Wiley Series on Bioinformatics:
Computational Techniques and Engineering
Yi Pan and Albert Y. Zomaya, Series Editors

Biological Knowledge Discovery Handbook

PREPROCESSING, MINING, AND POSTPROCESSING OF BIOLOGICAL DATA

MOURAD ELLOUMI • ALBERT Y. ZOMAYA

IEEE
IEEE computer society
WILEY

BIOLOGICAL KNOWLEDGE DISCOVERY HANDBOOK

Wiley Series on

Bioinformatics: Computational Techniques and Engineering

A complete list of the titles in this series appears at the end of this volume.

BIOLOGICAL KNOWLEDGE DISCOVERY HANDBOOK

Preprocessing, Mining, and Postprocessing of Biological Data

Edited by

MOURAD ELLOUMI
Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE) and University of Tunis-El Manar, Tunisia

ALBERT Y. ZOMAYA
The University of Sydney

WILEY | IEEE computer society

Cover Design: Michael Rutkowski
Cover Image: ©iStockphoto/cosmin 4000

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Elloumi, Mourad.
Biological knowledge discovery handbook : preprocessing, mining, and postprocessing of biological data / Mourad Elloumi, Albert Y. Zomaya.
pages cm. – (Wiley series in bioinformatics; 23)
ISBN 978-1-118-13273-9 (hardback)
1. Bioinformatics. 2. Computational biology. 3. Data mining. I. Zomaya, Albert Y. II. Title.
QH324.2.E45 2012
572.80285–dc23
2012042379

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To my family for their patience and support.

Mourad Elloumi

To my mother for her many sacrifices over the years.

Albert Y. Zomaya

CONTENTS

SECTION I BIOLOGICAL DATA PREPROCESSING

PART A: BIOLOGICAL DATA MANAGEMENT

PART B: BIOLOGICAL DATA MODELING

PART C: BIOLOGICAL FEATURE EXTRACTION

PART D: BIOLOGICAL FEATURE SELECTION

SECTION II BIOLOGICAL DATA MINING

PART E: REGRESSION ANALYSIS OF BIOLOGICAL DATA

PART F: BIOLOGICAL DATA CLUSTERING

PART G: BIOLOGICAL DATA CLASSIFICATION

PART H: ASSOCIATION RULES LEARNING FROM BIOLOGICAL DATA

PART I: TEXT MINING AND APPLICATION TO BIOLOGICAL DATA

PART J: HIGH-PERFORMANCE COMPUTING FOR BIOLOGICAL DATA MINING

SECTION III BIOLOGICAL DATA POSTPROCESSING

PART K: BIOLOGICAL KNOWLEDGE INTEGRATION AND VISUALIZATION

PREFACE

With the massive developments in molecular biology during the last few decades, we are witnessing an exponential growth of both the volume and the complexity of biological data. For example, the Human Genome Project provided the sequence of the 3 billion DNA bases that constitute the human genome. Consequently, we are provided too with the sequences of about 100,000 proteins. Therefore, we are entering the postgenomic era: After having focused so many efforts on the accumulation of data, we now must to focus as much effort, and even more, on the analysis of the data. Analyzing this huge volume of data is a challenging task not only because of its complexity and its multiple and numerous correlated factors but also because of the continuous evolution of our understanding of the biological mechanisms. Classical approaches of biological data analysis are no longer efficient and produce only a very limited amount of information, compared to the numerous and complex biological mechanisms under study. From here comes the necessity to use computer tools and develop new in silico high-performance approaches to support us in the analysis of biological data and, hence, to help us in our understanding of the correlations that exist between, on one hand, structures and functional patterns of biological sequences and, on the other hand, genetic and biochemical mechanisms. *Knowledge discovery and data mining* (KDD) are a response to these new trends.

