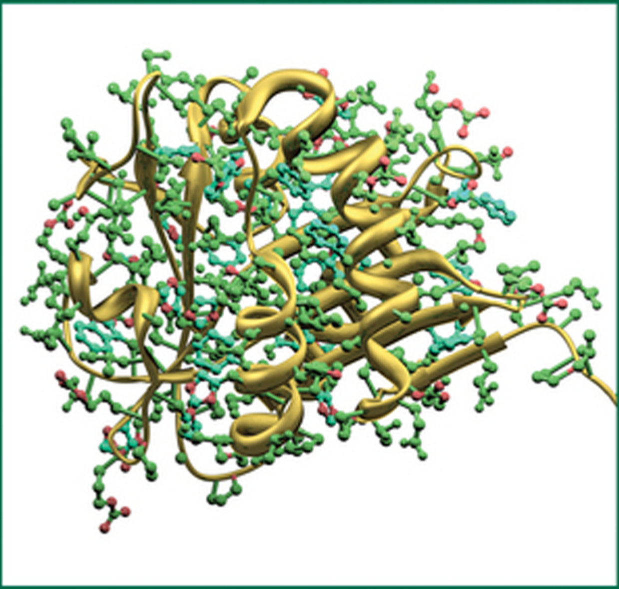
Wiley Series on Bioinformatics: Computational Techniques and Engineering

Yi Pan and Albert Y. Zomaya, Series Editors

Algorithmic and Artificial Intelligence Methods for Protein Bioinformatics



YI PAN
MIN LI
JIANXIN WANG

IEEE

IEEE computer society

WILEY

ALGORITHMIC AND ARTIFICIAL INTELLIGENCE METHODS FOR PROTEIN BIOINFORMATICS

Wiley Series on

Bioinformatics: Computational Techniques and Engineering

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

ALGORITHMIC AND ARTIFICIAL INTELLIGENCE METHODS FOR PROTEIN BIOINFORMATICS

Edited by

YI PAN
Georgia State University

JIANXIN WANG
Central South University, China

MIN LI
Central South University, China

WILEY | IEEE computer society

Cover design: John Wiley & Sons, Inc.
Cover illustration: © Martin McCarthy/iStockphoto

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 877-762-2974, 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:

Algorithmic and AI methods for protein bioinformatics / [compiled by] by Yi Pan, Jianxin Wang,
Min Li.
pages cm
ISBN 978-1-118-34578-8 (pbk.)
1. Proteins–Analysis–Mathematics. 2. Bioinformatics. 3. Artificial intelligence. I. Pan, Yi,
1960- editor of compilation. II. Wang, Jianxin, 1969- editor of compilation. III. Li, Min,
1978- editor of compilation. IV. Title: Algorithmic and artificial intellingence methods for protein
bioinformatics.
QP551.A4335 2012
572′.6–dc23

2012040076

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

III PROTEIN STRUCTURE ALIGNMENT AND ASSESSMENT

IV PROTEIN–PROTEIN ANALYSIS OF BIOLOGICAL NETWORKS

V APPLICATION OF PROTEIN BIOINFORMATICS

PREFACE

Proteins are any of a group of complex organic macromolecules that contain carbon, hydrogen, oxygen, nitrogen, and usually sulfur and are composed of one or more chains of amino acids. Proteins are fundamental components of all living cells and include many substances, such as enzymes, hormones, and antibodies, which are essential for the proper functioning of an organism. *Protein bioinformatics* is a newer name for an already existing discipline. It encompasses the techniques and methodologies used in bioinformatics that are related to proteins. Proteins can be described as a sequence, a two-dimensional (2D) structure, or a three-dimensional (3D) structure. In addition, interactions among proteins can be described as a network or a graph. Hence, many traditional algorithmic techniques such as graph algorithms, heuristic algorithms, approximate algorithms, parameterized algorithms, and linear programming can be applied to analyze protein interaction networks. On the other hand, because of the large amount of data available from wet labs and experiments with proteins, traditional algorithmic methods may not be sufficiently powerful and intelligent to be applied. Hence, we can use many mature machine learning or artificial intelligence (AI) methods to analyze protein data such as predicting protein structures based on existing databases or datasets. These AI techniques include support vector machines (SVMs), hidden Markov models (HMMs), neural networks, decision trees, reinforcement learning, genetic algorithms, pattern recognition, clustering, and random forests. Combinations of traditional algorithms such as graph algorithms, statistical methods, and AI techniques such as SVM have been used in protein structure prediction and protein interaction networks, and many good results have been achieved. The objective of this book is to promote collaboration between computer scientists working on algorithms and AI and biologists working on proteins by presenting cutting-edge research topics and methodologies in the area of protein bioinformatics.

