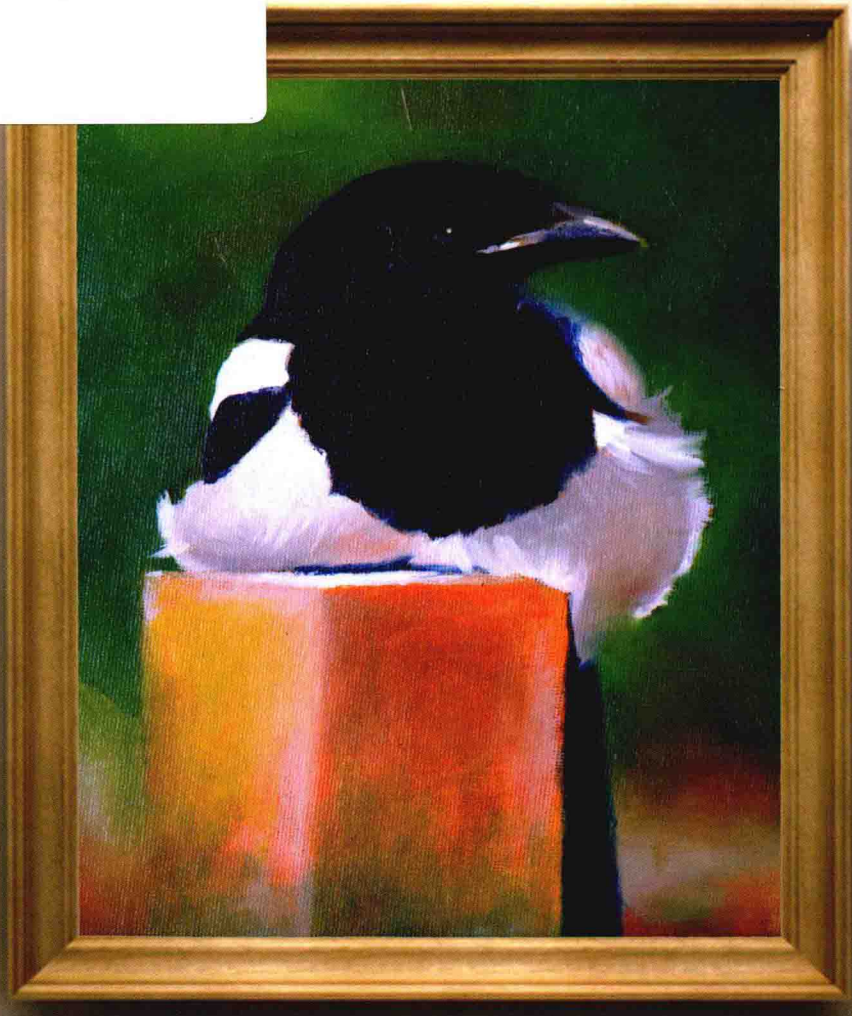
Chapman & Hall/CRC
Mathematical and Computational Biology Series

Genome Annotation

Jung Soh, Paul M.K. Gordon,
and Christoph W. Sensen



CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC Mathematical and Computational Biology Series

Genome Annotation

Jung Soh, Paul M.K. Gordon,
and Christoph W. Sensen

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
Version Date: 20120801

International Standard Book Number: 978-1-4398-4117-4 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Genome Annotation

CHAPMAN & HALL/CRC
Mathematical and Computational Biology Series

Aims and scope:
This series aims to capture new developments and summarize what is known over the entire spectrum of mathematical and computational biology and medicine. It seeks to encourage the integration of mathematical, statistical, and computational methods into biology by publishing a broad range of textbooks, reference works, and handbooks. The titles included in the series are meant to appeal to students, researchers, and professionals in the mathematical, statistical and computational sciences, fundamental biology and bioengineering, as well as interdisciplinary researchers involved in the field. The inclusion of concrete examples and applications, and programming techniques and examples, is highly encouraged.

Series Editors

Proposals for the series should be submitted to one of the series editors above or directly to:
CRC Press, Taylor & Francis Group
4th, Floor, Albert House
1-4 Singer Street
London EC2A 4BQ
UK

Published Titles

Algorithms in Bioinformatics: A Practical Introduction
Wing-Kin Sung

Bioinformatics: A Practical Approach
Shui Qing Ye

Biological Computation
Ehud Lamm and Ron Unger

Biological Sequence Analysis Using the SeqAn C++ Library
Andreas Gogol-Döring and Knut Reinert

Cancer Modelling and Simulation
Luigi Preziosi

Cancer Systems Biology
Edwin Wang

Cell Mechanics: From Single Scale-Based Models to Multiscale Modeling
Arnaud Chauvière, Luigi Preziosi, and Claude Verdier