Knowledge discovery is a field where we combine techniques from algorithmics, soft computing, machine learning, knowledge management, artificial intelligence, mathematics, statistics, and databases to deal with the theoretical and practical issues of extracting *knowledge*, that is, new concepts or concept relationships, hidden in volumes of raw data. The knowledge discovery process is made up of three main phases: *data preprocessing*, *data processing*, also called *data mining*, and *data postprocessing*. Knowledge discovery offers the capacity to automate complex search and data analysis tasks. We distinguish two types of knowledge discovery systems: *verification systems* and *discovery* ones. Verification systems are limited to verifying the user's hypothesis, while discovery ones autonomously predict and explain new knowledge. Biological knowledge discovery process should take into account both the characteristics of the biological data and the general requirements of the knowledge discovery process.

Data mining is the main phase in the knowledge discovery process. It consists of extracting nuggets of information, that is, pertinent patterns, pattern correlations, and estimations or rules, hidden in huge bodies of data. The extracted information will be used in the verification of the hypothesis or the prediction and explanation of knowledge. Biological data mining aims at extracting motifs, functional sites, or clustering/classification rules from biological sequences.

Biological KDD are complementary to laboratory experimentation and help to speed up and deepen research in modern molecular biology. They promise to bring us new insights into the growing volumes of biological data.

This book is a survey of the most recent developments on techniques and approaches in the field of biological KDD. It presents the results of the latest investigations in this field. The techniques and approaches presented deal with the most important and/or the newest topics encountered in this field. Some of these techniques and approaches represent improvements of old ones while others are completely new. Most of the other books on biological KDD either lack technical depth or focus on specific topics. This book is the first overview on techniques and approaches in biological KDD with both a broad coverage of this field and enough depth to be of practical use to professionals. The biological KDD techniques and approaches presented here combine sound theory with truly practical applications in molecular biology. This book will be extremely valuable and fruitful for people interested in the growing field of biological KDD, to discover both the fundamentals behind biological KDD techniques and approaches, and the applications of these techniques and approaches in this field. It can also serve as a reference for courses on bioinformatics and biological KDD. So, this book is designed not only for practitioners and professional researchers in computer science, life science, and mathematics but also for graduate students and young researchers looking for promising directions in their work. It will certainly point them to new techniques and approaches that may be the key to new and important discoveries in molecular biology.

This book is organized into 11 parts: Biological Data Management, Biological Data Modeling, Biological Feature Extraction, Biological Feature Selection, Regression Analysis of Biological Data, Biological Data Clustering, Biological Data Classification, Association Rules Learning from Biological Data, Text Mining and Application to Biological Data, High-Performance Computing for Biological Data Mining, and Biological Knowledge Integration and Visualization. The 48 chapters that make up the 11 parts were carefully selected to provide a wide scope with minimal overlap between the chapters so as to reduce duplication. Each contributor was asked that his or her chapter should cover review material as well as current developments. In addition, the authors chosen are leaders in their respective fields.

MOURAD ELLOUMI AND ALBERT Y. ZOMAYA

CONTRIBUTORS

Jad Abbass, Faculty of Science, Engineering and Computing, Kingston University, London, United Kingdom and Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon

Muhammad Abulaish, Center of Excellence in Information Assurance, King Saud University, Riyadh, Saudi Arabia and Department of Computer Science, Jamia Millia Islamia (A Central University), New Delhi, India

Syed Toufeeq Ahmed, Vanderbilt University Medical Center, Nashville, Tennessee

Shiva Akbari-Birgani, Laboratory of Systems Biology and Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

Ali Al Mazari, School of Information Technologies, The University of Sydney, Sydney, Australia

Mohamed Al Sayed Issa, Computers and Systems Department, Faculty of Engineering, Zagazig University, Egypt

Yazdan Asgari, Laboratory of Systems Biology and Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

Wassim Ayadi, Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE) and LERIA, University of Angers, Angers, France

Haider Banka, Department of Computer Science and Engineering, Indian School of Mines, Dhanbad, India

Laure Berti-Équille, Institut de Recherche pour le Développement, Montpellier, France

Gianluca Bontempi, Machine Learning Group, Computer Science Department, Université Libre de Bruxelles, Brussels, Belgium