This book comprises chapters written by experts on a wide range of topics that are associated with novel algorithmic and AI methods for analysis of protein data. This book includes chapters on analysis of protein sequences, structures, and their interaction networks using both traditional algorithms and AI methods. It comprehensively summarizes the most recent developments in this exciting research area. Protein bioinformatics plays a key role in life science, including protein engineering via designing tailor-made proteins, drug design based on finding docking molecules to kill disease cells, and improvement of protein effectiveness through modifying biocatalysts. Because of the many advantages of protein bioinformatics compared to traditional wet lab experiments, applications of protein bioinformatics are also described in this book. The important work of some representative researchers in protein bioinformatics is brought together for the first time in one volume. The topic is treated in depth and is related to, where applicable, other emerging technologies such as data mining and visualization. The goal of the book is to introduce readers to the most recent work and results in protein bioinformatics in the hope that they will build on them to make new discoveries of their own. It also arms the readers with the analysis tools and methods used in protein bioinformatics to enable them to tackle these problems in the future. The key elements of each chapter are briefly summarized below.

This is the first edited book dealing with the topic of protein bioinformatics and its applications in such a comprehensive manner. The material included in this book was carefully chosen for quality, coverage, and relevance. This book also provides a mixture of algorithms, AI methods, data preparation, simulation, experiments, evaluation methods, and applications, which provide both qualitative and quantitative insights into the rich field of protein bioinformatics.

This book is intended to be a repository of case studies that deal with a variety of protein bioinformatics problems and to show how algorithms and AI methods are used (sometimes together) to study the protein biological data and to achieve a better understanding of the data. It is hoped that this book will generate more interest in developing more efficient and accurate methodologies and solutions to protein bioinformatics problems and applications. This should enable researchers to handle more complicated and larger protein data once they understand the theories of the algorithms and AI methods described in this book and how to apply them. Although the material contained in this book spans a number of protein bioinformatics topics and applications, the chapters are presented in such a way that makes the book self-contained so that the reader does not have to consult external sources. This book offers (in a single volume) a comprehensive coverage of a range of protein bioinformatics applications and how they can be analyzed and used through the use of algorithms and AI methods to achieve meaningful results and interpretations of the protein data more accurately and efficiently.

The goal of this edited book is to provide an excellent reference for students, faculty, researchers, and people in the industry in the fields of bioinformatics, computer science, statistics, and biology who are interested in applying algorithms and AI methods to solve biological problems. This book is divided into five parts: (I) From Protein Sequence to Structure, (II) Protein Analysis and Prediction, (III) Protein Structure Alignment and Assessment, (IV) Protein–Protein Analysis of Biological Networks, and (V) Application of Protein Bioinformatics. The chapters are briefly summarized as follows:

- Chapter 1 discusses scaling of similarity sensitivity in remote homology modeling on yeast species and how the candidate genes are searched; these studies are important for different stages of embryogenesis of model plant species *Arabidopsis thaliana* in light of the concept of *dynamical patterning modules*.

- Understanding the biological term *sequence motif* is an important task in modern bioinformatics research, and these motif patterns may be able to predict the structural or functional area of other proteins. Protein sequence motif discovery is discussed in Chapter 2.
- Chapter 3 introduces methods for identifying calcium binding sites in proteins. Three methodologies for predicting calcium binding sites in proteins are reviewed and compared using different algorithms and AI methods.
- Chapter 4 proposes an imbalance learning method for protein methylation prediction using ensemble SVMs. It focuses on computational predictions of a particular posttranslational modification (PTM)–protein arginine methylation.
- Chapter 5 studies the prediction of protein posttranslational modification sites. By taking advantage of the large magnitude of experimentally verified PTM sites and utilizing a comprehensive machine learning method, a useful bioinformatics software system for PTM site prediction is provided.
- Chapter 6 describes an effective and a reliable tool using data mining and machine learning techniques for predicting local protein structure.
- In Chapter 7, a novel effective approach for predicting the boundaries of protein structure elements instead of individual residues structures using SVM is proposed.
- The states of the art of different machine learning-based RNA binding site prediction methods are overviewed in Chapter 8.
- In Chapter 9, many sequence-based and mass spectrometry data-based frameworks for determining disulfide bonds are presented.
- Chapter 10 gives the most recent update on protein contact order prediction. A new contact order web server is described that can predict the contact order by structure and sequence homology contrarily to the existing servers.
- Chapter 11 surveys about 15 computational methods for cysteine oxidation state prediction developed since the early 1990s.
- Chapter 12 addresses the computational methods in cryoelectron microscopy 3D structure reconstruction and its multilevel parallel strategy on the GPU platform.
- Chapter 13 gives a brief introduction to the biological, mathematical, and computational aspects of making pairwise comparisons between protein structures.
- To discover protein structures for optimal structure alignment, methods for using vector space model and suffix trees for efficient string matching and querying and how to index 3D protein structure are explained in Chapter 14. Furthermore, a protein similarity algorithm is explained in detail.

