu.edu

[6]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: mahaffy@math.sdsu.edu

[7]Department of Biology, San Diego State University, San Diego, California, 92182-4614.E-mail: mya@sunstroke.sdsu.edu

[8]Department of Biology, San Diego State University, San Diego, California, 92182-4614.E-mail: forest@sunstroke.sdsu.edu

Table 1 shows that the Power Law and the Lognormal best described the observed contig spectra. In contrast, the ecological models did a very poor job of explaining the observed data.

	Scripps Pier	Mission Bay	Fecal Data	Mission Bay Sediment
Power Law	1.81	2.11	9.20	0.0104
Exponential Law	12.1	16.2	59.9	0.0126
Logarithmic	2.51	2.81	10.3	0.0104
Broken Stick	10.7	14.6	51.6	0.0156
Niche Preemption	29.5	38.1	145	ND
Lognormal	1.89	2.31	9.66	0.0104

Table 1: Errors for the given species abundance model and the environmental sample above. ND = not determined

In the environmental samples the phage community structure is best described by a Power Law or Lognormal distribution. This does not mean that all phage community structures will be a Power Law or Lognormal distribution, but it does give some idea of what the actual mathematical distribution may look like. The distinction between the phage community falling under actual Power Law distribution or a Lognormal distribution needs to be determined in the future because the absolute number of species calculated by the Lognormal is approximately 3X as much as that predicted by the Power Law. The shape of the community distribution is also important for modeling how phage and their bacterial host interact. Currently, we are performing higher coverage sequencing to differentiate between these two functions.

References

[4] Breitbart, M., B. Felts, et al. (2004). Diversity and population structure of a nearshore marine sediment viral community. Proceedings of the Royal Society B, in press.

[5] Breitbart, M., I. Hewson, et al. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 85(20): 6220-6223

[6] Breitbart, M., P. Salamon, et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 99(22): 14250-14255.

[3] Lander, E. S. & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2, 231-239.

[7] MacArthur, R. H. (1957). On the relative abundance of bird species. Proc. *Nat. Acad. Sci*. Wash. 43: 293-295.

[2] Sequencher http://www.genecodes.com/

[1] Wommack, K. and R. Colwell (2000). "Virioplankton: Viruses in aquatic ecosystems." *Microbiol Mol Biol Reviews* 64(1): 69-114.