Nigel P. Brown, BioQuant, University of Heidelberg, Heidelberg, Germany

Giulia Bruno, Dipartimento di Ingegneria Gestionale e della Produzione, Politecnico di Torino, Torino, Italy

David Campos, DETI/IEETA, University of Aveiro, Aveiro, Portugal

Jessica Andrea Carballido, Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC), Dept. Computer Science and Engineering, Universidad Nacional del Sur, Bahía Blanca, Argentina

Luciano Cascione, Department of Clinical and Molecular Biomedicine, University of Catania, Italy

Ümit V. Çatalyürek, Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio

Carlo Cattani, Department of Mathematics, University of Salerno, Fisciano (SA), Italy

Meghana Chitale, Department of Computer Science, Purdue University, West Lafayette, Indiana

Young-Rae Cho, Department of Computer Science, Baylor University, Waco, Texas

Kwok Pui Choi, Department of Statistics and Applied Probability, National University of Singapore, Singapore

Matteo Comin, Department of Information Engineering, University of Padova, Padova, Italy

Francesca Cordero, Department of Computer Science, University of Torino, Turin, Italy

Suresh Dara, Department of Computer Science and Engineering, Indian School of Mines, Dhanbad, India

Bhaskar DasGupta, Department of Computer Science, University of Illinois at Chicago, Chicago, Illinois

Hasan Davulcu, Department of Computer Science and Engineering, Ira A. Fulton Engineering, Arizona State University, Tempe, Arizona

Mourad Elloumi, Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE) and University of Tunis-El Manar, Tunisia

Juan Esquivel-Rodríguez, Department of Computer Science, Purdue University, West Lafayette, Indiana

Alfredo Ferro, Department of Clinical and Molecular Biomedicine, University of Catania, Italy

Alessandro Fiori, Dipartimento di Automatica e Informatica, Politecnico di Torino, Torino, Italy

Adelaide Valente Freitas, DMat/CIDMA, University of Aveiro, Portugal

Terry Gaasterland, Scripps Genome Center, University of California San Diego, San Diego, California

Cristian Andrés Gallo, Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC), Dept. Computer Science and Engineering, Universidad Nacional del Sur, Bahía Blanca, Argentina

Roger J. Garsia, Department of Clinical Immunology, Royal Prince Alfred Hospital, Sydney, Australia

Raffaele Giancarlo, Department of Mathematics and Informatics, University of Palermo, Palermo, Italy

Rosalba Giugno, Department of Clinical and Molecular Biomedicine, University of Catania, Italy

Jin-Kao Hao, LERIA, University of Angers, Angers, France

Ayat Hatem, Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio

Heiko Horn, Department of Disease Systems Biology, The Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Ting Hu, Computational Genetics Laboratory, Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire

Kun Huang, Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio

Zina M. Ibrahim, Social Genetic and Developmental Psychiatry Centre, King's College London, London, United Kingdom

Dino Ienco, Institut de Recherche en Sciences et Technologies pour l'Environnement, Montpellier, France

Costas S. Iliopoulos, Department of Informatics, King's College London, Strand, London, United Kingdom and Digital Ecosystems & Business Intelligence Institute, Curtin University, Centre for Stringology & Applications, Perth, Australia

Jahiruddin, Department of Computer Science, Jamia Millia Islamia (A Central University), New Delhi, India

Laetitia Jourdan, INRIA Lille Nord Europe, Villeneuve d'Ascq, France

Lakshmi Kaligounder, Department of Computer Science, University of Illinois at Chicago, Chicago, Illinois

Radha Krishna Murthy Karuturi, Computational and Mathematical Biology, Genome Institute of Singapore, Singapore

Khairul A. Kasmiran, School of Information Technologies, The University of Sydney, Sydney, Australia

Ioannis Kavakiotis, Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Kamer Kaya, Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio

Catharina Maria Keet, School of Computer Science, University of KwaZulu-Natal, Durban, South Africa