- Chapter 15 discusses several issues of structural alignment and methods that are implemented for sequence-order-independent structural alignment at both the global and local surface levels.
- Chapter 16 describes the methods used to study the prediction of protein structure classes and functions and measures, such as physicochemical features of amino acids, *Z*-curve representation, and the chaos game representation of proteins.
- Chapter 17 describes a new machine learning algorithm that uses a support vector machine (SVM) technique that understands structures from the Protein Databank (PDB) and, when given a new model, predicts whether it belongs to the same class as the PDB structures.
- The characteristics, strengths, and shortages of many network algorithms for clustering biological networks are discussed in Chapter 18. It includes various algorithms to cluster on protein–protein interaction networks (PPINs) based on the features of PPINs.
- Chapter 19 describes different algorithms applied to identify protein complexes, including methods based solely on PPIN data, methods combined with multiple information sources, and new trends in prediction of protein complexes on dynamic networks.
- To detect functional modules from protein–protein interaction networks, an ant colony optimization (ACO)-based algorithm with the topology of the network for the functional module detection is proposed and discussed in Chapter 20.
- Chapter 21 gives a brief overview of current state of the art in metabolic pathway/network alignment and how it can be used in automatic data curating.
- Chapter 22 starts by providing some background information on how PPI networks can be modeled on different PPI network alignments, and then focuses on local PPI network alignment algorithms and global PPI network alignment algorithms. Coarse-grain comparison is also addressed in that chapter.
- Among many machine learning techniques proposed for quantitative structure–activity relationship (QSAR) analysis and drug activity comparison, Chapter 23 focuses on the design and results of SVMs used for protein-related drug activity comparison.
- The main goal of Chapter 24 is to analyze how the general problem of finding repetitions in biological data evolved from sequences to networks data, by focusing on the open challenges and specific applications in biological networks.
- Chapter 25 gives a brief overview of an online resource and prediction server named MeTaDoR that provides comprehensive structural and functional information on membrane targeting domains.

- Chapter 26 gives a brief review of network-based identification and integration of gene signature of complex disease. In particular, it focuses on breast cancer gene signature in protein interaction networks using graph centrality.

We would like to express our sincere thanks to all authors for their effort and important contributions. We highly appreciate the reviews and corrections done by Ms. Tammie Dudley, which have improved the manuscript tremendously. We would also like to extend our deepest gratitude to Simone Taylor (senior editor) and Diana Gialo (editorial assistant) from Wiley for their guidance and help in finalizing this book. Finally, we would like to thank our families for their support, patience, and love. Without the collective effort of all of the above mentioned individuals, this book might be still in preparation. We hope that our readers will enjoy reading this book and give us feedback for future improvements.

Yi Pan
Department of Computer Science, Georgia State University, Atlanta, Georgia, USA
Email: pan@cs.gsu.edu

Jianxin Wang
School of Information Science and Engineering, Central South University, Changsha, China
Email: jxwang@mail.csu.edu.cn

Min Li
School of Information Science and Engineering, State Key Laboratory of Medical Genetics,
Central South University, Changsha, China
Email: limin@mail.csu.edu.cn

CONTRIBUTORS

Ajith Abraham, Machine Intelligence Research Labs, MIR Labs (Global Network)

Gulsah Altun, University of California, San Diego (UCSD), CA 92037

Vo Anh, School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia

Anamika Basu, Assistant Professor, Gurudas College, Kolkata, India

Piotr Berman, Department of Computer Science, Georgia State University, Atlanta, GA 30303

Nitin Bhardwaj, Department of Bioengineering, University of Illinois at Chicago, IL 60607

Curtis Harrison Bollinger, Computer Science Department, C.S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211

Mark Brandt, Department of Chemistry and Biochemistry, Rose-Hulman Institute of Technology, Terre Haute, IN 47803

Bernard Chen, University of Central Arkansas, Conway, AR 72032

Gang Chen, School of Information Science and Engineering, Central South University, Changsha 410083, China

Luonan Chen, Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Qiong Cheng, Department of Computer Science, Georgia State University, Atlanta, GA 30303

Anjum Chida, Georgia State University, Atlanta, GA 30303

Wonhwa Cho, Department of Chemistry, University of Illinois at Chicago, IL 60607

Bhaskar DasGupta, Department of Computer Science, University of Illinois at Chicago, IL 60607

Hai Deng, Sigma-Aldrich Bioinformatics, St. Louis, MO 63103

Zejin Ding, Georgia State University, Atlanta, GA 30303

Aiguo Du, Dow Chemical Company, 2301 N. Brazosport Blvd, Freeport, TX 77541

Joseph Dundas, Department of Bioengineering, University of Illinois at Chicago, IL 60607

