CHEMICAL REVIEWS

Volume 107, Number 8

Introduction: Advances in Genomics and Proteomics

John Yates received his Ph.D. in Chemistry at the University of Virginia under Professor Donald Hunt. His graduate research involved the development and application of tandem mass spectrometry for sequence analysis of proteins. Following a Biotechnology Fellowship at the California Institute of Technology, he moved to the Department of Molecular Biotechnology at the University of Washington, where he attained the tenured rank of Associate Professor. He is now a Professor in the Department of Cell Biology at The Scripps Research Institute. His research interests include development of integrated methods for tandem mass spectrometry analysis of protein mixtures, bioinformatics using mass spectrometry data, and proteomics. He is the lead inventor of the SEQUEST software for correlating tandem mass spectrometry data to sequences in the database and principle developer of the shotgun proteomics technique for the analysis of protein mixtures. He has received the American Society for Mass Spectrometry research award, the Pehr Edman Award in Protein Chemistry, the American Society for Mass Spectrometry Biemann Medal, the HUPO Distinguished Achievement Award in Proteomics, the Herbert Sober Award from the ASBMB, and the Christian Anfinsen Award from The Protein Society. He is an Associate Editor at *Analytical Chemistry*, a member of the Editorial Boards of the *Journal of Proteome Research*, *Molecular Systems Biology*, *Molecular Oncology*, and *Bioconjugate Chemistry*, and a member of the board of reviewing editors at *Science*.

Andrei Osterman received his Ph.D. in Biochemistry at Moscow University in Russia. The main focus of his research, including postdoctoral studies in the laboratory of Meg Phillips at UT Southwestern in Dallas, was in the field of mechanistic and structural enzymology. In 1999 he joined Integrated Genomics, a start-up biotech firm in Chicago, to lead the effort on genomics-driven discovery of metabolic enzymes and pathways. He joined the faculty of the Burnham Institute for Medical Research in La Jolla, CA, as an Associate Professor in the Bioinformatics and Systems Biology Program in 2003. His research group combines bioinformatics and experimental techniques to explore fundamental and applied aspects of the key metabolic pathways and networks in a variety of species, from bacteria to human.

Genomics and *Proteomics*, the two topics of this issue of *Chemical Reviews*, are deeply connected with each other at various levels. Beyond many obvious links, both are technology-driven *fields in exponential growth where* (quoting from R. Overbeek, one of the contributors in this issue) *every stage is just a warm-up for the next*. Therefore, a humble, albeit exciting, goal of this thematic issue is to provide a snapshot of the current experimental and computational techniques in both fields. Genomic and proteomic technologies, despite their apparent differences, are both truly integrative, with equally important components provided by chemistry, engineering, and computing. However, the most important common denominator of both technologies is their holistic and highly parallel (as opposed to *one-by-one*) approach to the analysis of genes and proteins of entire organisms.

A genomic revolution triggered by the rapid development of new technologies led to a burst of new terminology propagating rapidly across all branches of the life sciences. Indeed, it took more than 60 years for the field of *genomics* (introduced in mid-1980s) to evolve from the *genome* (a chimera of *gen*es and chromos*ome* introduced in early 1920s). In just a few years after the Big Bang of genome sequencing, many other "-omes" (transcriptome, proteome, reactome, interactome, and metabolome, to name just the most prominent ones) and respective "-omics" emerged including *proteomics*, the second topic of this issue. The connotations of the term *genomics* rapidly expanded from focused efforts on mapping and sequencing of complete genomes to a nearly all-inclusive conglomerate of genome-scale experimental and computational enterprises. *Comparative genomics*, one of the most insightful approaches to interpretation of genomic data, was born almost immediately upon landmark reports of completely sequenced bacterial

10.1021/cr068201u CCC: $65.00
Published on Web 07/21/2007

genomes of *Haemophilus ifluenzae* and *Mycoplasma genitalium* in 1995, followed by the first archaeal (*Methanococcus jannaschii*) and eukaryotic (*Saccharomyces cerevisiae*) genomes in 1996. With the development of high-throughput platform technologies, the term *functional genomics* drifted from its original broader scope toward genome-scale analysis of gene and protein expression as well as protein−protein and protein−DNA interactions. *Structural genomics, pharmacogenomic*s, and *chemical genomics* were spawned, followed by endless other variations on the theme. Although a certain degree of skepticism toward some of these newcomers (as driven by winds of fashion) is not without merit, it is more important to recognize in this cross-disciplinary phenomenon a perception of the strong impact of genome-scale approaches, which is beyond any doubt.

Genomics has changed in the most profound way the way we plan and perform our research. Who would have thought, just a decade ago, that the quickest and cheapest way to identify and clone a single gene of interest would be to sequence and analyze the entire genome first? Of no less importance, advents of enabling genomic technologies have strongly affected our perspective of the most fundamental problems of biology, from origin of Life to human health and disease. It would be unthinkable to provide even a brief account of these innumerable applications and implications under the cover of any single edition. Therefore, the focus of this issue is primarily on the key technological aspects of contemporary genome (Section 1) and proteome (Section 2) analysis.

