Machine Learning Techniques in Plant Biology

Chapter · March 2015

DOI: 10.1007/978-81-322-2172-2_26

3 authors:

Khwaja Osama
Integral University
10 PUBLICATIONS **58** CITATIONS
SEE PROFILE

Bhartendu Nath Mishra
Institute of Engineering & Technology - Lucknow
89 PUBLICATIONS **939** CITATIONS
SEE PROFILE

Pallavi Somvanshi
TERI School of Advanced Studies,New Delhi, India
80 PUBLICATIONS **400** CITATIONS
SEE PROFILE

Debmalya Barh · Muhammad Sarwar Khan
Eric Davies *Editors*

PlantOmics: The Omics of Plant Science

Springer

Debmalya Barh
Muhammad Sarwar Khan • Eric Davies
Editors

PlantOmics: The Omics of Plant Science

Springer

Editors
Debmalya Barh
Department of Genomics
Institute of Integrative Omics and Applied Biotechnology (IIOAB)
Nonakuri, West Bengal, India

Muhammad Sarwar Khan
Centre of Biochemistry and Biotechnology
University of Agriculture
Faisalabad, Pakistan

Eric Davies
Department of Plant Biology
North Carolina State University
Raleigh, NC, USA

ISBN 978-81-322-2171-5 ISBN 978-81-322-2172-2 (eBook)
DOI 10.1007/978-81-322-2172-2

Library of Congress Control Number: 2015930013

Springer New Delhi Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer (India) Pvt. Ltd. is part of Springer Science+Business Media (www.springer.com)

My loving Gul: Shaurya Shree

Foreword

Plants are the primary source of "most of our food, fuel, fibers, fabrics, and pharmaceuticals". Stresses (biotic and abiotic) are the major threats to plants, being the primary cause of crop yield losses worldwide. On the other hand, with the global population expected to reach nine billion by 2050, an increase in crop productivity and quality will be needed to meet the requirements. Each of the 29 chapters of this *PlantOmics: The Omics of Plant Science* book opens a door to exciting cutting-edge omics approaches and their applications to meet the future demands.

The flow of the chapters in the book is highly scientific and strategically organized to be easy to go. It starts with the topic omics approaches in model plants and their applications in improvement of maize and rice like major cereal crops. Chapters 2, 3, and 4 describe very important technologies such as spectroscopy (NIR, MIR, Raman), next generation sequencing (NGS), and functional genomics and their applications in current plant science. Chapters 5 and 6 deal with technical advancements and applications of cyto-mutagenomics and epigenomics in crop improvement. Chapter 7 gives a detailed account on plant miRNA biology, associated technologies, and their tailor-made applications to improve plant stress response.

Each topic dealt in Chaps. 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, and 25 is a unique imprint of this book. These chapters cover well established and several budding omics areas in plant science such as Plant Proteomics, Metabolomics, Glycomics, Lipidomics, Secretomics, Phenomics, Cytomics, Physiomics, Signalomics, Thiolomics, Organelle Omics, Micromorphomics, Microbiomics, Cryobionomics, Nanobiotechnology, and Plant Pharmacogenomics. Each of these chapters describes the latest technologies and applications of the respective omics in a very comprehensive way; therefore, they are up to date, easy to understand, and can be spontaneously adopted to expand the area of our research and development.

Chapters 26, 27, and 28 deal with computational and systems biology approaches in plant science making the book more useful to any kind of plant biology research, whether in a wet lab or *in silico*. The last chapter (Chap. 29) is very brief but interesting where the editors have provided valuable insights on the future directions of omics and plantomics. They have proposed several new areas in omics which we must explore towards development of an integrated meta-omics strategy to ensure the world and earth's health and related issues.

Overall, it is a great effort by Dr. Barh, his editorial team, and 90 expert contributors from 15 countries to make this highly resourceful, up-to-date, thought provoking, and worthwhile unique book for students and researchers in the field of cutting-edge plant omics sciences. I highly recommend the book for keeping you up to date in the field.

Professor Ana Paulina Barba de la Rosa, PhD
Molecular Biology Division
Head, Unit of Proteomics and Molecular Biomedicine
Instituto Potosino de Investigación
Científica y Tecnológica (IPICyT), Mexico
President, Mexican Proteomics Society, Mexico

Preface

The term "*omics*" depicts completeness. In the last two decades, the term has been suffixed with several biological topics to provide complete information on the subject. With the advent of new technologies, the arena of "*omics*" is increasing rapidly. However, most of the currently available books that deal with omics technologies and their applications are mainly focused on animal system. To fill this gap, we have introduced this *PlantOmics: The Omics of Plant Science* book to provide a complete spectrum of plant related omics to the students and researchers working in the field of cutting-edge plant molecular biotechnology. Equal focus has been given to the technological advancements as well as their specific applications. Therefore, the book provides a comprehensive account of the state-of-the-art latest developments and trends of *omics* approaches in plant science. Several topics have also highlighted the integrative omics strategies enabling the cost-effective development of superior plants for various purposes.

The book consists of *29 chapters* written by *90 experts* from *15 countries* that represent three-fourths of the globe. In the introductory chapter (Chap. 1), Dr. Agrawal and colleagues have described the omics of model plants where genomics, proteomics, transcriptomics, and metabolomics of model plants such as *Arabidopsis*, rice, and maize are dealt in detail. Further, this chapter also provides how these technology derived knowledge can be used for transgenomics, mapping for biotic and abiotic stresses, and marker assisted selection for crop improvement. In Chap. 2, Dr. Cozzolino's group has given a nice overview on the most commonly used spectroscopy techniques such as NIR, MIR, and Raman in plant omic analysis. To make the chapter more resourceful, Dr. Cozzolino has also demonstrated instrumentations and analytic software for these spectroscopy techniques. The hot topic, next generation sequencing (NGS), its technologies, various platforms, algorithms, and *de novo* assembly, annotation, and analysis of plant genome are given by Dr. Tiwary in Chap. 3. Chapter 4, by Dr. Jha and his colleagues, provides a comprehensive account of techniques associated with plant functional genomics and their applications. Drs. Talukdar and Sinjushin in Chap. 5 have described various techniques of cytogenomics and mutagenomics and their cost effective applications in plant breeding and biology. This chapter has also highlighted the mutations that cause alterations in antioxidant defense response to withstand diverse abiotic stresses to reveal intrinsic cellular and metabolic events towards sensitivity of seed plants to salinity, drought, metal toxicity