Valeria Fionda, Free University of Bozen-Bolzano, Italy

Jianjiong Gao, Computer Science Department, C.S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211

Robert W. Harrison, Department of Computer Science, Georgia State University, Atlanta, GA 30303

Jieyue He, Southeast University, China

Allen Holder, Department of Mathematics, Rose-Hulman Institute of Technology, Terre Haute, IN 47803

Morten Källberg, Department of Bioengineering, University of Illinois at Chicago, IL 60607

Timothy Lee, San Francisco State University, San Francisco, CA 94132

Xiujuan Lei, Shanxi Normal University, Shanxi, China

Min Li, School of Information Science and Engineering, Central South University, Changsha 410083, China; State Key Laboratory of Medical Genetics, Central South University, Changsha 410078, China

Jie Liang, Department of Bioengineering, University of Illinois at Chicago, IL 60607

Guohui Lin, Department of Computing Science, University of Alberta, Edmonton, Canada

Hui Liu, Computer Science Department, Missouri State University, Springfield, MO 65897

Zhi-Ping Liu, Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Zhiyong Liu, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Hui Lu, Department of Bioengineering, University of Illinois at Chicago, IL 60607

William Murad, San Francisco State University, San Francisco, CA 94132

Tomáš Novosád, Department of Computer Science, VSB—Technical University of Ostrava, Ostrava 70833, Czech Republic

Suely Oliveira, Department of Computer Science, University of Iowa, Iowa City, IA 52242

Yi Pan, Department of Computer Science, Georgia State University, Atlanta, GA 30303

Xiaoqing Peng, School of Information Science and Engineering, Central South University, Changsha 410083, China

Simona E. Rombo, Universitá degli Studi di Palermo, Department of Mathematics, Computer Science Section, 90123 Palermo, via Archirafi 34, Italy

Anasua Sarkar, Assistant Professor, Government College of Engineering and Leather Technology, Kolkata, India

Lei Shi, State University of New York at Buffalo, NY 14214

Yi Shi, Department of Computing Science, University of Alberta, Edmonton, Canada

Yosi Shibberu, Department of Mathematics, Rose-Hulman Institute of Technology, Terre Haute, IN 47803

Rahul Singh, San Francisco State University, San Francisco, CA 94132

Václav Snášel, Department of Computer Science, VSB—Technical University of Ostrava, Ostrava, 70833, Czech Republic

Phang C. Tai, Georgia State University, Atlanta, GA 30303

Xiaohua Wan, Institute of Computing Technology and Graduate University, Chinese Academy of Sciences, Beijing, China

Jianxin Wang, School of Information Science and Engineering, Central South University, Changsha 410083, China

David S. Wishart, Department of Computing Science, University of Alberta, Edmonton, Canada

Dong Xu, Computer Science Department, C.S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211

Jack Y. Yang, Harvard University, Cambridge, MA 02140

Jian-Yi Yang, Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109

Qiuming Yao, Computer Science Department, C.S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211

Zu-Guo Yu, Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Hunan, China; Key Laboratory of Intelligent Computing and Information Processing, Ministry of Education of China, Xiangtan University, Hunan, China; School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia

Alexander Zelikovsky, Department of Computer Science, Georgia State University, Atlanta, GA 30303

Aidong Zhang, State University of New York at Buffalo, NY 14214

Fa Zhang, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Yan-Qing Zhang, Georgia State University, Atlanta, GA 30303

Wei Zhong, University of South Carolina Upstate, Spartanburg, SC 29303

Jianjun Zhou, Department of Computing Science, University of Alberta, Edmonton, Canada

Shao-Ming Zhu, Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Hunan, China

PART I

FROM PROTEIN SEQUENCE TO STRUCTURE

CHAPTER 1

EMPHASIZING THE ROLE OF PROTEINS IN CONSTRUCTION OF THE DEVELOPMENTAL GENETIC TOOLKIT IN PLANTS

ANAMIKA BASU and ANASUA SARKAR

The diversity in land plants due to size, shape, and form is the combined result of different developmental processes for adult plant formation from the zygote and their evolution. The ancestral patterning toolkit designed by Floyd [1] for land plants was constructed by using developmentally important gene families that are responsible for flowering plant growth, patterning, and differentiation. In this chapter we search for the candidate genes, which are important for different stages of embryogenesis of the model plant species *Arabidopsis thaliana* in light of the *dynamical patterning modules* concept. This is a novel idea that Newman [2] applied to build a *developmental genetic toolkit* by using a set of genes that mobilize the physical processes that are important in metazoan development. In this chapter we focus on a small number of gene families, which are related to physical characteristics, such as turgor pressure, asymmetric cell division, asymmetric distribution of cellular components, anisotropic expansion, and dynamics in merstemic cell maintenance and finally establish their evolutionary developmental (evo-devo) roles in land plant development with a functional/biological analysis.