Daisuke Kihara, Department of Computer Science, Purdue University, West Lafayette, Indiana and Department of Biological Sciences, Purdue University, West Lafayette, Indiana

Gaurav Kumar, Department of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, Australia

Chee Keong Kwoh, School of Computer Engineering, Nanyang Technological University, Singapore

Giuseppe Lancia, Department of Mathematics and Informatics, University of Udine, Udine, Italy

Hee-Jin Lee, Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

Juntao Li, Computational and Mathematical Biology, Genome Institute of Singapore, Singapore and Department of Statistics and Applied Probability, National University of Singapore, Singapore

Wentian Li, Robert S. Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore LIJ Health Systems, Manhasset, New York

Yehua Li, Department of Statistics and Statistical Laboratory, Iowa State University, Ames, Iowa

Charles Lindsey, StataCorp, College Station, Texas

Giosué Lo Bosco, Department of Mathematics and Informatics, University of Palermo, Palermo, Italy and I.E.ME.S.T., Istituto Euro Mediterraneo di Scienza e Tecnologia, Palermo, Italy

Nashat Mansour, Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon

Ali Masoudi-Nejad, Laboratory of Systems Biology and Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

Sérgio Matos, DETI/IEETA, University of Aveiro, Aveiro, Portugal

Patrick E. Meyer, Machine Learning Group, Computer Science Department, Université Libre de Bruxelles, Brussels, Belgium

Debahuti Mishra, Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar, Odisha, India

Sashikala Mishra, Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar, Odisha, India

Ahmed Mokaddem, Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE) and University of Tunis-El Manar, El Manar, Tunisia

Kartick Chandra Mondal, Laboratory I3S, University of Nice Sophia-Antipolis, Sophia-Antipolis, France

Jason H. Moore, Computational Genetics Laboratory, Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire

Fouzia Moussouni, Université de Rennes 1, Rennes, France

Mohamed Nadif, LIPADE, University of Paris-Descartes, Paris, France

Radhika Nair, Department of Computer Science and Engineering, Ira A. Fulton Engineering, Arizona State University, Tempe, Arizona

Jean-Christophe Nebel, Faculty of Science, Engineering and Computing, Kingston University, London, United Kingdom

Alioune Ngom, School of Computer Science, University of Windsor, Windsor, Ontario, Canada

Thuy Diem Nguyen, School of Computer Engineering, Nanyang Technological University, Singapore

Oleg Okun, SMARTTECCO, Stockholm, Sweden

José Luis Oliveira, DETI/IEETA, University of Aveiro, Portugal

Hatice Gülçin Özer, Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio

Evangelos Pafilis, Institute of Marine Biology Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Crete, Greece

Jong C. Park, Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

Nicolas Pasquier, Laboratory I3S, University of Nice Sophia-Antipolis, Sophia-Antipolis, France

Chintan Patel, Department of Computer Science and Engineering, Ira A. Fulton Engineering, Arizona State University, Tempe, Arizona

Yudi Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden

Ruggero G. Pensa, Department of Computer Science, University of Torino, Turin, Italy

Giuseppe Pigola, IGA Technology Services, Udine, Italy

Luca Pinello, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts; and I.E.ME.S.T., Istituto Euro Mediterraneo di Scienza e Tecnologia, Palermo, Italy

Solon P. Pissis, Department of Informatics, King's College London, Strand, London, United Kingdom

Alberto Policriti, Department of Mathematics and Informatics and Institute of Applied Genomics, University of Udine, Udine, Italy

Ignacio Ponzoni, Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC), Dept. Computer Science and Engineering, Universidad Nacional del Sur, Bahía Blanca, Argentina and Planta Piloto de Ingeniería Química (PLAPIQUI) CONICET, Bahía Blanca, Argentina

Alfredo Pulvirenti, Department of Clinical and Molecular Biomedicine, University of Catania, Italy

Shoba Ranganathan, Department of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, Australia

Hendrik Rohn, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Haifa Ben Saber, Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE) and University of Tunis, Tunisia