and other stresses, prospecting to formulate effective breeding strategies in different agro-climatic conditions. Epigenomics technologies and their potential applications in crop improvement are summarized in Chap. 6 by Drs. Shafiq and Khan. Especially, this chapter highlights the roles of chromatin remodeling mechanisms in response to environmental stimuli and their role in crop improvement. Chapter 7, by Dr. Boopathi, on Plant miRNomics gives a comprehensive account to explain how the miRNAs fine tune the gene expression and play key roles in developmental timing and patterning of structures in response to external and internal stimuli in plants. This chapter also provides how the miRNAs can be used to improve plant stress responses. Chapter 8, by Dr. Agrawal and his group, describes the recent technological progresses in plant proteomics and highlights the achievements made in understanding the plant proteomes and their applications. In Chap. 9, Dr. Sangwan and colleagues explain various technology platforms in plant metabolomics research and how the metabolomics is used in monitoring and assessing gene functions, stress responses, and to characterize post-genomic processes from a broad perspective along with the challenges the domain is facing. Dr. Khurana's group in Chap. 10 overviews the chemistry and technologies in plant glycomics. This chapter also gives summary of applications of glycomics in biopharming and several biological processes such as plant signaling, stress responses, and immunity. In the next chapter (Chap. 11), Dr. Namasivayam elucidates the chemistry and analytic technologies, lipid signaling in plants, lipidomes in plant defense mechanisms, and several other aspects of plant lipidomics. The comprehensive mechanisms regulating constitutive and induced secretome of diverse plants and their habitat along with technological approaches are discussed by Dr. Yadav and her group in Chap. 12. In Chap. 13, Dr. Rahman and colleagues give a detailed account on integrated-omics approaches in phenomics and its applications in plant and agriculture. Chapter 14, by Drs. Davies and Stankovic, describes how novel methods based on super-fast and super-resolution microscopy can be used in describing proteins, nucleic acids, cytoskeleton, and small molecules of major interest to plants. In Chapter 15, Dr. Karpiński and colleagues educate us on plant physiomics. The chapter provides insights on how the combined molecular-physiological events drive plant growth, development, acclimatization, and defense responses. Dr. Vian et al., in Chap. 16, have introduced the term "Signalomics" and have shown how novel methods can be used to analyze systemic signals including electrical and hydraulic signals in plants. In Chap. 17, Dr. Talukdar and colleagues elucidate the use of latest cutting-edge functional genomics tools to understand the plant thiol metabolism from source (soil) to sink (grains) in diverse arenas of "thiolomics". The next three chapters (Chaps. 18, 19, 20) are dedicated to organelle omics. Chapter 18, by Dr. de Luna Valdez et al., explores how chloroplasts organize their genomes and regulate their transcriptomes, proteomes, and metabolomes, trying to focus on classical knowledge and reviewing new datasets obtained through large-scale research projects and systems approaches that shed light on chloroplast functionality under the chloroplast omics chapter. In Chap. 19, Dr. Khan summarizes the developments from plastid genomics to gene expression and briefly describes how transplastome facilitates expression of

vaccines, therapeutics, and plantibodies, in addition to tailoring agronomic traits in plants. Plant mitochondrial omics (Chap. 20), by Dr. Mustafa and his colleagues, describes a detailed account on regulation of mitochondrial genes at transcriptional, post-transcriptional (splicing and RNA editing), translational, and post-translational levels in omics perspective. Chapter 21 describes "Micromorphomics", a term coined by Dr. Tulika Talukdar to explain how plants combat environmental stresses through collective morphological manifestations in their organs architectures. Chapter 21 is dedicated to microbiomics. In this chapter, Dr. Sharma's team has discussed technologies to identify new groups of microorganisms involved in plant diseases from microbiome of rhizosphere and roles of microbiome in plant health and related areas. Drs. Martinez-Montero and Harding in Chap. 22 (Cryobionomics) intend to explore the connections between stability and cryogenic/non-cryogenic stress factors with a view to aiding protocol improvement, optimization, and validation for plant genetic resources conservation with several examples. Chapter 24, by Dr. Kazi and colleagues, focuses on the development and use of "nanotechnology" for formulating agriculturally important chemicals (fertilizers) with more useful properties and their direct delivery as well as their applications in various agricultural sectors. Chapter 25, by the same group, systemically analyzes the recent developments in plant pharmacogenomics and its contributions in the field of molecular and pharmaceutical sciences. Dr. Somvanshi and colleagues in Chap. 26 have attempted to describe several machine learning approaches and their applications in plant biology in a very simple way. Similarly, in Chap. 27, Dr. Sarika's team has emphasized on a number of applications of bioinformatics in agriculture in view of functional genomics, data mining techniques, genome-wide association studies, high-performance computing facilities in agriculture, and various bioinformatics tools/databases important for breeders, biotechnologists, and pathologists. Chapter 28 (Plant systems biology), by Drs. Bhardwaj and Somvanshi, describes recent insights and advancements in systems biology approaches in order to understand how plant systems work. In the brief concluding chapter (Chap. 29), we, the editors, have proposed several omics terms under "*Futuromics*" centraling Plantomics to direct the future perspectives of plant omics in meta-omics era.

We believe that this book will be a valuable resource to all who are working on cutting-edge plant omics. We appreciate your comments and suggestions to improve the next edition.

Nonakuri, India — Debmalya Barh
Faisalabad, Pakistan — Muhammad Sarwar Khan
Raleigh, NC, USA — Eric Davies

Contents

Contributors

Lalit Agrawal, Ph.D. National Botanical Research Institute, Lucknow, India

Pawan Kumar Agrawal, Ph.D. Vivekananda Parvatiya Krishi Anusandhan Sansthan (VPKAS), Almora, India

Mustafeez Mujtaba Babar, Ph.D. Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

B. Kalyana Babu, Ph.D. Vivekananda Parvatiya Krishi Anusandhan Sansthan (VPKAS), Almora, India

Debmalya Barh, Ph.D. Department of Genomics, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, West Bengal, India

Tulika Bhardwaj, Ph.D. Department of Biotechnology, Teri University, New Delhi, India

Jayant S. Bhat, Ph.D. Regional Research Centre, Indian Agriculture Research Institute, Dharwad, Karnataka, India

N. Manikanda Boopathi, Ph.D. Department of Plant Molecular Biology and Bioinformatics, Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India

Vijaya Brahma, Ph.D. Toronto General Research Institute, University Health Network, Toronto, ON, Canada

Paweł Budiak, Ph.D. Department of Genetics, Breeding and Biotechnology, Warsaw University of Life Sciences, Warszawa, Poland

Sushma Chauhan, Ph.D. Department of Chemical and Biochemical Engineering, Dongguk University, Seoul, South Korea

Daniel Cozzolino, Ph.D. School of Agriculture, Food and Wine, The University of Adelaide, Osmond, SA, Australia

Eric Davies, Ph.D. Department of Plant Biology, North Carolina State University, Raleigh, NC, USA

L.A. de Luna-Valdez, Ph.D. Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

R.K. Dubey, Ph.D. Department of Floriculture and Landscaping, Punjab Agricultural University, Ludhiana, Punjab, India

S. Encarnación-Guevara, Ph.D. Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

Alberto Fassio, Ph.D. National Institute for Agricultural Research, INIA La Estanzuela, Colonia, Uruguay

Piotr Gawroński, Ph.D. Department of Genetics, Breeding and Biotechnology, Warsaw University of Life Sciences, Warszawa, Poland

Mustafa Malik Ghulam, Ph.D. Agriculture Biotechnology, National Institute for Biotechnology and Genetic Engineering (NIBGE), Faisalabad, Pakistan

Magdalena Górecka, Ph.D. Department of Genetics, Breeding and Biotechnology, Warsaw University of Life Sciences, Warszawa, Poland

A.A. Guevara-García, Ph.D. Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

Deepti Bhushan Gupta, Ph.D. TERI University, Vasant Kunj, New Delhi, India

Keith Harding, Ph.D. Damar Research Scientists, Damar, Cuparmuir, Fife, Scotland, UK

Firoz Hossain, Ph.D. Division of Genetics, Indian Agriculture Research Institute, New Delhi, India