1.1 INTRODUCTION

In the evolutionary history of the plant kingdom, the plants evolved through increasing levels of complexity, from a freshwater green alga, through bryophytes, lycopods, ferns, and gymnosperms to the complex angiosperms of

Algorithmic and Artificial Intelligence Methods for Protein Bioinformatics. First Edition.
Edited by Yi Pan, Jianxin Wang, Min Li.

today. Between about 480 and 360 million years ago (mya), from a simple plant body consisting of only a few cells, land plants (liverworts, hornworts, mosses, and vascular plants) evolved to an elaborate two-phase lifecycle with complex organs and tissue systems [3]. As mentioned above, the diversity of plant kingdom due to size, shape, and form is the combined result of different developmental processes for adult plant formation from the zygote, which evolved progressively [4].

Identifying the genes, which act in the developmental pathways and consequently in determining how they are modified during evolution, is the focus of the field of evolutionary developmental (evo-devo) biology. The fundamental aspects of the plant body plan have been found to be remarkably consistent within the plant kingdom irrespective of vast diversification [5]. Graham et al. [5] identified nine fundamental body plan features that originated during radiation of algae and were inherited by the plant kingdom. In light of these fundamental features, we analyze the evo-devo roles of embryogenesis in plants. In this chapter we select some candidate genes to construct a genetic toolkit for land plants, and establish their biological/functional significances with an evolutionary analysis.

1.2 EVOLUTIONARY DEVELOPMENTAL (EVO-DEVO) ROLES IN EMBRYOGENESIS OF PLANTS (IN DEVELOPMENTAL PLANT GENETIC TOOLKIT FORMATION)

To explore the evo-devo biology of embryogenesis in plants, we select some candidate genes and relate them to their physical properties to depict the developmental genetic toolkit of land plants, following studies by Newman [2]. Table 1.1 shows the candidate genes, their physical principles, and their relevant evo-devo roles.

1.3 PHASES IN EMBRYOGENESIS IN *Arabidopsis Thaliana*

In *embryogenesis*, a multicellular organism forms from a single cell. In *Arabidopsis thaliana*, embryogenesis is a continuous process that can be divided into three major phases, described as early, mid, and late. In this section, we review all these phases starting from the early phase of embryogenesis.

The early phase consists of pattern formation, morphogenesis, defining the axes of the plant body plan, and forming the organ systems. Embryogenesis involves three basic steps: (1) cell growth, (2) cell differentiation, and (3) morphogenesis. The embryo grows up to certain limit, and then differentiates into cells that differ from their mother cell in structure and function. Thus different morphological structures such as stem, root, or flower are formed, enabling the formation of a total plant.

TABLE 1.1 Relationship between Candidate Genes and Their Physical Properties and Evolutionary Developmental (Evo-Devo) Roles

Characteristic Molecules	Physical Principles	Evo-Devo Roles
Expansin	Turgor pressure	Cell wall expansion and organ initiation
Extensin	Turgor pressure	Growth initiation by cell plate formation
GNOM	Asymmetric cell division	Apical–basal axis formation
TORMOZ	Longitudinal cell division	Pattern formation
PIN	Asymmetric distribution	Tissue polarization
ACTIN	Anisotropic expansion	Cellular polarity determination
G proteins	Dynamics in meristem maintenance (differential cell division)	Lateral organ formation
NPH4 TF	Differential cell expansion	Aerial tissue formation
AGO10	Complementary sequence binding	Adaxial–abxaial polarity formation
FT protein	Signaling	Floral morphogenesis
LEAFY	Intercellular communication through gradient formation	Flower patterning

1.3.1 Cell Growth Phase

The growth in plants is defined as an irreversible increase in cell mass [6]. Since the cell mass value factors both in cell volume and cell number, there are two processes relevant to the plant growth: (1) an irreversible increase in cell size, known as *cell elongation*, and (2) an increase in cell number, defined as *cell division*.

1.3.1.1 Cell Elongation Phase Cell expansion, driven by the turgor pressure, mediates the plant growth. During this process, cells increase manyfold in volume and become highly vacuolated. The cell membrane and the cell wall surround the plant cell, whereas for an animal cell, only cell membrane is present. Since the structure of the cell wall is more rigid than that of the cell membrane, the cell growth mechanism differs in plants and animals. Because of the presence of the rigid cell wall, no cell migration occurs during plant development. During embryogenesis, the zygote elongates 3 times before cell division as a result of the extension of the cells along the sidewalls. Before cell division, the cell nucleus moves to the position of the new cell walls to be formed [7]. Two cells are formed from the zygote following the asymmetric cell division: a small apical cell and a larger basal cell with different cell contents. The dense cytoplasm is present in the apical cell and in most of the basal cell vacuole (see Fig. 1.1).

