Joshua S. Weitz, School of Biology & School of Physics, Georgia Institute of Technology, Atlanta, GA, USA
Bart Haegeman, INRIA Research Team MERE, UMR MISTEA, Montpellier, France

A Neutral Theory of Genome Evolution and the Frequency Distribution of Genes

The gene content of genomes of closely related bacteria can differ significantly. For example, pair-wise comparisons of genome sequences from isolates of the same species often do not share a substantial fraction of their gene content. When a large number of genomes within a bacterial "species" are sequenced, the gene content variability can be summarized as a gene frequency distribution: given G sequenced genomes, some genes are found in k of these genomes where k ranges from 1 to G. Empirically, such gene frequency distributions possess a characteristic U-shape, such that there are many genes that only appear in one genome, fewer genes which appear in an intermediate number of genomes, and many genes which appear in all genomes. Genes within each of these three categories have been labeled accessory, character and core genes, respectively[1]. It would seem that U-shaped gene frequency distributions can be used to infer the essentiality and/or importance of a gene to a species. Instead, we ask: is it possible to recapitulate findings of U-shaped gene frequency distributions in the absence of selective forces driving genomic and population composition?

Here, we answer this question in the affirmative by proposing a simple and analytically tractable neutral model of genome evolution that explicitly accounts for gene composition of genomes. In this model, genomes undergo birth-death processes in a neutral sense and also acquire and lose genes. The model differs from most previous efforts to analyze genome evolution by self-consistently treating the dynamics at two scales: population level drift and genomic level change. We analyze our model using coalescent theory and derive closed form solutions for gene frequency distributions. We find that gene frequency distributions in the model possess a characteristic U-shape even in the absence of selective forces driving genome and population structure. We fit model predictions to empirical data from 6 bacterial pathogens: *B. anthracis*, *E. coli*, *Staph. aureus*, *Strep. pneumoniae*, *Strep. pyogenes* and *N. meningitides*, using a bioinformatics pipeline for assessing gene-genome composition[2]. In so doing, we find a reasonable correspondence between our neutral model and data from six distinct bacterial species with sequenced genomes from multiple isolates. However, our model assuming constant population sizes predicts gene frequency distributions with systematically fewer rare genes than the empirical distributions.

Hence, we also consider variations to our base model. These variations include cases of constant and exponentially growing population sizes as well as two alternative models which contain a "rigid" and "flexible" core component of genomes. All of these models can improve fits to empirical distributions. In addition, all of these models make a number of other predictions regarding the scaling of sample core and pan genome sizes in accord with observations. Together, these models suggest that U-shaped gene frequency distributions provide less information than previously suggested regarding gene essentiality. Hence, we discuss the need to find patterns of genome composition variation other than gene frequency distributions that can be explained by neutral models and identify those patterns or deviations from patterns that cannot be explained by neutral models. In addition, we briefly highlight the need for additional theory to disentangle the roles of evolutionary mechanisms operating within and amongst individuals in driving the dynamics of gene distributions.

References

1. Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. *Trends in Genetics* 25:107–110.
2. Kislyuk AO, Haegeman B, Bergman N, Weitz JS (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* 12:32.