M.A. Iquebal, Ph.D. Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi, India

N. Jagadeeshselvam, Ph.D. Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India

Kaliaperumal Jagatheesh, Ph.D. Department of Biotechnology, Periyar University, Salem, India

Sarika Jaiswal, Ph.D. Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi, India

Johnravindar D, Ph.D. Department of Biotechnology, Periyar University, Salem, India

Richard L. Jayaraj, Ph.D. Department of Biotechnology, Periyar University, Salem, India

Uday Chand Jha, Ph.D. Indian Institute of Pulses Research (IIPR), Kanpur, India

Stanisław Karpiński, Ph.D. Department of Genetics, Breeding and Biotechnology, Warsaw University of Life Sciences, Warszawa, Poland

Alvina Gul Kazi, Ph.D. Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Maria Khalid, Ph.D. Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Abdul Rehman Khan, Ph.D. Department of Environmental Sciences, COMSATS I.I.T., Abbottabad, Pakistan

Muhammad Sarwar Khan, Ph.D. Centre for Agricultural Biochemistry and Biotechnology (CABB), University of Agriculture, Faisalabad, Pakistan

S.M. Paul Khurana, Ph.D. Amity Institute of Biotechnology, Amity University Harayana, Gurgaon, Haryana, India

Sumaira Kousar, Ph.D. Industrial Biotechnology, National Institute for Biotechnology and Genetic Engineering (NIBGE), Faisalabad, Pakistan

R. Kowsalya, Ph.D. Department of Biotechnology, Periyar University, Salem, India

Milena Kulasek, Ph.D. Department of Genetics, Breeding and Biotechnology, Warsaw University of Life Sciences, Warszawa, Poland

Dinesh Kumar, Ph.D. Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi, India

Amit K. Kushwaha, Ph.D. Department of Metabolic and Structural Biology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India

P. León-Mejía, Ph.D. Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

P.K. Malhotra, Ph.D. Schools of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, Punjab, India

K. Manigandan, Ph.D. Department of Biotechnology, Periyar University, Salem, India

Marcos E. Martinez-Montero, Ph.D. Bioplantas Center, Plant Breeding Laboratory, University of Ciego de Ávila, Ciego de Avila, Cuba

Bhartendu Nath Mishra, Ph.D. Department of Biotechnology, Institute of Engineering and Technology, G.B. Technical University, Lucknow, India

Siddhartha Kumar Mishra, Ph.D. School of Biological Sciences, Dr. Harisingh Gour Central University, Sagar, India

C.S. Mukhopadhyay, Ph.D. School of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana, Punjab, India

Raveendran Muthurajan, Ph.D. Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India

Elangovan Namasivayam, Ph.D. Department of Biotechnology, Periyar University, Salem, India

Senthil Natesan, Ph.D. Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India

Khwaja Osama, Ph.D. Department of Biotechnology, Institute of Engineering and Technology, G.B. Technical University, Lucknow, India

Pavan Kumar Padarthi, Ph.D. Department of Biotechnology, Periyar University, Salem, India

Basavanagouda S. Patil, Ph.D. Regional Research Centre, Indian Agriculture Research Institute, Dharwad, Karnataka, India

Hifzur Rahman, Ph.D. Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India

Anil Rai, Ph.D. Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi, India

Sathishraj Rajendran, Ph.D. Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India

Mahendran Ramachandran, Ph.D. Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India

Valarmathi Ramanathan, Ph.D. Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India

Sasikala Ramasamy, Ph.D. Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India

Saleha Resham, Ph.D. Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Ernesto Restaino, Ph.D. National Institute for Agricultural Research, INIA La Estanzuela, Colonia, Uruguay

Navinder Saini, Ph.D. Vivekananda Parvatiya Krishi Anusandhan Sansthan (VPKAS), Almora, India

Neelam S. Sangwan, Ph.D. Department of Metabolic and Structural Biology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India

Rajender Singh Sangwan, Ph.D. Department of Metabolic and Structural Biology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India

Center for Innovative and Applied Bioprocessing (formerly BioProcessing Unit), (An Autonomous Institute under Department of Biotechnology, Govt. of India), Mohali, Punjab, India

Chiranjib Sarkar, Ph.D. Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi, India

Sarfraz Shafiq, Ph.D. Department of Environmental Sciences, COMSATS I.I.T., Abbottabad, Pakistan

Anamika Sharma, Ph.D. Department of Dental Microbiology, Theerthankar Mahaveer University, Moradabad, Uttar Pradesh, India

Pankaj Sharma, Ph.D. Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, Punjab, India

Shubhendu Shekhar, Ph.D. National Institute of Plant Genome Research, New Delhi, India

G.S. Sidhu, Ph.D. Schools of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, Punjab, India

Andrey Sinjushin, Ph.D. Department of Genetics, M.V. Lomonosov Moscow State University, Moscow, Russia

Pallavi Somvanshi, Ph.D. Department of Biotechnology, Teri University, New Delhi, India

Bratislav Stankovic, Ph.D. University for Information Science and Technology "St. Paul the Apostle", Ohrid, Republic of Macedonia

Pamidimarri D.V.N. Sudheer, Ph.D. Department of Chemical and Biochemical Engineering, Dongguk University, Seoul, South Korea

Magdalena Szechyńska-Hebda, Ph.D. Department of Genetics, Breeding and Biotechnology, Warsaw University of Life Sciences, Warszawa, Poland

The Franciszek Górski Institute of Plant Physiology, Polish Academy of Sciences, Kraków, Poland

Dibyendu Talukdar, Ph.D. Department of Botany, R.P.M. College, University of Calcutta, Uttarpara, Hooghly, West Bengal, India

Tulika Talukdar, Ph.D. Department of Botany, Krishnagar Government College, Nadia, Krishnagar, West Bengal, India

Department of Botany, APC Roy Govt. College, Siliguri, Darjeeling, West Bengal, India

Pragya Tiwari, Ph.D. Department of Metabolic and Structural Biology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India

Basant K. Tiwary, Ph.D. Centre for Bioinformatics, Pondicherry University, Pondicherry, India

Swati Tripathi, Ph.D. Vegetable Research Division, National Institute of Horticultural and Herbal Science, Rural Development Administration, Suwon, Republic of Korea

Najam us Sahar Sadaf Zaidi, Ph.D. Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Harsh Vardhan, Ph.D. Faculty of Medicine, Department of Immunology/ Service of pneumology, University of Sherbrooke, Sherbrooke, QC, Canada

Alain Vian, Ph.D. UMR 1345 IRHS (Université d'Angers, Agrocampus Ouest, INRA), SFR 4207 Quasav, Faculté des Sciences, Angers cedex 01, France

Esteban Vicente, Ph.D. National Institute for Agricultural Research, INIA Salto Grande, Salto, Uruguay

Dinesh K. Yadav, Ph.D. Amity Institute of Biotechnology, Amity University Harayana, Gurgaon, Haryana, India

Neelam Yadav, Ph.D. Amity Institute of Biotechnology, Amity University Harayana, Gurgaon, Haryana, India

Ritesh K. Yadav, Ph.D. Department of Metabolic and Structural Biology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India

Sarika Yadav, Ph.D. Amity Institute of Biotechnology, Amity University Harayana, Gurgaon, Haryana, India

About the Editors

Debmalya Barh (M.Sc., M.Tech., M.Phil., Ph.D., PGDM) is a well-known molecular biotechnologist who is an expert in integrative *omics*–based cutting-edge R&D. He is the founder of the Institute of Integrative Omics and Applied Biotechnology (IIOAB) – a first-of-its-kind research organization in India that provides a global platform for multidisciplinary research and advocacy. He works with 400+ well-regarded researchers from 40+ countries and has 150+ high-impact international publications, several books, and book chapters in the field of biomedical and agricultural omics. Dr. Barh is a globally branded editor for editing omics related research reference books from Springer and Taylor & Francis and an editorial and review board member for a number of highly reputed international journals. Due to his significant contributions in the field and in promoting applied biological, agricultural, and bio-medical sciences using unique research strategies, in the year 2010 he has been recognized by *Who's Who in the World* and in 2014 he has been entered into the *Limca Book of Records,* "the Indian equivalent to the Guinness Book of World Records".