The steps listed below (1–6) in generation of the embryo proper and the suspensor from the apical and the basal cell, respectively, are shown in Figure 1.2a:

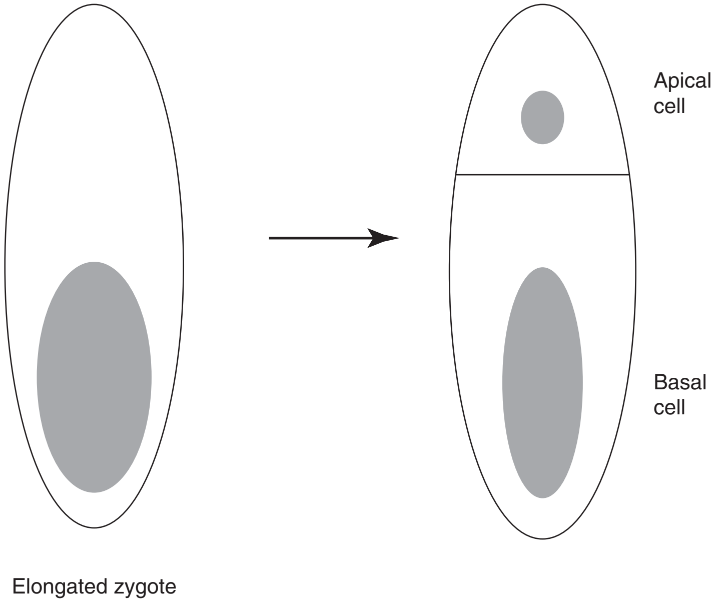


Figure 1.1 Asymmetric cell division of the zygote.

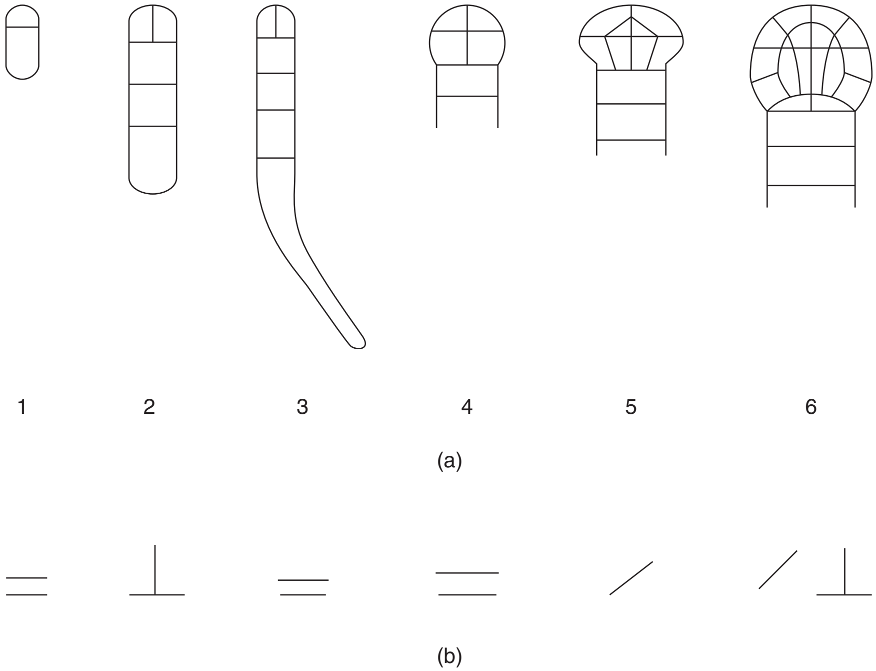


Figure 1.2 Generation of the embryo proper and the suspensor: (a) the steps involved in formation of the embryo proper and the suspensor; (b) the plane of cell division in each step: horizontal —, longitudinal ⊥, and transverse /.

1. The zygote undergoes a cell division along the horizontal axis, producing a smaller apical cell (1) and a larger basal cell.
2. The apical cell undergoes a cell division along the longitudinal ⊥ axis to form the embryo proper, and the basal cell develops the suspensor by cell division along the horizontal axis.
3. Cell expansion occurs in the suspensor.
4. The embryo proper divides by another horizontal division, perpendicular to the plane of the previous division. The basal portion of the suspensor is not shown.
5. The four quadrants of the embryo proper have divided by transverse divisions to produce an eight-cell embryo proper.
6. The cells of the embryo proper have divided by some transverse and two longitudinal cell divisions to form a 16-cell embryo proper.

The topmost cell of the suspensor has divided transversely to produce the hypophysis.

