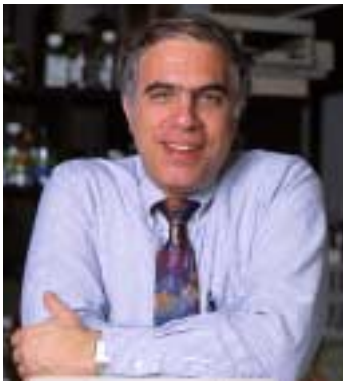
Microbial Genomics: From Sequence to Function



Ira Schwartz
New York Medical College, Valhalla, New York, USA

Guest Editor, Series on Genomics

Dr. Schwartz is professor of biochemistry and molecular biology and medicine. His research focuses on emerging tick-borne infections, primarily Lyme disease and human granulocytic ehrlichiosis (HGE). Studies in his laboratory demonstrated that the infectious agent of HGE was present in ticks collected in 1984—10 years before the first description of clinical cases and human coinfection with *Borrelia burgdorferi* and the agent of HGE. More recently, his laboratory has reported on Lyme disease risk from an individual tick bite in the six lower Hudson Valley, New York, counties and the identification of a *B. burgdorferi* subtype that is more frequently associated with disseminated infection in early Lyme disease patients. Dr. Schwartz is now using genomic approaches to identify genes and proteins involved in *B. burgdorferi* pathogenesis.

The era of genomics (the study of genes and their function) began a scant dozen years ago with a suggestion by James Watson that the complete DNA sequence of the human genome be determined. Since that time, the human genome project has attracted a great deal of attention in the scientific world and the general media; the scope of the sequencing effort, and the extraordinary value that it will provide, has served to mask the enormous progress in sequencing other genomes. Microbial genome sequencing, of particular interest to the community studying emerging infectious diseases, prompted the series of articles presented in the following pages. These articles review technological and scientific advances that have occurred since publication of the *Haemophilus influenzae* genome sequence in July 1995 (1); that was the first demonstration that an entire genome sequence could be deciphered by a "shotgun" approach, i.e., the sequencing and assembly of random fragments of the genome. This is now the method of choice for sequencing of most other genomes, including human (as performed by Celera Genomics).

The articles by Fraser et al. (this issue, pp. 505-12) and by Weinstock (pp. 496-504) briefly describe some of the sequencing methods and annotation of the completed sequences. As of this writing (late June 2000), 23 bacterial genomes have been fully sequenced. More than 70 other microbial genome projects are under way; a regularly updated listing is available on the Internet (http://www.tigr.org/tdb/mdb/mdbinprogress.html). For some species, several strains have been examined, facilitating whole genome comparisons that provide insights not available by other methods (Fraser et al., this issue).

Generally, the first analysis of a completed, fully assembled genome consists of determining all the putative open reading frames (ORFs), which may constitute protein coding regions. These derived amino acid sequences are searched against sequence databases to determine the relationship to previously sequenced genes. There can be three results: a "hit" to a gene of known function, a hit to a gene of unknown function (usually referred to as a conserved hypothetical protein), or no database match. In the first instance, the newly sequenced gene is generally annotated as a homolog of the best hit. When the first bacterial genome sequences were elucidated, it was not surprising that a significant percentage (35%-45%) of identified ORFs either were of unknown function or had no database match. More surprising is that these numbers have not changed substantially as more and more sequences have been determined. Thus, close to half of all bacterial ORFs identified to date have no known function, half of which again are unique to the given

Address for correspondence: Ira Schwartz, Department of Biochemistry and Molecular Biology, New York Medical College, Valhalla, NY 10595 USA; fax: 914-594-3345; e-mail: schwartz@nymc.edu.

species. This represents an enormous storehouse of unrecognized metabolic potential, and it appears obvious that many novel biochemical reactions and pathways are yet to be discovered and characterized. A recent example along these lines is the construction of 6,144 individual yeast strains, each containing an expression clone of all identified yeast ORFs. This strain collection was used for high throughput identification of three previously unrecognized enzymatic activities (2). Similar approaches should provide fertile ground for many future biochemical investigations.

Homology searching and ORF identification have been especially useful in revealing the overall metabolic capability of an organism, identifying potential targets for antimicrobial therapy, and elucidating candidate virulence genes. This information also provides a framework for comparative studies of closely related bacterial species and different strains (e.g., virulent and avirulent) of the same species. Examples of each are provided in the ensuing articles by Fraser et al. and by Weinstock. Despite the obvious value of such analyses, however, some caution must be exercised. Ultimately, the caliber of the bioinformatics tools employed for sequence homology analysis and ORF annotation determines the quality of the data. Unquestionably, errors in annotation exist, and these can result in erroneous classification of newly identified genes. Often, annotated genes are arranged into commonly identified metabolic pathways, and certain key activities are "missing." This may suggest that the organism under study may only contain a portion of the particular pathway. However, given the large reservoir of genes with unknown function, it is equally plausible that another protein has evolved to catalyze the missing reaction. This process has been referred to as non-orthologous gene displacement (3). How common such gene displacement is will only become clear through biochemical studies of the type described above.

Complete bacterial genome sequencing has revealed more extensive genetic exchange between species than suspected. Lateral or horizontal gene transfer has been inferred from differences in guanine-cytosine content or codon preference in specific regions of a genome relative to the entire genome. The best known examples of such lateral exchange are the acquisition of antibiotic resistance genes and pathogenicity islands. The extent of lateral transfer has a profound impact on the inference of phylogenetic relationships by use of specific protein sequences and is the subject of substantial debate (4-6). Fraser et al. touch on these issues and present an approach to phylogeny based on comparative genomics.