Muhammad Sarwar Khan, a Doctorate from the University of Cambridge, is a highly regarded Molecular Biologist from Pakistan who was a Founding Head of Biotech Interdisciplinary Division at NIBGE and is currently serving as the Director of Center of Agricultural Biochemistry and Biotechnology (CABB), University of Agriculture, Faisalabad, Pakistan. He has several awards to his credit, including Civil Award (Presidential Medal for Technology), Gold Medal in Agriculture (Pakistan Academy of Sciences), Performance Gold Medal (Pakistan Atomic Energy Commission), and Biotechnologist Award (National Commission on Biotechnology), and is a Life Fellow of Cambridge Commonwealth Society. He has several high-impact publications in scientific journals including *Nature* and *Nature Biotechnology*, and is an author of a number of books and book chapters. Dr. Khan has made colossal contributions in the field of chloroplast genetic engineering and is pioneer in expressing GFP in plant chloroplasts, developing transplastomic rice and sugarcane. His research focus also includes expression of foreign genes in the chloroplasts to confer agronomic traits such as insect-pest resistance, salinity and herbicide tolerance, and overexpression of antigenic and therapeutic proteins in chloroplasts to develop cost-effective therapeutics and vaccines.

Machine Learning Techniques in Plant Biology

Khwaja Osama, Bhartendu Nath Mishra, and Pallavi Somvanshi

Contents

Abstract

There is an increasing amount of various genome-sequencing projects and advancement in generation of plant ESTs has resulted in generation of large quantities of data from different fields of plant biology in the public domain. Therefore, a need arises in the analysis of the available data and integrating them with several information of plant biology like crop improvement, nutrigenomics, biochemical engineering, etc. The biological data are mostly complex and vague, analysis of these data is difficult, and interpretation of interaction in different elements cannot be done by simple mathematical functions. Complex computing approaches like artificial intelligence are being applied to understand and interpret these data. The definition of intelligence is debatable for a long period of time; however, intelligence can be vaguely defined as the ability to learn from previous experiences and to adapt accordingly in relatively

K. Osama, Ph.D. • B.N. Mishra, Ph.D.
Department of Biotechnology, Institute of Engineering and Technology, G.B. Technical University, Sitapur Road, Lucknow 226021, India

P. Somvanshi, Ph.D. (✉)
Department of Biotechnology, Teri University, 10 Institutional Area, Vasant Kunj, New Delhi 110070, India
e-mail: psomvanshi@gmail.com

D. Barh et al. (eds.), *PlantOmics: The Omics of Plant Science*,
DOI 10.1007/978-81-322-2172-2_26, © Springer India 2015

new situations. Artificial intelligence uses machine learning algorithm in which the system generates some adaptive learning approaches in order to achieve some goal of environment. Several machine learning approaches have been applied in plant biology till date. In this chapter we will discuss few machine learning approaches and their applications in plant biology.

Keywords

Artificial neural network • Support vector machine • Artificial intelligence • Hidden Markov model • Genetic algorithm

Introduction

One of the most distinctive qualities of human beings that distinguishes them from other animals is their desire to understand and control the process of nature. This basic nature of humans gives birth to the field of science. Over the years of learning and understanding, humans have developed a majestic structure of knowledge that helps us to understand and predict to varying extent various natural phenomena. Though, there is a limit in our abilities to predict and hence several complex resources have been developed by us to understand and control many aspects of life. Over the course of time and our repeated interactions with great uncontrollable natural forces, we have learned the extent in which we can control some aspects of life and extent to which many aspects are uncontrollable. The aim of creating artificial intelligence and artificial life can be traced back to the very beginnings of the Computer Age. Although the study of intelligence is a more than 2000-year-old discipline, artificial intelligence is one of the newest disciplines formally started in 1956. Artificial intelligence is a vast field comprising large areas like logical reasoning, computation, and probability. Different scientists have defined artificial intelligence in different ways; however, these definitions can be broadly categorized in four classes (Russell et al. 1995) as:

1. Systems that think rationally
2. Systems that act rationally
3. Systems that think like human beings
4. Systems that act like human beings

The machine learning field is a part of the broad field of artificial intelligence and a direct successor of statistical model fitting with useful information from pile of data. The only staggering difference between statistical approach and machine learning is that the former regards description of data to be handled in mathematical terms of probability measure and not in terms of deterministic function such as cluster assignments, prediction functions, etc. The tasks to be solved are practically equivalent. In this area, learning methods are also known as estimation methods. Many researchers have long time ago found the basic philosophy and idea of machine learning to be very closely related to nonparametric forms of estimation. Estimation, unlike the statistical approach, does not require learning frameworks or other related things to statistical approach as the former does not have to do with probabilistic model of any data. As an alternative, it only assumes interest in prediction of new instances, which is a far less ambitious work and requires lesser examples to help the point in order to achieve a required performance.

Human beings learn through their life experience and human brain can solve very complex problems based on that learning. On the contrary, machines follow a set of rules or algorithm to solve a problem. The difference in the workings of a human brain and machine can be explained by a simple example of a 5-year-old child who can easily differentiate between a cow, goat, and sheep but will not be able to solve a mathematical equation. However, a computer can solve the equation quite easily, while differentiating between animals would be much difficult (Fig. 1). In the past few years, various studies have shown that machine learning and statistical approach are unlike approaches but converge at some point of time. It is often possible to express the methods of machine learning in probabilistic framework, and vice versa performance of these methods in view of theoretical study is immensely based on the similar assumption or postulation and has been followed as probability theory. It is inspired by the biological brain; the word "fitting" is exchanged with "learning." The "learning"