The zygote cell initially divides along apical–basal axis. Then the cell growth or expansion occurs. By a few successive cell divisions along different planes, the apical cell and the basal cell, respectively, form the embryo proper and the suspensor. In the embryo proper, eight cells are formed by two vertical and one horizontal cell divisions, whereas the basal cell divides along vertical axis only. The embryo proper will form most of the mature embryo, which are cotyledons, shoot meristem, hypocotyl regions, and part of the radical (or embryogenic) root. From the basal cell, the suspensor and the hypophysis are formed. The hypophysis contributes to the root meristem. *Thus, cell growth and expansion in a specific direction and selection of a division plane, all play important roles in plants morphogenesis*.

The irreversible increase in cell volume and surface area, as in the cell wall, are determinants of plant cell growth. Both intrinsic and extrinsic factors are required for cell growth. Both light and gravitational forces act as extrinsic factors, whereas the turgor pressure and many other biomolecules are intrinsic factors. Turgor pressure can be generated by water uptake of growing cells in their vacuoles. Since this pressure is homogeneous and multidirectional, the cell wall expands uniformly over the whole cell. Cell wall expansion consists in two processes: (1) stretching of the cell wall resulting from turgor pressure and (2) deposition of new material by the cell membrane in a specific direction. These two processes are related. During cell growth, the cell wall is stretched by turgor pressure because interpolymeric bonds in the wall naturally break and re-form. Thus the polymers forming cell wall, under tension from the turgor pressure, tend to slip past each other irreversibly to enlarge the cell wall. When the cell wall is stretched (normally 10–100-fold), it does not become thinner because new material is deposited to resist the strain of turgor pressure [8]. So the growing cell secretes substances to form a polymeric structure of crystalline cellulose microfibrils, embedded in a hydrophilic matrix, which are composed of hemicelluloses and pectins.

1.3.1.1.1 Role of Expansin and Extensin Proteins in Cell Elongation Cosgrove [9] proposed the role of protein *expansins* in cell wall expansion by slippage or rearrangement of matrix polymers. Plant cell wall expansion occurs more rapidly at low pH. This event is known as *acid growth*. The primary cell wall of plant is composed of xylans (homopolymers of xylose), xyloglucans (heteropolymers whose backbone consists of glucose and sidechains with xylose), and mixed linked glucans (homopolymers of glucose). Xlycan (polysaccharide) molecule are present in the cell wall. They can stick to each other, as well as connect the cellulose microfibrils to each other. Expansins disrupt both types of bonding: between xlycans and between xlycan and cellulose. In presence of turgor pressure, expansins cause the displacement of wall polymers and a slippage occurs at the point of polymer adhesion. In shoot apical meristem, the expansins are expressed, and these proteins are present in primordium forming cells. During leaf initiation, the cell wall is altered by expansins. Thus, in presence of the turgor pressure, the cell wall of meristem expands in a specific direction; For example, it may bulge outward to form primodium. This is the regulatory role of the cell wall protein (expansin) in leaf initiation (plant morphogenesis) [10].

Extensins are a family of hydroxyproline-rich glycoproteins (HRGPs), which are found in the cell walls of higher plants. They play important roles in determining the cell shape and formation of the division plane during cell division. In dicots, extensins have a repeated pentapeptide Ser–$(Hyp)_4$ structure. It has been proposed that a positively charged extensin scaffold reacts with acidic pectin to form extensin pectate. This can act as the template for a newly synthesized cell plate during cell division. Thus extensins play an essential role in growth initiation of plants.

The *RSH* gene is a lethal mutant in *Arabidopsis* embryogenesis that encodes extensin AtEXT3, a structural glycoprotein located in the nascent cross wall or “cell wall” and also in mature cell walls. *RSH* is essential for the correct positioning of the cell plate during cytokinesis in cells of the developing embryo. *RSH* is detected in the first asymmetric cell division of the zygote [11]. In the *RSH* mutant, both apical and basal cells are formed, but the position of the plane of the newly synthesized cell wall is different, which may be compared to that of wild types. As a result, this event forms either same sized apical and basal cells or larger apical cells. This type of mutant also affects embryonic development in later stages, such as the globular stage. Cell division occurs during this stage, but the specific plane of division, as shown in Figure 1.2b, is not maintained. Therefore, cell divisions occur randomly. These stop the normal bilateral symmetry of embryogenesis.

