Metagenomics versus Moore's law

Metagenomics sprang from advances in sequencing technology, and continued improvements are providing data in quantities unimaginable a few years ago. But without concerted efforts, the amount of data will quickly outpace the ability of scientists to analyze it.

As Craig Venter sails the oceans collecting seawater samples to profile microbial communities by high-throughput sequence analysis, microbiologists around the world are busy collecting their own samples. The diversity of locations—from Antarctic lakes to human armpits—highlights the reality that microscopic organisms represent a significant fraction of the Earth's ecosystem.

Any population this large is certain to have profound influences on its environment. Yet our knowledge of these communities and their functions is rudimentary, partly owing to our inability to culture the vast majority of microbes. The arrival of high-throughput capillary Sanger sequencing in the 1990s, with its ability to analyze collective microbial genomes, opened a new window onto these communities and spawned the field of metagenomics.

Metagenomics takes a number of forms depending on the question being asked and the available resources. Information on phylogenies and frequency of community members can be obtained by targeted sequencing of a reference gene such as the 16S rRNA. Entire genomes of common community members can be assembled by shotgun approaches if the sequencing is deep enough, but most analysis relies on contigs assembled from a few reads. These are sufficient for gene identification, and classification of the contigs can still provide community phylogenies.

Microbiologists are now flocking to second-generation sequencing platforms that provide orders of magnitude more sequence per dollar than the Sanger technique. But although these platforms provide enormous amounts of data, their use comes with challenges. Notably, read lengths drop drastically compared to Sanger sequencing—by about 50% for pyrosequencing and 90% for Illumina and SOLiD platforms. Most metagenomics analysis pipelines are designed for Sanger sequencing data, so the short read lengths and error profiles of the new methods present challenges for data analysis and interpretation.

Reports on pages 639 and 673 and an accompanying News and Views on page 636 illustrate some of the dangers and challenges involved and describe new algorithms to deal with them. More work is needed to assess the new technologies and develop optimized analysis pipelines, and these efforts are well underway.

But even as these problems are being solved, a larger problem has taken the community off-guard: the exponentially increasing amount of sequence data. Just over three years ago, the first two second-generation sequencing platform–based shotgun metagenomes were reported—each less than 40 megabases. Today there are over 4,000 sequenced metagenomes, and their size and number are increasing. Each new pyrosequenced metagenome is 200–500 megabases, and those generated on Illumina platforms are 20–50 gigabases. To analyze these metagenomes using established pipelines would take tens of years on a single processor and weeks to months on machines with up to 1,000 processors. The rate of increase in sequence generation is far outpacing Moore's law, and the cost of analyzing the largest datasets already exceeds the cost of generating them.

Analysis of new metagenomes requires assembly, gene prediction and other computationally intensive operations. The August update of the Integrated Microbial Genomes database will have 6.5 million genes, and integrating this with existing metagenomes will give 25 million genes. Some projections suggest we will reach 250 million genes in two years. At current database sizes all-versus-all comparisons are already impossible without a supercomputer. Development of more efficient algorithms will help, but this will not solve the basic problem of too little computing power. Individual access to supercomputers or cloud computing would help, at least temporarily.

Ultimately, major initiatives are needed to avoid metagenome-analysis gridlock. The publicly available MG-RAST (metagenome rapid annotation using subsystem technology) service, which provides automatic annotation of metagenomes, is an excellent start but it does not solve the wider problems. Two things need to happen. Firstly, funding agencies need to realize that biology is closing in on fields like physics in terms of dataset sizes and computational demands; this calls for increased support for data analysis. Secondly, the community needs to decrease computational demands by improving data sharing through standards and centralized coordination and by aggregating computationally intensive operations.

This summer, after discussions at the International Conference on Systems for Intelligent Molecular Biology, community members formed the M5 (metagenomics, metadata, metaanalysis, multiscale-models and metainfrastructure) Consortium under the roof of the Genomics Standards Consortium to devise a solution to the coming gridlock. Their proposed 'M5 Platform'—to be announced later this year—deserves the support of the community, funding agencies and those who hold the keys to the high-performance computing centers. Unless major efforts are taken immediately, researchers will find they have a wealth of data but no way to interpret it.