Introduction

Genome Map

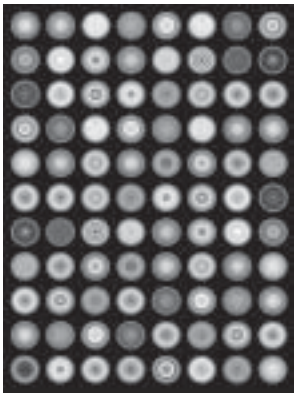
Genomic science is essential for understanding all levels of biological organization from single cells, to the biosphere, and beyond. This will become increasingly apparent as the emphasis in the genomics field shifts from primarily one of generating DNA sequence information to data management and interpretation. There are several ways in which genomics can impact our understanding of microbial life on Earth: (1) Whole genome sequencing of organisms that can be grown in culture has already demonstrated its power in revealing the biochemical and metabolic capabilities of these species; (2) partial genome sequencing can be used to interrogate various environments to provide a preliminary catalog of information on gene identity and function; (3) functional genomics approaches (genome-enabled science) can now begin to elucidate how individual cells and communities work together in an integrated fashion; and (4) databases and bioinformatics tools can allow investigators to fully exploit ever-increasing amounts of genome data to generate new hypotheses about the relationships between organisms and the properties of cells and communities; promote advances in the biotechnology industry, in food production, and in drug design; and understand the origin of life and how it continues to evolve.

Microbes of agricultural importance, animal pathogens, and microbial extremophiles are largely underrepresented in the more than 120 microbial sequencing projects that have been completed or funded.

More than 120 microbial genome sequencing projects have been completed or funded. The organisms represented in this list are heavily weighted toward human pathogens, with smaller numbers of non-pathogenic bacteria and archaea included. Superficial phylogenetic coverage of the major bacterial groups has been achieved in the selection of species for whole genome analysis; however, microbes of agricultural importance, animal pathogens, and species that inhabit extreme environments have been underrepresented. Moreover, there have been no projects funded to date to tackle unculturable microbial species, which presumably represent the greatest diversity in the microbial world. For example, less than 1% of the bacterial species in the oceans have been cultured, and there are roughly a million of these unknown cells per ml of seawater.

DNA sequencing technology has evolved rapidly over the past ten years and the methods for generating sequence have become faster and more efficient. As a result, the costs for generating a finished DNA sequence has dropped from ~$1 per base pair in the early 1990s to less than $0.20 per finished base pair in large sequencing centers. DNA sequencing costs are far less of an obstacle in identifying organisms for genome analysis than they were even three or four years ago, and as a result additional species and organisms with larger genomes (>4 million base pairs) have now been targeted for sequencing.

Identification of genes in prokaryotic genomes has advanced to the stage where nearly all protein coding regions can be identified with confidence. Computational gene finders now routinely find over 99% of protein coding regions and RNA genes. Once the protein coding genes are located, the most challenging problem is determining their function. Typically, about 40-60% of



DNA Microarray Sequencing data

the genes in a newly sequenced bacterial genome display detectable sequence similarity to protein sequences whose function is at least tentatively known. This sequence similarity is the primary basis for assigning function to new proteins, but the transfer of functional assignments is fraught with difficulties. The remainder of the genes in a completed sequence don't resemble any genes of known function; in those cases, sequence alone doesn't reveal anything about the biological role of the protein product. Better databases and database management tools would go a long way towards maximizing the impact of DNA sequence data that is accumulating in public repositories. The field of bioinformatics – the marriage of biology and computer science – has exploded in recent years out of necessity. However, there has been little coordinated effort among groups that are involved in bioinformatics research, and as a result, no standards for genome annotation, curation, and database structure have emerged. As the amount of genome information continues to grow, it will become increasingly difficult to navigate all of the existing databases that have appeared.

"There has been little coordinated effort among groups involved in bioinformatics research, and as a result, no standards for genome annotation, curation, and database structure have emerged."

Complete microbial genome sequences represent just the beginning of a new field of genome-enabled science. Comparative genomics and *in silico* studies have begun to reveal insights into protein function and regulation that would not have been possible just a few years ago. It is now possible to think about biological investigations at the whole genome level. For example, with microarray analysis every gene in a microbial genome can be represented on a glass slide. Using this kind of approach, investigators can interrogate the level of expression of all genes in a species under various experimental conditions. This kind of approach harnesses the power of genomic technology in a way that investigators could not dream of only ten years ago. Other approaches such a two-hybrid analysis or genome-wide knock-out experiments are also providing large-scale biological insights into the workings of microbial cells.

The microbial genomics field is at an important crossroads. Our vision of the future must be built upon today's efforts to generate complete genome sequence information, to develop databases in which to store and make available information from genomics and functional genomics projects, and to develop novel methods for follow-up biological work on a genome-wide scale. While progress in the genomics field has proceeded at a dizzying pace, it is now time to reevaluate where the field is going and how to best achieve the collective goals of the community. Genomic approaches to microbiology have unlocked so many new avenues of investigation. One of the challenges going forward will be how to best leverage available funding to maximize the return on this research investment.