Lee Sael, Department of Computer Science, Purdue University, West Lafayette, Indiana and Department of Biological Sciences, Purdue University, West Lafayette, Indiana

Ali Salehzadeh-Yazdi, Laboratory of Systems Biology and Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

Rodrigo Santamaría, Department of Computer Science and Automation, University of Salamanca, Salamanca, Spain

Bertil Schmidt, Institut für Informatik, Johannes Gutenberg University, Mainz, Germany

Falk Schreiber, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany and Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany

Khedidja Seridi, INRIA Lille Nord Europe, Villeneuve d'Aseq, France

Kailash Shaw, Department of CSE, Gandhi Engineering College, Bhubaneswar, Odisha, India

Simon J. Sheather, Department of Statistics, Texas A&M University, College Station, Texas

Stephen A. Smith, Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan

Junilda Spirollari, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ

Alexandros Stamatakis, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

El-Ghazali Talbi, INRIA Lille Nord Europe, Villeneuve d'Ascq, France

Kean Ming Tan, Department of Statistics, Purdue University, West Lafayette, Indiana

Xin Lu Tan, Department of Statistics, Purdue University, West Lafayette, Indiana

Bahar Taneri, Department of Biological Sciences, Eastern Mediterranean University, Famagusta, North Cyprus and Institute for Public Health Genomics, Cluster of Genetics and Cell Biology, Faculty of Health, Medicine and Life Sciences, Maastricht University, The Netherlands

Mingjie Tang, Department of Computer Science, Purdue University, West Lafayette, Indiana

Ahmed Y. Tawfik, Information Systems Department, French University of Egypt, El-Shorouk, Egypt

Sukru Tikves, Department of Computer Science and Engineering, Ira A. Fulton Engineering, Arizona State University, Tempe, Arizona

George Tzanis, Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Filippo Utro, Computational Genomics Group, IBM T.J. Watson Research Center, Yorktown Heights, New York

Davide Verzotto, Department of Information Engineering, University of Padova, Padova, Italy

Francesco Vezzi, Department of Mathematics and Informatics and Institute of Applied Genomics, University of Udine, Udine, Italy

Alessia Visconti, Department of Computer Science, University of Torino, Turin, Italy

Ioannis Vlahavas, Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Jason T. L. Wang, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ

Penghao Wang, School of Mathematics and Statistics, The University of Sydney, Sydney, Australia

Dongrong Wen, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ

Pengyi Yang, School of Information Technologies, University of Sydney, Sydney, Australia

Jean Yee-Hwa Yang, School of Mathematics and Statistics, University of Sydney, Sydney, Australia

Yaning Yang, Department of Statistics and Finance, University of Science and Technology of China, Hefei, China

Zejun Zheng, Singapore Institute for Clinical Sciences, Singapore

Ling Zhong, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ

Bing B. Zhou, School of Information Technologies, University of Sydney, Sydney, Australia

Albert Y. Zomaya, School of Information Technologies, University of Sydney, Sydney, Australia

SECTION I

BIOLOGICAL DATA PREPROCESSING

PART A

BIOLOGICAL DATA MANAGEMENT

CHAPTER 1

GENOME AND TRANSCRIPTOME SEQUENCE DATABASES FOR DISCOVERY, STORAGE, AND REPRESENTATION OF ALTERNATIVE SPLICING EVENTS

BAHAR TANERI[1,2] and TERRY GAASTERLAND[3]
[1]Department of Biological Sciences, Eastern Mediterranean University,
Famagusta, North Cyprus
[2]Institute for Public Health Genomics, Cluster of Genetics and Cell Biology,
Faculty of Health, Medicine and Life Sciences, Maastricht University, The Netherlands
[3]Scripps Genome Center, University of California San Diego, San Diego, California

1.1 INTRODUCTION

Transcription is a critical cellular process through which the RNA molecules specify which proteins are expressed from the genome within a given cell. DNA is transcribed into RNA and RNA transcripts are then translated into proteins, which carry out numerous functions within cells. Prior to protein synthesis, RNA transcripts undergo several modifications including 5′ capping, 3′ polyadenylation, and splicing [1]. Premature messenger RNA (pre-mRNA) processing determines the mature mRNA's stability, its localization within the cell, and its interaction with other molecules [2]. In addition to constitutive splicing, the majority of eukaryotic genes undergo alternative splicing and therefore code for proteins with diverse structures and functions.