Perhaps the greatest value of complete genome sequence information is its use in generating hypotheses that can be further tested by biological ("wet") experiments. Weinstock describes how complete genome sequences may be scrutinized for clues to pathogenic mechanisms and emphasizes that this is merely a starting point for subsequent studies. In many cases, putative virulence determinants can be identified by homology to previously characterized proteins. This approach works reasonably well for pathogens closely related to those that have been extensively studied. However, for many organisms (e.g, the spirochetes *Borrelia burgdorferi* and *Treponema pallidum*), the sequence provides markedly less insight into potential pathogenic processes (7,8). Ultimately, definitive demonstration that any candidate virulence factor plays a role in pathogenesis is best accomplished by genetic manipulation. For example, disruption of a candidate gene should abolish the ability of the mutated pathogen to elicit disease in an animal model of infection, and reintroduction of the wild-type gene should reestablish virulence. For many pathogens, such genetic tools are not yet available, but one hopes that the genetics will soon catch up to the genomics.

One of the most exciting outcomes of the genomics revolution is the ability to probe an organism's global gene expression under a specified set of physiologic conditions. Variously referred to as transcription or gene expression profiling or monitoring, this technology is facilitated by highly parallel analysis of mRNA content in a cell using oligonucleotide chips or cDNA microarrays. The review by Cummings and Relman (this issue, pp. 513-25) describes some technical aspects of the technology and its specific application for the study of host-pathogen interactions. Since the technique is not truly quantitative, it is usually applied to measuring the differences in expression of "all the genes" in the organism under two different growth conditions—for example, environmental (changes in pH or temperature) or nutritional (rich vs. minimal media). The genes that are differentially expressed are assumed to be responsive to the physiologic state of the cell

under these differing conditions. For eukaryotes, this can also be used to uncover differences between normal and diseased cells or tissues. A particularly interesting extension of this approach is the study of alterations in host gene expression on exposure to, or infection with, a bacterial pathogen; this is discussed extensively by Cummings and Relman. Like all enabling technologies, DNA microarray analysis has manifold applications, and new ones will surely be developed. In the context of bacterial genomics, two additional uses are worthy of note. As already described, close to half of annotated ORFs have no known function, and some percentage of these may not be genes at all. Microarray analysis can elucidate the true nature of the "expressed genome" by confirming the expression of genes of unknown function. Of course, as with all experimental data, a positive result is meaningful, but a negative result must be interpreted with caution since a particular "gene" may be expressed only under a very selective set of conditions (perhaps one not amenable to facile experimental analysis). Microarrays can also be employed in high-throughput detection or diagnostic applications based on ribosomal RNA hybridization (9,10).

This series of review articles on genomics, like any other series covering a broad and rapidly evolving area of investigation, cannot provide comprehensive coverage of all topics. For example, information on the development of novel antimicrobial drugs and vaccines based on whole genome sequencing data (11-13) has not been included. Finally, much of the early emphasis in genomics has been on accumulating and annotating raw sequence data. While certain fundamental insights have been gained from these data (e.g., the extent of lateral gene exchange and existence of novel genes), ultimately, the most profound advances will result from using sequence information to drive the study of microbial biology, i.e., how the genome determines function. Expression profiling is the first step along this path. However, this technique measures mRNA, the "message," rather than the protein gene product. Proteomics—describing the complete protein complement of an organism—has thus developed as the adjunct to genomics. Proteomics has been made feasible by major advances in high throughput mass spectrometry, despite the fact that the core component of the technology remains two-dimensional gel electrophoresis. Descriptions of the technique and an example of its application to a microbial system can be found in recent publications (14-16).

Acknowledgments

The author acknowledges the support of the G. Harold & Leila Y. Mathers Charitable Foundation and National Institutes of Health (AI45801).

References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496-512.
2. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, et al. A biochemical genomics approach for identifying genes by the activity of their products. Science 1999; 286: 1153-5.
3. Koonin EV, Mushegian AR, Bork P. Non-orthologous gene displacement. Trends Genet 1996; 12: 334-6.
4. Doolittle WF. Lateral genomics. Trends Cell Biol 1999; 9:M5-8.
5. Doolittle WF. Phylogenetic classification and the universal tree. Science 1999; 284: 2124-8.
6. Technical comments on W.F. Doolittle, Science 1999; 284:2124. Science 1999; 286:1443a.
7. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, et al. Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*. Nature 1997; 390: 580-6.
8. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 1998; 281: 375-88.
9. Wilson WJ, Viswanathan V, Macht M, Wilson KH, Andersen GL. Development of an Affymetrix™ 16S rRNA GeneChip for bacterial identification. Abstracts of the 100th General Meeting of the American Society for Microbiology 2000; Abstract Q-30.
10. Lepp PW, Brown PO, Relman DA. Development of a high density DNA microarray for investigation of microbial diversity in periodontal health and disease. Abstracts of the 100th General Meeting of the American Society for Microbiology 2000; Abstract N-28.
11. Moir DT, Shaw KJ, Hare RS, Vovis GF. Genomics and antimicrobial drug discovery. Antimicrob Agents Chemother 1999; 43: 439-46.
12. Hoffman SL, Rogers WO, Carucci DJ, Venter JC. From genomics to vaccines: Malaria as a model system. Nat Med 1998; 4:1351-3.
13. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 2000; 287:1816-20.
14. Blackstock WP, Weir MP. Proteomics: quantitative and physical mapping of cellular proteins. Trends Biotechnol 1999;17:121-7.
15. Yates JR. Mass spectrometry: from genomics to proteomics. Trends Genet 2000; 16: 5-8.
16. Hirose I, Sano K, Shioda I, Kumano M, Nakamura K, Yamane K. Proteome analysis of *Bacillus subtilis* extracellular proteins: a two-dimensional protein electrophoretic study. Microbiology 2000; 146:65-75.