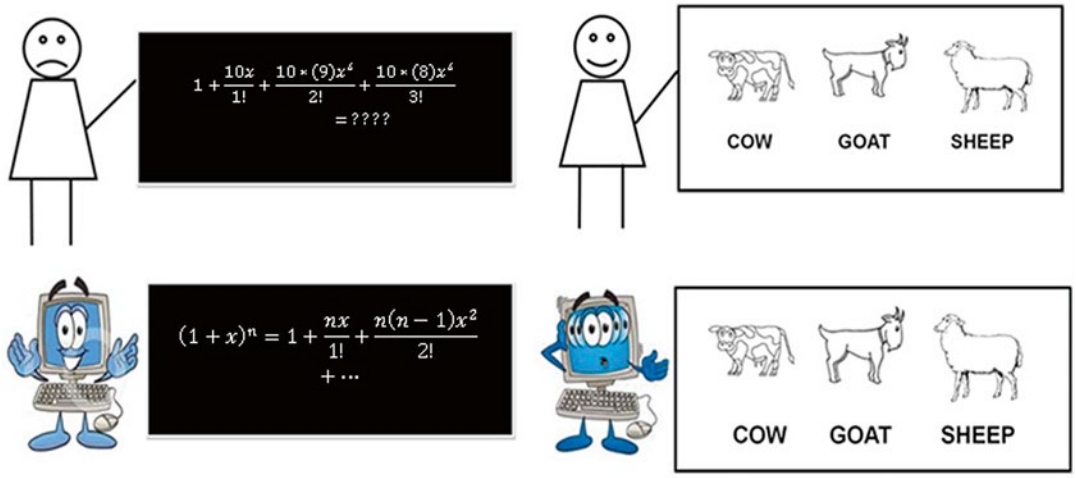

Fig. 1 Illustration of difference in the workings of the human brain and machines

process in machine learning can be classified into three categories.

(a) *Supervised Learning:* In this type of learning, a sample of input–output pair is provided to the machine for learning (training data set). Each input set in the training data set is associated with the output set. The task of the machine is to find a deterministic function that maps each input with its associated target values in order to minimize the error in future prediction. While solving a given problem using supervised learning, some of the steps are to be considered (Fig. 2). First, we need to determine the number of variables involved in defining the problem; second, we need to select a training data set that describes the problem completely. In the third step, the training data is presented to the system in a form understandable to it. The "machine" or the model is then trained with the data. The learning takes place by adjusting weights of connections according to training error calculated by comparing model output and the actual output in training data set. The trained model is then validated for its robustness and accuracy with the validation data set which contains data that were not present in the training data set. Based upon the nature of the target values' type of deterministic function, changes and different types of learning can be performed like classification learning (the aim is to find whether the two elements in output space are the same or not); preference learning (where the aim is to find whether two elements in output space are equal or not and if not, which one is preferred over the other), an example of such learning are search results of queries on web search engines; and function learning (where the aim is to optimize a function for a given process).

(b) *Unsupervised Learning:* Unsupervised learning is a machine learning technique in which the data set used for training the system does not contain target vectors. Instead of which training data set contains input vectors and a cost function which is to be minimized during the process of learning. The aim of machine in unsupervised learning is to develop representations of input data that can be used for solving problems like decision making, predicting future inputs, etc. Unsupervised learning is mostly used in the field of estimation problem like dimensionality reduction, clustering, statistical modeling, etc. One simple example of unsupervised learning is clustering where we try to cluster different types of data based upon the input data. The inputs in training data set are used by machine to learn pattern, and any new data which lies beyond the limit of those patterns is considered noise.

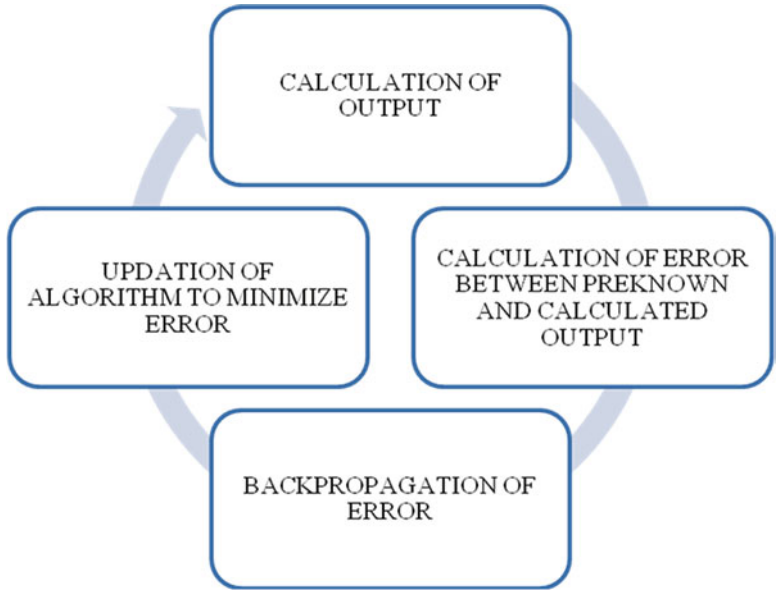

Fig. 2 Steps taken in supervised learning

(c) *Reinforced Learning:* In reinforced learning, the data is not provided to the machine. Instead, the machine interacts with the environment by taking some actions and is rewarded or punished by the reactions from the environment. The machine uses these rewards, corresponding to its actions to learn to act in a way to maximize its future rewards (and minimize punishments). The reinforced learning is concerned about how the machine learns to live with the environment for long term while maximizing its rewards. It defines a function that needs to be maximized during the learning process, and then it finds a strategy to get maximum reward. The reinforced learning uses several algorithms to find this strategy like native brute-force algorithm, value function approaches, direct policy estimation, etc. Reinforced learning is being successfully used in robotics, games, telecommunications, etc.

Machine learning is called black box model as compared to mathematical modeling which is considered to be white box model (Fig. 3). Machine learning is considered black box because in these models are based upon the information or data available from the process but very little theory is known, while in mathematical modeling, models are based upon theoretical knowledge of the process. It will classify or optimize test set according to training set, but it will not provide information of which variables are involved (Mozer and Smolensky 1989; Andrews et al. 1995; Tickle et al. 1998; Alexander and Mozer 1999).

Ideally, machine learning approaches are best suited where abundant amount of data are available, but very less is known about the process. Thus, machine learning is used in fields which are rich in information. These fields are rich in data but the theoretical knowledge is not sufficient for building a model. Thus, these fields solve the problems by principle of induction in inference. However, building a model with the available data and no theoretical knowledge is a difficult task because often these data are incomplete and noisy (Baldi and Brunak 2001). Biological science is essentially an information-rich field and since biological processes are too complex, so very little has been discovered about them; thus, biological science in general or computational biology in particular is an ideal field for application of machine learning.

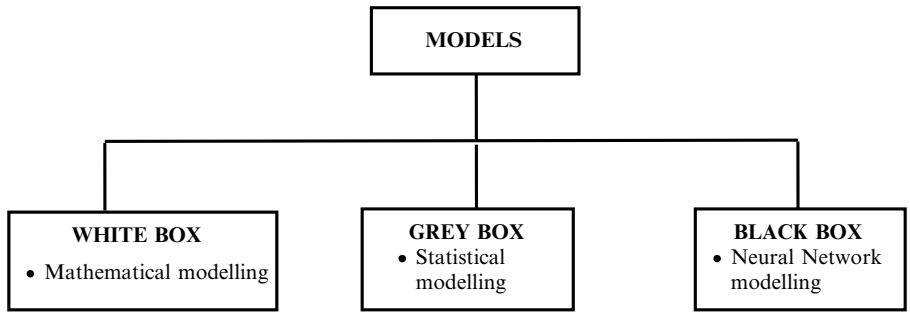

Fig. 3 Classification of models

Problems in Modeling Biological Process

Biological processes vary with time, are nonlinear in nature, and are complex due to composition of many different and interacting elements governed by nondeterministic rules and influenced by external factors (Coruzzi et al. 2009; Gallego et al. 2011). Commonly, most of biological interactions cannot be elucidated by simple stepwise algorithm or a precise formula, particularly when the data set is complex, noisy, vague, uncompleted, or formed by different kinds of data (Prasad and Gupta 2006; Gallego et al. 2011). Many times, behavior of a biological system over a time period is difficult to understand and interpret; additionally, genetic and environmental factors also show biological responses (Karim et al. 1997).