Clustering in Bioinformatics and Drug Discovery
John D. MacCuish and Norah E. MacCuish

Combinatorial Pattern Matching Algorithms in Computational Biology Using Perl and R
Gabriel Valiente

Computational Biology: A Statistical Mechanics Perspective
Ralf Blossey

Computational Hydrodynamics of Capsules and Biological Cells
C. Pozrikidis

Computational Neuroscience: A Comprehensive Approach
Jianfeng Feng

Computational Systems Biology of Cancer
Emmanuel Barillot, Laurence Calzone, Philippe Hupe, Jean-Philippe Vert, and Andrei Zinovyev

Data Analysis Tools for DNA Microarrays
Sorin Draghici

Differential Equations and Mathematical Biology, Second Edition
D.S. Jones, M.J. Plank, and B.D. Sleeman

Dynamics of Biological Systems
Michael Small

Engineering Genetic Circuits
Chris J. Myers

Exactly Solvable Models of Biological Invasion
Sergei V. Petrovskii and Bai-Lian Li

Gene Expression Studies Using Affymetrix Microarrays
Hinrich Göhlmann and Willem Talloen

Genome Annotation
Jung Soh, Paul M.K. Gordon, and Christoph W. Sensen

Glycome Informatics: Methods and Applications
Kiyoko F. Aoki-Kinoshita

Handbook of Hidden Markov Models in Bioinformatics
Martin Gollery

Introduction to Bioinformatics
Anna Tramontano

Introduction to Bio-Ontologies
Peter N. Robinson and Sebastian Bauer

Introduction to Computational Proteomics
Golan Yona

Introduction to Proteins: Structure, Function, and Motion
Amit Kessel and Nir Ben-Tal

An Introduction to Systems Biology: Design Principles of Biological Circuits
Uri Alon

Kinetic Modelling in Systems Biology
Oleg Demin and Igor Goryanin

Knowledge Discovery in Proteomics
Igor Jurisica and Dennis Wigle

Meta-analysis and Combining Information in Genetics and Genomics
Rudy Guerra and Darlene R. Goldstein

Methods in Medical Informatics: Fundamentals of Healthcare Programming in Perl, Python, and Ruby
Jules J. Berman

Published Titles (continued)

Modeling and Simulation of Capsules and Biological Cells
C. Pozrikidis

Niche Modeling: Predictions from Statistical Distributions
David Stockwell

Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems
Qiang Cui and Ivet Bahar

Optimal Control Applied to Biological Models
Suzanne Lenhart and John T. Workman

Pattern Discovery in Bioinformatics: Theory & Algorithms
Laxmi Parida

Python for Bioinformatics
Sebastian Bassi

Quantitative Biology: From Molecular to Cellular Systems
Sebastian Bassi

Spatial Ecology
Stephen Cantrell, Chris Cosner, and Shigui Ruan

Spatiotemporal Patterns in Ecology and Epidemiology: Theory, Models, and Simulation
Horst Malchow, Sergei V. Petrovskii, and Ezio Venturino

Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition
Sorin Drăghici

Stochastic Modelling for Systems Biology, Second Edition
Darren J. Wilkinson

Structural Bioinformatics: An Algorithmic Approach
Forbes J. Burkowski

The Ten Most Wanted Solutions in Protein Bioinformatics
Anna Tramontano

Preface

The year 1995 saw the arrival of the first completed microbial genomes, *Haemophilus influenzae* and *Mycoplasma genitalium*. Several years of struggle for a complete genome was ended by the group at The Institute for Genomic Research (TIGR). From today's point of view, 16 years and more than a thousand completed genomes later (including, of course, the human genome), it may be hard to understand how much of an accomplishment this was. At the time, however, most laboratories around the world were still sequencing on slab gels using radioisotopes as the label for the fragments, which represent the DNA sequence.

When we sit down to "browse" genomes today, we do not often remember the days before e-mail and the Internet, or the days before automated DNA sequencing became a commodity. But it is certainly worthwhile to take a look back, as many of the design decisions that were made in the last 16 years influence the way we deal with genomic information today.

When the first genomes were presented at a by-invitation-only meeting in Worcester, England, in 1995, the only tool that was capable of handling such a large file was a word processor. Therefore, the sequence was first presented to the scientists at the meeting as a character file, which was scrolling on the screen behind the speaker. One of the major problems with handling a large DNA sequence file at the time was that most bioinformatics software was only tailored for DNA fragments of a size much less than a complete microbial genome, typically no more than approximately 100 kilobase pairs. The first automated genome analysis and annotation systems were barely emerging in 1995, and thus the handling of a complete genome with a size of more than a million base pairs all at once was impossible.