In this chapter, we describe the process of RNA splicing and focus on RNA alternative splicing. As described in detail below, splicing removes noncoding introns from the pre-mRNA and ligates the coding exonic sequences to produce the mRNA transcript. Alternative splicing is a cellular process by which several different combinations of exon–intron architectures are achieved with different mRNA products from the same gene. This process generates several mRNAs with different sequences from a single gene by making use of alternative splice sites of exons and introns. This process is critical in eukaryotic gene expression and plays a pivotal role in increasing the complexity and coding potential of genomes. Since alternative splicing presents an enormous source of diversity and greatly

Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data,
First Edition. Edited by Mourad Elloumi and Albert Y. Zomaya.

elevates the coding capacity of various genomes [3–5], we devote this chapter to this cellular phenomenon, which is widespread across eukaryotic genomes.

In particular we explain the databases for Alternative Splicing Queries (dbASQ), a computational pipeline we used to generate alternative splicing databases for genome and transcriptome sequences of various organisms. dbASQ enables the use of genome and transcriptome sequence data of any given organism for database development. Alternative splicing databases generated via dbASQ not only store the sequence data but also facilitate the detection and visualization of alternative splicing events for each gene in each genome analyzed. Data mining of the alternative splicing databases, generated using the dbASQ system, enables further analysis of this cellular process, providing biological answers to novel scientific questions.

In this chapter we provide a general overview of the widespread cellular phenomenon alternative splicing. We take a computational approach in answering biological questions with regard to alternative splicing. In this chapter you will find a general introduction to splicing and alternative splicing along with their mechanism and regulation. We briefly discuss the evolution and conservation of alternative splicing. Mainly, we describe the computational tools used in generating alternative splicing databases. We explain the content and the utility of alternative splicing databases for five different eukaryotic organisms: human, mouse, rat, frutifly, and soil worm. We cover genomic and transcriptomic sequence analyses and data mining from alternative splicing databases in general.

1.2 SPLICING

A typical mammalian gene is a multiexon gene separated by introns. Exons are relatively short, about 145 nucleotides, and are interrupted by much longer introns of about 3300 nucleotides [6, 7]. In humans, the average number of exons per protein coding gene is 8.8 [7]. Both introns and exons of a protein-coding gene are transcribed into a pre-mRNA molecule [1]. Approximately 90% of the pre-mRNA molecule is composed of the introns and these are removed before translation. Before the mRNA molecule transcribed from the gene can be translated into a protein molecule, there are several processes that need to take place. While in total an average protein-coding gene in human is about 27,000 bp in the genome and in the pre-mRNA molecule, the processed mRNA contains only about 1300 coding nucleotides and 1000 nucleotides in the untranslated regions (UTRs) and polyadenylation (poly A) tail. The removal of introns and ligation of exons are referred to as the splicing process or the RNA splicing process [1, 7]. Splicing takes place in the nucleus. Final products of splicing which are the ligated exonic sequences are ready for translation and are exported out of the nucleus [1].

1.2.1 Mechanism of Splicing

Simply, splicing refers to removal of intervening sequences from the pre-mRNA molecule and ligation of the exonic sequences. Each single splicing event removes one intron and ligates two exons. This process takes place via two steps of chemical reactions [1]. As shown in Figure 1.1, within the intronic sequence there is a particular adenine nucleotide which attacks the 5′ intronic splice site. A covalent bond is formed between the 5′ splice site of the intron and the adenine nucleotide releasing the exon upstream of the intron. In the second chemical reaction, the free 3′-OH group at the 3′ end of the upstream exon ligates with the 5′ end of the downstream exon. In this process, the intronic sequence, which contains an RNA loop, is released.