The major component of the plant cell wall is polysaccharides (80% present in *Arabidopsis*). Pectin and hemicellulose are the main polysaccharides. In the mature cell wall, pectin is present in middle lamella, which adheres two adjacent plant cells to each other. Hemicellulose is involved in cell expansion, cell growth, and thus in cell shape formation. Both pectin and hemicellulose are synthesized and modified in Golgi complex. After packing into different vesicles (specific for each component), they are targeted to different domains of the cell wall. This type of physical segregation of components of cell wall is important for the normal growth and the development of plants. Extensin proteins, as mentioned earlier, are modified in *cis*-Golgi network and through secretary vesicles delivered to *trans*-Golgi network and ultimately to cell wall. Thus cell growth occurs. This explains the importance of selecting expansin and Extensin proteins in the genetic toolkit for plant cell elongation.

1.3.1.2 Cell Division Phase After elongation in a specific direction, cell division occurs in a specific plane, so that cell partition occurs along its longest axis. This is observed in the first cell division of the zygote, where apical–basal axis is formed by an asymmetric cell division.

1.3.1.2.1 Role of GNOM for Normal Cell Division From the very first cell division, Shevell et al. [12] observed that orientation of the plane of cell division, the rate of cell division, and GNOM gene products control the direction and amount of cell expansion. In the GNOM mutant, instead of asymmetric cell division in the zygote, two nearby equal-sized cells are formed. As a result of the distorted orientation of the plane of cell division, an altered octant embryo is formed with twice the number of cells found in the wild type. No cotyledonary primodia forms from apical cell. The root development is inhibited in the GNOM mutant because the basal cell fails to form the hypophysis. Very little elongation occurs in mutant cells with abnormal vacuoles [12]. In dark, wild-type hypocotyls elongate in the longitudinal direction. But for the dark-grown GNOM mutant, expansion occurs in both horizontal and longitudinal directions [12]. Thus, in the GNOM mutant, regulation on the direction of cell elongation is absent. Another important observation is that GNOM mutant cells can be separated easily from one another. This may be due to the inappropriate deposition of pectin or its derivatives in cell wall. From these observations, we can conclude that GNOM is essential for normal cell division, cell expansion, and cell adhesion in *Arabidopsis*.

ARF (ADP ribosylation factor) proteins are small guanine–nucleotide binding proteins. They are interconverted between two forms, such as GTP-bound ARFs and GDP-bound ARFs, by guanine–nucleotide exchange factors (GEFs) and by GTPase activating proteins (GAPs), respectively [13]. GNOM is a guanine–nucleotide exchange factor of ARF class proteins. It is present in the cytosolic face of endosomes. Generally ARFs are responsible for formation of vesicle coats, which are necessary for the formation of transport vesicles from donor compartments (i.e., vesicle budding) and cargo selection for transmembrane proteins. Here, GNOM proteins are involved in coat recruitment to endosomes for PIN1 protein targeting. PIN1 is a transmembrane protein, which is responsible for auxin hormone transport. GNOM proteins, as a regulator of intracellular vesicle trafficking, controls the rapid cycling of PIN1 proteins between the basal domain of plasma membrane and the endosomal compartments. Thus the embryos lacking the GNOM protein fail to establish the coordinated polar localization of the auxin–efflux regulator PIN1. This explains our selection of GNOM protein for the evolutionary–developmental (evo-devo) role of cell division in the plant genetic toolkit.

1.3.2 Cell Differentiation Phase

The term *differentiation* with respect to the plant cell refers to the property of plant cells to form quantitatively different specialized cell groups. During this phase, a number of cells that are derived from a single progenitor cell or a group of cells are qualitatively different in their contents and are specialized for different functions [6].

Several different mechanisms are involved in plant cell differentiation:

1. Signaling through cell wall–associated determinants (e.g., PIN)
2. Polarity by F-actin
3. Asymmetric cell division
   a. Differential cell division by G protein
   b. Asymmetric cleavages in different cell planes by TORMOZ
   c. Differential expansion rate by NPH4 transcription factor
4. Micro-RNA regulation by ZWILLE

In the next section, for each of these mechanisms, we evaluate the importance of associated proteins, to form the plant genetic toolkit.

1.3.2.1 Role of PIN for Signaling through the Cell Wall (Associated Determinants)

Cell polarity is an asymmetric distribution of the cellular components with respect to an arbitrary axis inside the cell [14]. An important example of plant polarity is the apical–basal polarity of the PIN family of auxin efflux facilitators, which forms the organization of the entire plant body. In plants, the matured embryo contents a main axis of polarity, with the shoot meristem flanked by the cotyledons (embryonic leaves) at the top end and separated by hypocotyl (embryonic stem) and root from the root meristem at the opposite pole [15]. This is due to tissue polarization (in addition to cell polarization) during plant development, which is necessary for organogenesis. Auxin hormone as a key regulator of tissue polarization controls its own polar transport in response to internal factors (various transcription factors) and external factors (light, gravity). Thus auxin is distributed asymmetrically along matured embryo.

Auxin is a multifunctional phytohormone that controls various developmental processes, such as cytoskeleton organization, intracellular membrane trafficking,