Two articles on advances in genome sequencing technology set the stage for a new *quantum leap* in data acquisition and open the Genomics Section of this issue. In France they say: *l'appétit viens en mangeant.* Indeed, not only did our appetite for getting more and more sequences not get any smaller after hundreds of complete genomes, but it continues to grow. The community wish-list is expanding in many dimensions—hundreds of isolates and strains of individual pathogens (we need to understand variations in virulence, antibiotic resistance, etc.), thousands of novel microbial species (we barely scratched the surface of diversity of prokaryotic Life), millions of unculturable microorganisms in various ecosystems including the human body (*community genomics* or *metagenomics* is one of the most recent and most important additions to the *omics* family). And this is only in the world of microbes! We also need more diverse multicellular eukaryotes and, certainly, thousands and thousands of human genomes, just for starters.

The first article by J. Leamon and J. Rothberg provides us with a thorough overview of the main technical aspects of *454 sequencing*, including elegant chemistry, nanofluidics, and software solutions that transformed the great invention of pyrosequencing into a robust mega-throughput technology. Due to relatively short reads and some other limitations, this technology cannot yet fully replace but rather supplements a conventional Sanger sequencing in the analysis of novel, especially large, genomes. At the same time, 454 sequencing appears to be securing a dominant position in the analysis of genomic variations as well as in metagenomics venues.

The next review by P. Rabinowicz introduces quite a different approach to genomic data acquisition, which is of particular importance for the analysis of humongous plant genomes oversaturated by "junk" DNA, extensive repeats, and intergenic regions that in some cases exceed "useful" (gene-coding) segments by orders of magnitude. Various *gene enrichment* and *filtration* techniques that take advantage of the differences in certain chemical properties between useful and parasitic DNA are the key technological components contributing to the utility and mere feasibility of many ongoing plant sequencing projects.

Notwithstanding the importance and the challenges of expanding genomic sequencing efforts, most people recognize even bigger challenges associated today with computational (*bioinformatic*) analysis of genomic data. Therefore, the primary focus of the following articles in this section is on various aspects of *computational genomics* contributing to a broad range of tasks, from genome assembly to functional interpretation.

The review by H. Tang provides a detailed survey of computational techniques and software tools to address the most daunting problems in assembly and primary analysis of complex genomes such as those of multicellular eukaryotes. In addition to its obvious practical importance, the analysis of repeats and genome rearrangements sheds light on such fundamental problems as genome instability and evolution.

The next level of complexity associated with the primary analysis of mammalian and other eukaryotic genomes, identification of encoded proteins, is reflected upon in the review by I. Artamonova and M. Gelfand. The authors extensively use an evolutionary approach to analyze various fundamental and practical aspects of exon−intron arrangement and shuffling across a growing number of available eukaryotic genomes. This article also illustrates that genome evolution and evolutionary concepts are not only the *means* of comparative genomics, but they are often among its key *goals* and deliverables, providing new insights to the mystery of origination and diversification of species.

The subject of accurate and consistent genome annotation is of paramount importance for the entire field of genomics and for its innumerable applications in biology, biotechnology, and medicine. In recognition of this obvious notion, as well as of the widespread discontent with the quality and utility of annotations available in various public archives, we included in this issue two articles that together reflect upon a broad spectrum of problems and solutions in this important area.

The article by R. Overbeek, D. Bartels, V. Vonstein, and F. Meyer starts from the detailed analysis of the computational approaches, algorithms, and software tools for accurate *gene calling*, identification of genome regions encoding proteins and functional RNAs (such as tRNA and rRNA) in prokaryotes. Although the absence of introns makes this task much more straightforward compared to protein prediction in eukaryotic genomes (as discussed in the previous review), many problems with identification of gene starts as well as with short genes and pseudogenes lead to a substantial level of errors. We should notice that the rapidly growing amount and quality of proteomics data (as reflected in Section 2 of this thematic issue) will certainly contribute to a significant improvement of gene calling techniques. The second part of this article spells out the key principles of functional annotation of prokaryotic genes. It provides a historic perspective and an outlook of the major developments aimed to significantly improve the quality of automated gene functional assignment in hundreds (soon-to-be thousands) of bacterial and archaeal genomes. Among those are establishment of a controlled vocabulary to describe aspects of gene functions projectable across species, development and

curation of functionally homogeneous and consistent protein families, reconstruction of pathways and subsystems across multiple diverse species, integration of various types of gene functional coupling evidence to support functional assignments, and help in predicting functions for previously uncharacterized protein families.

The next review by D. Frishman provides a very detailed survey of various approaches and software tools for manual and automated genome annotation. A nearly comprehensive and candid analysis of various types and sources of errors rapidly propagating across genomes and databases (particularly valuable when provided by people who contribute such a substantial effort to genome annotation) sets the stage for the insightful discussion of computational and experimental ways to reduce such errors. The feasibility and importance of this task is discussed with respect to both the exponentially growing number of sequenced genomes and the growing demand of high-quality annotations by emerging genome-scale functional studies (Systems Biology) of multiple diverse species.