The modeling of these systems is challenging and of extreme importance for scientists and engineers for purposes such as prediction and simulation. Typically, researchers look to create models with two main goals in mind. First, the model should accurately map the input variables to the output variables as observed in real-life situations. Secondly, the model should be a fitting representation of the system's underlying physical characteristics (Resop 2006).

Deterministic mathematical models also known as white box models have been traditionally developed from either physical principles or by statistical regression (Salas et al. 2000). Physical models consist of systems of ordinary or partial differential equations. These models try to represent the underlying physical relationship between variables involved. The benefit of physical models is that they are based on a deep and thorough understanding of the system. However, limitations of these models include the difficulty of setting up and solving complex differential equations analytically, as well as determining equation coefficients and initial and boundary conditions (Coppola et al. 2005). Usually, these equations must be solved using numerical methods, such as the finite element method.

Statistical models on the other hand are designed by finding the equation that best fits a set of experimental data. These models are useful and are generally simple to solve. Statistical regression equations limit the user by requiring a large amount of sample data to estimate parameters of equation and to find the data trend. Also, there are difficulties that arise when manually determining the optimal structure of the statistical equation (Hill et al. 1994).

Artificial intelligence technologies have same or even better potential than traditional statistics in modeling nonlinear relationships in biological data and also have superior prediction powers (Gago et al. 2010a). Recent studies have demonstrated that AI technologies show the same or even better performance than traditional statistics for modeling complex nonlinear relationships hidden in the data and offer superior prediction powers (Landín et al. 2009; Gago et al. 2010a).

Algorithms of Machine Learning

In this section we will briefly discuss some of the highly used algorithms of machine learning.

Dynamic Programming

Dynamic programming is used for problems which can be divided into smaller subproblems and the solution of the bigger problem is found by combining the solutions of smaller problems. Dynamic programming is used almost everywhere in sequence analysis. Sequence alignment analysis algorithms like Needleman–Wunsch and Smith–Waterman are some examples of dynamic programming. Dynamic programming is very well known and the origin of many predictable algorithms for series analysis. Reinforcement or fortification learning algorithms are another very important class of algorithms which can be analyzed as simplification of ideas for dynamic programming.

Gradient Descent

Gradient descent also known as steepest descent is one of the most important breakthroughs in machine learning. It is used for building a best model that minimizes the error. It is simple and easy to use and guarantees to find a minimum of a function if present. Gradient descent is often used with back propagation of information like in back-propagation neural networks. In complex functions where a number of local minima are present, gradient descent mostly ends up finding the local minima rather than global. Therefore, gradient descent algorithm is generally run in multiple iterations with different starting points and learning rates.

Gradient descent is slow and often less effective at small step size, and while convergence speed could be increased with increasing step size, larger step size results in large error. Speed of convergence and efficiency of gradient descent are improved using various variations like conjugate gradient descent (in which the weights are adjusted in directions conjugate to the gradient in order to get fast convergence) and gradient descent with adaptive (in which the learning rate is adjusted during the training in order to produce an optimum convergence rate and error) line search algorithms (Stanimirovic and Miladinovic 2010).

Expectation–Maximization Algorithms

In computational biology, the data available for training probabilistic are often incomplete. Expectation–maximization algorithm is used for parameter estimation in such models. It is a generalization of maximum likelihood estimation in incomplete data case, but expectation–maximization addresses more difficult problems than maximum likelihood.

In the expectation–maximization algorithm, hidden or missing variables are estimated using known or present parameters (the E step), and then these completions are used to reestimate hidden parameters. This step is called the M step because this can be thought as maximization of expected log-likelihood of data. In the complete data case, there is only one global optimum, but in the incomplete data case, there are multiple local optima. Expectation–maximization algorithm cuts the problem into simple subproblems which have single global optima (Chuong and Serafim 2008).

Many models in computational biology have hidden variables. These hidden variables are due to missing or non-recordable or corrupted data. Expectation–maximization algorithms are used in many applications like hidden Markov models, neural networks, etc. and sequence analysis like gene expression clustering (D'haeseleer 2005), motif finding (Lawrence and Reilly 1990), haplotype inference problem (Excoffier and Slatkin 1995), etc.

Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo methods belong to an important class of stochastic methods, which are strongly related to statistical physics and are highly sought methods for machine learning and Bayesian inference. These methods use probability distribution to solve problems. Markov Chain Monte Carlo (MCMC) method aims to find the solution of two basic problems: first, to use a probability distribution to generate random inputs in a defined domain and second, to estimate the expectations of a deterministic function over the distribution. This takes several steps, and after number of steps, the quality of the given sample improves as function for a number of steps to be taken at some point of time. If the distribution of inputs is not uniform, the approximation will not be accurate; also, size of input affects the accuracy of approximation; if a number of input samples are low, approximation will be poor. Typically, it is really not hard to create a Markov Chain with the preferred properties. In addition, the more difficult thing is to determine the steps, which are needed to converge to static distribution within a suitable error. A good form of chain will also have rapid mixing in which the immobile distribution is reached very quickly starting from any arbitrary position. The most common application of algorithm used in Markov Chain Monte Carlo methods is numerically calculating multidimensional integrals.

Simulated Annealing

Optimization of problem is a difficult task often practically impossible. As the problem gets large, area required to search for optimum also becomes large and a huge number of possible solutions were searched to find the optimum one. These are a large number of solutions still for modern computing. Often, while finding the global optima for a given problem, optimization algorithm gets stuck in local optima. Simulated annealing is a random search method for global optimization problem. This method is inspired by the annealing process of metals. Annealing involves heating and cooling of metals to change their physical properties. When the metal is heated, molecules in metal have high energy and they vibrate highly, but when it cools slowly, the vibration of metal molecules also slows down and metals' new structure gets fixed. In simulated annealing, the search of optimum is started at high energy and then it is lowered slowly. At high energy, the algorithm will accept solutions with greater frequency, accepting more solutions worse than the current solution. This provides algorithm the ability to jump local minimums. As the algorithm progresses, energy is slowly lowered, reducing the algorithm's frequency of finding solutions worse than the current one. Thus, the algorithm focuses on a search space to find global minima. Simulated annealing may become more efficient than other algorithms such as exhaustive enumeration, if the aim is merely to find an acceptably good solution in a fixed amount of time, or rather the best of all the possible solutions.

Evolutionary and Genetic Algorithms

Evolutionary algorithms are computer programs, which can solve complex and complicated mathematical and statistical problems using Darwin's theory of evolution. Several fixed-length vectors also known as individuals compete with each other to search for an optimal area. These creatures evolve with time to find the optimal solution. Evolutionary algorithms have been started with an initial population of individuals of finite size. Each individual is then associated with a fitness score. A fitness function is used to calculate the fitness score of each individual. The individuals with high fitness score represent the healthier solutions of the problems than that with individuals having low fitness score. After this initial phase, the main cycle of evolution begins. Each individual in the initial population generates one offspring using mutations. These offspring are then given a fitness value. Now, this first-generation children form a population which is considered as present population, and this cycle is repeated many times. These individuals evolve from generation to generation and compete with

each other in the same generation to be fittest scorer. The individual with the fittest score is considered to be the fittest individual and is selected to solve the problem.