The Web was a fledgling entity in 1995, with not much power and entirely based on the Hypertext Markup Language (HTML). It became clear very quickly that only large communities of scientists with a diverse

background could really make sense of the genomic information, provided that they were enabled to collaborate, and thus the Web quickly became the vehicle by which genome annotations were created and exchanged among scientists. The first automated genome analysis and annotation systems, which were Web-based, initially produced tabular output that listed the location of potential genes and gene functions, which were predicted mostly by database comparison. It became obvious very early that this was not sufficient for biologists, therefore graphical subsystems were added, which are today part and parcel of all genome analysis and annotation systems and are probably the only part of an automated genome analysis and annotation system that most users ever encounter.

Over time, the Web developed into the massive entity it is today, with many additions to the Web technologies, which were utilized in turn by the developers of today's genome analysis and annotation tools. The three most useful tools in this context were probably (1) the creation of the programming language Java by James Gosling, which allowed the development of truly platform-independent applications; (2) the introduction of Extensible Markup Language (XML), which could adequately be used for the description of biological and medical objects; and (3) the creation of Web services (for example, the BioMOBY system), which made distributed computing simple and easy, and allowed the transparent and seamless integration of new bioinformatics tools into Web-based bioinformatics applications.

DNA sequencing technology has progressed in several iterations to today's level, which is called "next-generation sequencing," but really represents the third or fourth generation of DNA sequencing technologies, with yet another generation just around the corner. The sheer amount of DNA sequence, which can be produced on a single device today, is mind-boggling. It has literally become possible to resequence genomes the size of the human genome within a few hours in a single laboratory and the $1000 human genome is on the horizon. At the time of this writing (2012), very few genome annotation pipelines are capable of dealing with this information volume and new strategies need to be developed to accommodate the needs of today's genome researchers.

In the near future, everyone will be able to carry their genome sequence on some kind of data storage device and diagnostics might become largely based on the results of genomics screens, which will be cheaper than today's advanced imaging technologies (MRI and CT scans, for example). Thoroughly annotated genomic information and the integration

of all information into a single model will be a prerequisite to successful approaches to individualized medicine, the development of advanced crops and the sustainable production of food, medical research and development, and the development of new and sustainable energy sources.

This book attempts to introduce the topic of automated genome analysis and annotation. The initial chapters take the reader through the last 16 years, explaining how the current analysis strategies were developed. This is followed by the introduction of up-to-date tools, which represent today's state of the art. The authors also discuss strategies for the analysis and annotation of next-generation DNA sequencing data. This book is intended to be used by professionals and students interested in entering the field.

We would like to thank Hershel Safer and the editorial team at CRC Press/Taylor & Francis for their patience while creating this book.

Authors

Jung Soh is a research associate at the University of Calgary, in Alberta, Canada. He was born in Seoul, Korea, and received a Ph.D. in computer science from the University at Buffalo, State University of New York, where he worked at the Center of Excellence for Document Analysis and Recognition (CEDAR). From 1992 to 2004, he worked as a principal research scientist at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea, in areas such as biometrics, human–computer interaction, robotics, video analysis, pattern recognition, and document analysis. Dr. Soh moved to Calgary in 2005. His current research interests include bioinformatics, machine learning, and biomedical data visualization.

Paul Gordon is the bioinformatics support specialist for the Alberta Children's Hospital Research Institute at the University of Calgary. He was born in Halifax, Nova Scotia, Canada. He received his bachelor of science (first class honors) and master's of computer science from Dalhousie University (Halifax). He worked at the Canadian National Research Council's Institute for Information Technology and at its Institute for Marine Biosciences

(1996–2001) before moving to Calgary. From 2002–2011, he was a lead programmer on several genome Canada-funded projects, supported by the University of Calgary. In the new era of the $1000 genome, Gordon's current work focuses on developing bioinformatics techniques for personalized medicine.

Christoph W. Sensen is a professor of bioinformatics at the University of Calgary, in Alberta, Canada. He was born in Oberhausen-Sterkrade, Germany, and studied biology in Mainz, Düsseldorf, and Köln. From October 1992 to November 1993, he worked as a visiting scientist at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. At the beginning of 1994, he moved to Canada. Until his move to Calgary in February 2001, he worked as a research officer at the National Research Council of Canada's Institute for Marine Biosciences (NRC-IMB). His research interests include genome research and bioinformatics.

Contributor

Stephen Strain was born in Szeged, Hungary. He studied visual arts and music in Debrecen, Szeged, and Budapest. In 1981, he emigrated to Calgary, Alberta, Canada, where he lives and works as an artist. He is known for his interpretation of flamenco music (www.flamencoelegante.com) and as a painter.

His painting *Magpie* is used as the cover design.

Contents