The last two reviews in Section 1 illustrate two different aspects of the next (postannotation) level of functional genome analysis. The article by D. Rodionov focuses on utilization of comparative genomics for identification, reconstruction, and prediction of bacterial regulons and regulatory networks. It provides a critical overview of existing computational techniques and algorithms for accurate delineation of conserved regulatory sites in the upstream regions of bacterial genes. Of no less importance, it contains the most comprehensive catalog of examples of computationally identified regulons, covering a substantial fraction of metabolic pathways in a broad spectrum of bacteria. These examples convincingly illustrate the power and utility of the comparative genomic approach for exploration of regulatory networks. This approach strongly complements and often exceeds the efficiency of experimental techniques (such as expression microarrays) traditionally dominating the field.

Synthetic genomics (or rather genome engineering), from a fundamental alteration of gene content in existing microbes up to creation of rationally designed and fully synthetic life forms, is yet another addition to the postgenomic gallery. *Frankencell* is not a new concept; what is definitely new is its migration from the world of fantasy to the world of real and already actively pursued technological challenges. The fundamental and practical importance of this long-range goal as well as the first steps in this direction by ongoing computational and experimental studies on genome minimization are discussed in the article by T. Fehér, B. Papp, C. Pál, and G. Pósfai, closing Section 1 of this issue. The concept of minimal gene-set and a related but distinct notion of the Core of Life were actively explored and debated from the first days of genomics. This review illustrates a fundamental revision of the early perceptions in this field instigated by accumulation of genomic data and progress of comparative and functional genomic studies, including large-scale projects on systematic gene deletion and genome reduction in model microbes.

Genomics has provided enabling information, data, and resources for a broad spectrum of other scientific areas. In particular, the sequencing of genomes has had a profound impact on the study of protein function. By defining the basic blueprint for all proteins in an organism, a resource is created that has been used to drive technologies for large-scale analysis of proteins. The drive to sequence genomes coincided with the invention of new mass spectrometry based technologies for the analysis of biological molecules. These techniques have evolved to use genome sequences for large-scale and rapid identification and quantification of proteins. Mass spectrometry based approaches for the analysis of proteins have flourished with the sequencing of genomes, and they have become a powerful analytical approach to derive a wide variety of information about proteins and biological systems.

Mass spectrometry methods have been developed to assess the structures of proteins that serve as an adjunct and supplementary tool to derive information that is hard to obtain using NMR and X-ray methods, or serve to fill in missing pieces of information. G. Xu and M. Chance review recent research to "footprint" the solvent-exposed regions of proteins to derive structural information based on solvent accessibility. In particular, they describe the recent efforts to derive higher resolution information using mass spectrometry-based methods. J. Benesch, B. Ruotolo, D. Simmons, and C. Robinson examine a remarkable area of research that studies intact protein complexes in the gas-phase of a mass spectrometer. By introducing complexes into the mass spectrometer, structural aspects of the complex can be studied including the stability of subcomplexes, the presence of ligands, and the stoichiometry of constituents. Steady advances are being made in the development of mass spectrometry based techniques for the study of protein and protein complex structures, and represent an exciting and important area of study.

A productive and prolific area in proteomics is the identification of proteins. Generally, these are done in the context of a biological question or discovery. The technology and approaches for large-scale proteomic experiments are dynamic and are constantly improving the type of information that can be derived. A key element to proteomic experiments is protein identification using mass spectrometry and computer algorithms. G. Lubec and L. Afjehi-Sadat review the limitations and pitfalls of the application of mass spectrometry for protein identification. G. Musso, Z. Zhang, and A. Emili further discuss experimental and computational procedures for the assessment of large-scale protein complex analysis. The identification of the components of protein complexes has been a strength of mass spectrometry techniques, and this technology has allowed large-scale studies to derive genomic scale data on the complexes of a cell or organism. An area seeking to translate proteomic technology to a clinical setting is the discovery of biomarkers of disease. Serum and plasma are key areas for analysis to identify biomarkers with diagnostic potential. This area has been very active over the past few years in both improvements to approaches and mass spectrometry based techniques. H. Issaq, Z. Xiao, and T. Veenstra provide a critical assessment of progress in this area.

The ability to derive data and thus biological information in proteomic studies is enabled by technology and methodology. T. Liu, M. Belov, N. Jaitly, W. Qian, and R. Smith describe recent advances in the use of accurate mass measurements as a means to identify proteins. Exciting new developments have emerged in this area with the development of new types of mass spectrometers capable of routine measurements of high resolution and high mass accuracy. M. Fournier, J. Gilmore, S. Martin-Brown, and M. Washburn

review the use of tandem mass spectrometry approaches to identify proteins through the analysis of digested protein mixtures. These methods often employ high resolution separation techniques such as a multidimensional liquid chromatography separation. Analysis of membrane proteins has traditionally been a challenge because of the hydrophobicity of proteins. A. Speers and C. Wu describe recent advances in the analysis of membrane proteins using "shotgun" proteomic methods that are leading to large-scale genomic type studies. These reviews are a snapshot of the dynamic research and progress occurring in the proteomics field and represent the excitement of discoveries to come.

John R. Yates III
Scripps Research Institute

Andrei L. Osterman
Burnham Institute for Medical Research

CR068201U