Artificial Neural Networks

Artificial neural networks are complex mathematical models, which mimic biological neural networks. An artificial neural network like any biological neural network is built by connecting several neurons. An artificial neuron (Fig. 4) is a simple mathematical model that does three functions: multiplication, addition, and activation (Krenker et al. 2011). The neuron first multiplies every input given by its corresponding weights. The network to memorize a given process uses weights corresponding to inputs and inter-node connections; these weights determine the conductivity of inputs through the network. The weighted inputs given to an artificial neuron are then added and passed through an activation function, also called a transfer to the next neuron. This function can be sigmoidal, linear, hyperbolic, tangent, or radial basis, and the type of activation can be selected according to specific problem. The most common neuronal nonlinear activation function used in biological systems is sigmoid in nature.

A single neuron is a simple mathematical function which is not useful in solving big problems. When these neurons are connected to each other to form an artificial neural network, the real potential of these models is visible. Unlike biological neural networks, artificial neurons are connected in a defined architecture. This arrangement of artificial neurons is called topology of network. Several standard network topologies have been defined by scientists in the past for different problems. For a particular problem, appropriate topology has to be decided and we need to fine-tune the topology itself and its parameters. The fine-tuning of network topology includes teaching the network in solving a given problem. Artificial neural networks like biological neural network can learn their behavior on the basis of inputs that they get from their environment (Kos et al. 2011).

This teaching or fine-tuning of neural network is called training of artificial neural network. Many different training mechanisms have been used in neural networks. These affect the accuracy of models and also influence speed at which the networks converge. The training of artificial neural networks can be classified as supervised learning and unsupervised learning. In supervised method, the desired output or target values are provided by an external source, and then the network output is matched with target values for optimizing the network weights and correcting network functioning. Artificial neural network uses a delta rule for training (Widrow and Hoff 1960). To train a network with a given set of training samples containing input data set x^p and target data set d^p, network calculates output y^p for every input values and subtract it from target values to calculate error ($d^p - y^p$):

$$y^p = \sum_j w_j x_j + \theta \tag{1}$$

where θ is the bias of the network.

Delta rule uses a cost function based on these errors to modify weights. The final error is then calculated using a cost or error function which can be mean square, root mean square, least mean

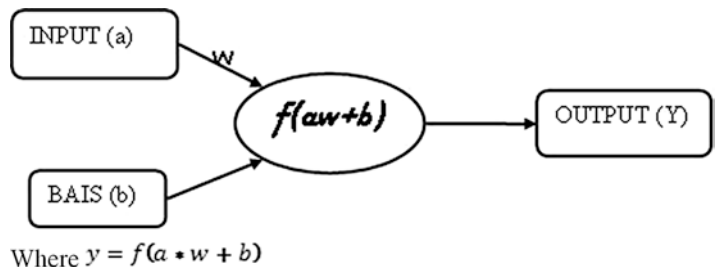

Fig. 4 Working principle of an artificial neuron

square, etc. The total error using least mean square error can be defined by

$$E = \sum_{p} \mathrm{E}^{\mathrm{p}} = \frac{1}{2} \sum_{p} \left(d^{\mathrm{p}} - y^{\mathrm{p}} \right)^2 \tag{2}$$

P represents the range of input data set, and E^{p} represents the error on the whole range of input data set. Weights are then adjusted to reduce error by gradient descent method and delta rule finds the value of new weights. The weights are changed proportionally to the negative of derivative of error measured for the current iteration with respect to each weight:

$$\Delta_{\mathrm{p}} w_j = -g \frac{\partial E^{\mathrm{p}}}{\partial w_j} \tag{3}$$

where γ is a constant of proportionality and $\Delta_{\mathrm{p}} w_j$ represents the change in target for pattern p. The derivative is

$$\frac{\partial E^{\mathrm{p}}}{\partial w_j} = \frac{\partial E^{p}}{\partial y^{\mathrm{p}}} \frac{\partial y^{\mathrm{p}}}{\partial w_j} \tag{4}$$

For a linear activation function,

$$\frac{\partial y^{p}}{\partial w_j} = \mathrm{x}_{\mathrm{j}} \tag{5}$$

and

$$\frac{\partial E^{p}}{\partial y^{p}} = -\left(d^{p} - \mathrm{y}^{\mathrm{p}} \right) \tag{6}$$

Thus,

$$\Delta_{p} W_j = \gamma \partial^{p} \mathrm{x}_{\mathrm{j}} \tag{7}$$

where x_{j} is the input vector and $\partial^{\mathrm{p}} = (d^{\mathrm{p}} - y^{\mathrm{p}})$ is the difference between the network output and the actual output or target for pattern p.

The delta rule modifies weights according to a proportionally negative derivative of error, i.e., if on increasing weight error decreases, then delta rule keeps on increasing weight till error reaches minimum or starts increasing, and if on increasing weight error decreases, delta rule decreases weights till error reaches minimum or starts decreasing.

In supervised learning technique, input variables and a cost function are provided to network, but no output variables are provided. In this technique, network parameters are set on the basis of input data and cost function. In pattern classification using unsupervised learning, a self-organizing network identifies the silent features of input data set; however, unlike supervised learning, In supervised learning there is no defined set of categories into which the patterns can be classified (Prasad and Gupta 2006).

A trained network is then validated for accuracy and robustness by simulating for a validation data set. Validation data set contains some input variables present in training data set and some new input variables. If network accurately simulates validation data set, then it is considered a trained or learned network.

Sometimes after learning if the network performance for training data is best but for test data set its performance is poor, it is called overlearning or over-fitting of network. Network size plays an important role in overlearning of a network; a large network over-fits small problem. So, an optimum size of network (number of nodes) has to be decided for a given problem. After training of network, it can be used for solving problems. Artificial neural networks are used for problems like function approximation, regression analysis, time series prediction, classification, pattern recognition, decision making, data processing, filtering, clustering, etc.

Structure of Artificial Neural Network

Artificial neural networks are inspired from biological nervous system and consist of a network of artificial neurons. An artificial neuron is a simple mathematical model which is not capable of solving complex real-life problems. The ability of artificial neural network in solving complex problems is due to arrangement of these neurons in the form of a network. The information in a neural network is processed through its building blocks in a nonlinear and parallel manner.

Arrangement of neurons in a network is called its topology or architecture of artificial neural network. An artificial neuron is called a node in a network; these nodes are arranged in the form of layers. Typically, the most commonly used artificial neural network is a three-layered network. The first layer is an input layer, the second layer is a hidden layer, and the third layer is an output layer. Generally, the three-layered neural network is shown as typical example because it is capable of solving practically all types of problems. Three-layered networks are capable of solving all problems but larger networks can solve these problems more efficiently.

