nature structural & molecular biology

Sequencing the microbial soup

Metagenomics has the potential to shed light on one of the least understood group of organisms on Earth, while also providing useful tools for many fields in biology.

Microorganisms are the most diverse and abundant life forms on the planet and have had a major role in the evolutionary success of multicellular organisms and life as it presently exists. The existence of microbes was discovered more than 300 years ago, thanks to the pioneering work of Anton van Leeuwenhoek, which marked the emergence of the field of microbiology. Despite their importance, microorganisms are still poorly understood, largely because of the inability to generate cultures of many microbial species. This limitation has allowed less than 1% of the total number of microbial species (estimated to be in the millions) to be studied in typical culture-dependent laboratory settings and has prevented a full understanding of their metabolic potential and contribution to the biosphere.

Now, work has been directed towards filling this large knowledge gap. This emerging field, called metagenomics, examines microbial communities as a whole rather than examining a cultured isolate of a single species. Because species within a community are interdependent—often a single species is capable of catalyzing only one step in a critical biosynthetic pathway—metagenomic analysis has the potential, for starters, to give insight into the capabilities of an entire community as well as how individual organisms serve the needs of the whole group.

Metagenomics is now possible because of better and cheaper sequencing technology, such as improved shotgun Sanger sequencing and pyrosequencing. These methods make it possible to reproduce megabases of DNA, even when very little of it is present in a sample. This is perfect for microbial populations, bypassing the need for cell cultures to generate abundant quantities of DNA for standard sequencing studies. In addition, faster computers with improved data storage capabilities make it easier to manipulate and piece together the vast number of sequence fragments being generated.

Many metagenomics projects begin with isolation of DNA collected from a sample environment (for example, soil, sewage, seawater or intestinal lining). The sample DNA can be sequenced directly or cloned into a DNA library sustainable in a bacterial strain. Generation of a metagenomic library can be bypassed if pyrosequencing technology is used, but the use of bacterial artificial chromosomes (BACs) has made it easier to clone and assemble these large libraries. Regardless of the technique, the overall goal is to generate sequence fragments that can be patched together to reflect the underlying genomes being sampled.

The metagenomes can be studied in sequence-based or function-based approaches. Sequence-based approaches can make use of large-scale sequencing of the metagenome, and may sometimes provide sequencing of an entire genome. Surveys of 16S ribosomal RNA genes present in the sample are frequently used to reveal the genetic diversity within a given population, which may not be fully realized through studies of organisms isolated and cultured from that same sample. Some metagenomes are simply too large to assess and targeted sequencing is used to test for the presence and abundance of particular functional genes, which would give information about the needs of the community being studied. Function-based approaches use bacterial translation machinery to produce the gene products from a metamicrobial DNA library. The proteins are then screened for activity, for example, vitamin production or resistance to antibiotics. From this, scientists can assess the capabilities of a community without having prior knowledge of the gene sequence or the organism from which it came.

There are, of course, hurdles that must be overcome. Isolation of metagenomic DNA often requires the use of a single purification technique that can either be too mild or too harsh for some members of the population, so that cells may not be lysed or DNA can be damaged or sheared, resulting in underrepresentation of certain species. In addition, metagenomic sequencing itself is more prone to errors, as it relies on less redundancy, often with only one fragment representing an underlying gene sequence. These projects generate much more data at the DNA as well as predicted protein and functional levels than standard genome projects, all of which must be made accessible in public databases that can be annotated and that can also house the metadata associated with the project. These are just a few of the many challenges facing existing and future metagenomic undertakings.

What insights can be gained from these metagenomic studies? One of the largest completed sequencing studies was on surface-water samples collected from the Sargasso Sea (Venter, J.C. *et al.*, *Science* **304**, 66–74, 2004), which yielded roughly 1.6 Gbp of sequence and identified more than 1.2 million genes from 1,800 species, including 148 new species. On page 177 of this issue, Batey and colleagues examine one Sargasso Sea gene, of the *S*-adenosylmethionine (SAM)-II riboswitch, an RNA regulatory element important for sulfur metabolism in α-proteobacteria. The data indicate that, although the overall architecture of the SAM-II riboswitch differs dramatically from that of the Gram-positive bacteria SAM-I riboswitch, they use similar chemical means to distinguish between cognate and noncognate ligands, thus providing the first example of convergent functional evolution in RNAs.

Many of the metagenome projects examine microbes that occupy niches in extreme environments—the Sargasso Sea is nutrient poor, and other projects examine microbial communities present in acid mine drainages and antibiotic-polluted soil, for example. These same species probably colonized different environments during Earth's history, allowing multicellular organisms to thrive later on. By examining these populations, we can learn how their metabolisms allow them to cope with such hostile conditions and how evolution has found different solutions for the same problem, which can potentially provide breakthroughs in biotechnological applications and insights for environmental and health sciences. ■