Book Review

Modern Genome Annotation

The advent and widespread availability of massively parallel DNA sequencing technology and the development of its applications [gene annotation, gene expression, protein–protein interaction (PPI) and pathway analysis] are leading to genomic era in which full genome sequencing of a variety of organisms will become universal in the global scientific community. The availability of genomic sequences has provided many excellent opportunities for interpretation and understanding of the expression of RNA and proteins. Many valuable computational tools have been developed to assist in the annotation of a variety of genome sequences. Further, older computational tools have been improved to adjust to the new environment.

‘Modern Genome Annotation’ is a comprehensive collection of numerous experimental and computational methods related to defining the biological role of molecules in all living organisms. This book consists of eight sections, made up of 19 chapters. The first section, consisting of two chapters (Chapters 1.1 and 1.2), focuses on computational eukaryotic gene prediction and quality control for correctly identifying genes in newly sequenced genomes. Chapter 1.1 (by Tyler Alioto and Roderic Guigo) explains the basic principles of gene prediction models, such as hidden Markov models and support vector machines, and introduces various tools that have implemented these theoretical frameworks. Chapter 1.2 (by Alinda Nagy et al.) focuses on the assessment and validation of the predicted genes and introduces the MisPred pipeline to identify mispredicted and abnormal genes/proteins from current knowledge database.

Section 2, which includes three chapters (Chapters 2.1, 2.2 and 2.3), deals with gene regulation and gene expression. Chapter 2.1 (by Olivier Sand et al.) evaluates predictions of cis-acting regulatory elements in genome sequences based on various biological and statistical assumptions. Chapter 2.2 (by Thomas Manke et al.) addresses recent efforts towards identifying protein–DNA interactions on a large scale based on a biophysical framework. This section concludes with Chapter 2.3 (by Richard Coulson et al.), which discusses computational techniques for identification of transcription factors and their binding sites as an approach to understand the gene regulation.

Section 3, Chapter 3 (by Richard Mott), briefly reviews methods for combining functional genomic data to detect complex diseases. Particular focus is given to the detection of the genetic association of disease through the use of single nucleotide polymorphisms, linkage disequilibrium, quantitative trait loci and recombinant inbred lines. This section emphasizes the need for an effective system to handle this huge amount of biological information. Any readers interested in catching up with the method to combine and apply genome data of the type described above will find these chapters very useful for the comprehensive content to find the genes in particular with a phenotype.

Section 4, composed of six chapters, covers the functional annotation of proteins. Chapter 4.1 (by Alan J. Bridge et al.) introduces numerous protein sequence databases that contain precise descriptions of the biological functions of individual proteins. Chapters 4.2 (by Claudine Médigue and Antonine Danchin) and 4.3 (by Irena Artamonova et al.) explain proper procedures and data mining for annotating bacterial genomes and the user-friendly application tools available for this purpose. Chapters from 4.4 (by Corin Yeats et al.) to 4.6 (by Bas Vroling and Gert Vriend) review modern genome annotations, providing a particular focus on the analysis of protein structures as a method to improve confidence in the assignment of proteins to particular families.

Section 5 deals with protein structure prediction. Chapter 5.1 (by Anna Tramontano et al.) begins with a discussion of folding recognition of protein structures and provides descriptions of various
methods for protein structure prediction. Chapter 5.2 (by Rita Casadio et al.) focuses on prediction of membrane protein structures, a class that comprises an average of 30% of all the protein sequences.

Section 6, comprising Chapters 6.1 and 6.2, discusses PPIs and pathways, focusing on the analysis of computational resources for the study of metabolism and the reconstruction of metabolic networks. Chapter 6.1 (by Pierre-Yves Bourguignon et al.) describes the currently available comprehensive metabolic databases and explains how to reconstruct metabolic networks that include mathematical properties and theoretical models. Chapter 6.2 (by Dmitrij Frishman et al.) focuses on computational methods for analyzing and predicting PPIs and introduces a diverse set of databases and tools for this purpose.

Section 7, comprising Chapter 7 (by Gabrielle A. Reeves et al.), briefly describes diverse resources for distributed protein annotation. In particular, this chapter introduces the distributed annotation system (DAS) as a practical solution for integration of all information derived from different proteome annotations. After reading the chapter of this book, the readers may understand how to use them and where to look for new applications and improvement without local installations.

Section 8 presents some applications of bioinformatics in two chapters. Chapter 8.1 (by Ben Adams et al.) discusses recent advances in viral bioinformatics and describes computational approaches for analyzing host–pathogen interactions. The final chapter, Chapter 8.2 (by Michael L. Tress et al.), addresses the analysis of splice variants identified in the genes experimentally studied by the pilot phase of the ENCODE project.

In summary, this book strives to provide a broad diversity of readers with a basic understanding of several technical and scientific issues related to genome annotation technologies in an interdisciplinary field. Although this book is aimed at individuals with computing experience, it contains sufficient background from both the biological and computing fields that students from either discipline should feel comfortable with the text. This book is too general or superficial for scientists fully involved in research and development in the areas of genome biology. This text could be very helpful to those interested in learning more about the application of annotation methods and how and why these technologies affect genome annotation.

Dae-Won Kim and Hong-Seog Park
Genome Research Center,
Korea Research Institute of Bioscience and Biotechnology,
Daejeon 305-806, Korea
E-mail: todaevon@kribb.re.kr, hspark@kribb.re.kr