Order from Chaos: Reconstructing Microbiomes and Metagenomes from Massive Amounts of Sequence Data

Andrew K. Benson

Department of Food Science and Technology, University of Nebraska-Lincoln

Microbiomes (collections of microbial species that are unique to specific ecosystems) numerically dominate soil, water, intestinal and other environments, and have attracted significant effort aimed at defining and manipulating them. Pioneering work in this field focused on sequencing 16S rRNA gene that is conserved among all Bacteria and creating large databases of sequences from known organisms that can be referenced to sort through a diverse microbial population. Advances in Next Generation DNA sequencing technologies (e.g. pyrosequencing) now make it routine for investigators to query composition of complex microbiomes by sequencing thousands to hundreds of thousands of 16S rRNA gene sequences from DNA extracted from environmental, clinical, or even food samples. This approach has already found many applications in medicine, agriculture, and the environment. These analyses are performed by deep pyrosequencing of PCR products amplified from the 16S rRNA gene; subsequently, each complete 16S rRNA sequence read must be assigned a taxonomic status by comparing it to a database of sequences from known and previously classified taxa. Word-based approaches, such as the CLASSIFIER algorithm, have developed to be very efficient at this task but suffer from the fact that 50% of more of the sequences from any given environment are unique and not present in the database, leaving us blind to classifying and quantifying significant portions of the microbiome. To circumvent this problem, alignment and distance estimates between millions of sequence reads can be used to groups similar sequences into Operational Taxonomic Units (OTUs). However, the hundreds of millions of computational operations that are required by this so-called OTU-picking approach make it computationally daunting, and current algorithms simply cannot handle such a volume of data. At the same time, microbiome studies are moving beyond descriptions of different microbiome compositions to quantifying factors that control it demanding high-throughput computational methods and databases to process and warehouse microbiome data. Furthermore, the blossoming of new capacities in microbial ecology has recently produced an explosion of interest in metagenomics, the study of the collective genetic capacity of all organisms in an ecosystem. Metagenomes are assembled from shotgun sequencing of environmental DNA samples, and contain sequence reads from fragments of all species in a given environment. This approach provides a less biased and more comprehensive view of both species content and genetic content than does the microbiome approach. Computationally, metagenome analysis depends on accurate assembly of shotgun sequences from random fragments in a complex mixture of organisms, a task that is exceedingly more complex than assembling a single genome. Moreover, as sequencing platforms progress in their capacity, the computational challenges will also progress from assembling complex and massive jumbles of random sequences into individual contigs to assembling these individual contigs into whole genomes. This challenge will be accompanied by a growing demand for high-throughput gene-finding and functional assignment.

Biosketch:

Dr. Andrew Benson received his Ph.D. in Microbiology in 1992 from the University of Texas Health Science Center at San Antonio. He is currently a Professor at the Department of Food Science and Technology and the Director of the Core for Applied Genomics and Ecology at University of Nebraska-Lincoln.