The connections between nodes and number of nodes per layer are defined by the approach, which is adopted to solve or interpret a given problem. Internodal connections direct direction of information flow through a neural network. On the basis of direction, the flow of information neural networks can be classified as: feed-forward neural network and cascade forward for unidirectional flow of information and recurrent or feedback for bidirectional flow of information. In feed-forward neural network, information flows from one layer to another in one direction, i.e., from input to hidden to output layer (Fig. 5). In cascade forward neural network, the flow of information is unidirectional except that there is an extra weight connection from input to each layer (Fig. 6). However, in networks having bidirectional flow of information (recurrent neural network), information flows both in forward and backward directions (Fig. 7). In a complete recurrent neural network, each node is connected to every node including itself. Recurrent neural networks are bulky and complex compared to feed-forward neural networks due to massive parallel processing of information. These networks require large computational space and are not easy to understand, unstable, and noise sensitive (Mandic and Chambers 2001). However, recurrent neural networks are closer to biological neural networks. Due to recurrent connections in these networks, output of a network can be used as input of the same network. This property provides the ability of prediction of future outcomes to recurrent neural networks. Several other networks are being designed using fuzzy logic and other techniques; some of these networks are self-organizing neural network, regression neural networks, fuzzy neural networks, etc. (Hayashi et al. 1993; Yao 1999; Yang 2006).

In 1980s back-propagation algorithm developed and helped in increasing popularity of artificial neural networks as function optimizers. Back-propagation algorithm is used to train networks using experimental data. Learning and updating of weights became easy with back-propagation algorithm. In this algorithm, network errors are back propagated to the network in order to train it. Back propagation can also be considered as a generalization of delta rule for nonlinear activation functions and multilayer networks (Kruschke and Movellan 1991).

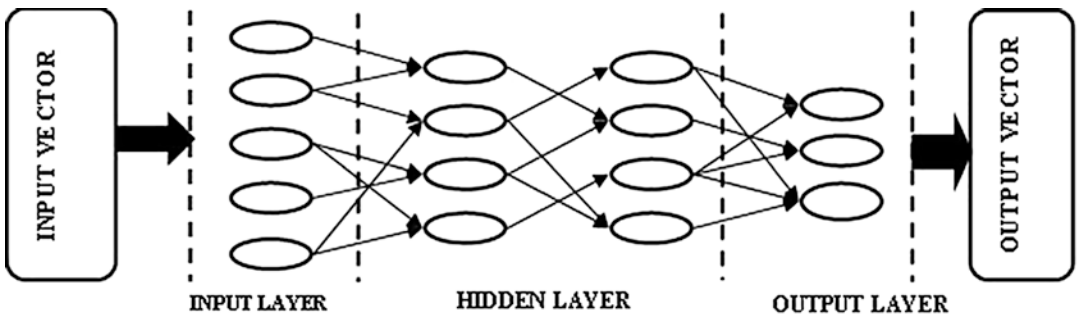

Fig. 5 Architecture of feed-forward neural network

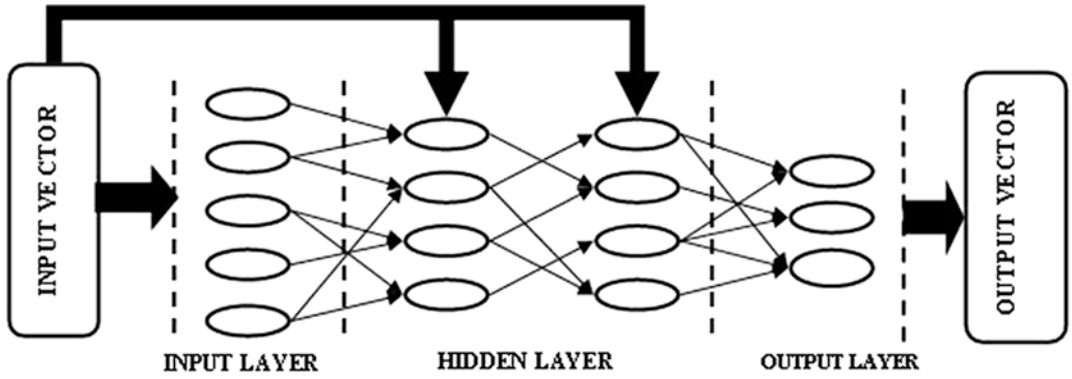

Fig. 6 Architecture of cascade forward neural network

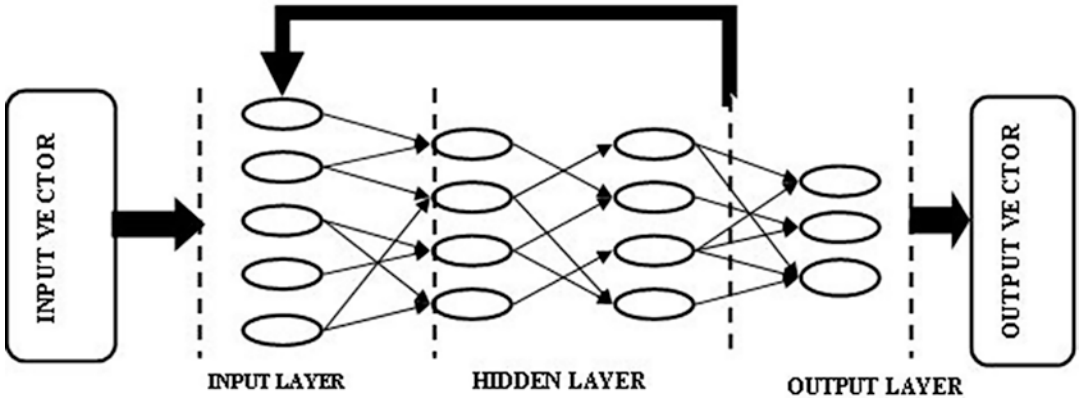

Fig. 7 Architecture of recurrent neural network

Key Steps in Applying Artificial Neural Networks

Data Preprocessing

Transformation and normalization are two widely used preprocessing methods. Transformation involves manipulating raw input data to reduce its dimensionality, while normalization is a transformation performed on input data to distribute the data evenly and scale it (mostly in range of −1 to 1 or 0 to 1) into an acceptable range for the network.

Network Selection

It involves selection of network model, number and size of hidden layers, initial weight matrix, etc.

Training Selection

It needs to start with network topology and initial weight, and train the network on your training data set. When the network reaches the satisfactory minimum error, it saves the weights.

Testing and Interpretation of Results

The trained network is applied on test data set to find the error. If it is not satisfactory, the training process and/or network architecture needs to be modified.

Applications of Artificial Neural Networks in Plant Biology

Artificial neural networks have been researched and used for applications in many different fields. Many of these areas are using artificial neural networks to solve problems, previously thought to be impossible or very difficult with traditional methods. Some of the reasons for the great interest shown in neural networks are their property of being a universal function approximator, i.e., their applicability on different types of problems (White 1992), and their flexibility, robustness, and no need of prior knowledge. A neural network can be applied to practically any problem unlike statistical methods where a mathematical relation between input and output variables is necessary and responses have to be determined for each problem (Zealand et al. 1999). Artificial neural networks prove to be a better choice than regression models for noisy data (Denton 1995). Thus, neural networks are powerful tools for modeling complex processes. They can model highly nonlinear complex systems like biological processes and other real-life problems (Gevrey et al. 2003) (Table 1).

Less applications of artificial neural network to plant biology are present in literature as compared to other fields, viz., pharmaceutical science (Colbourn 2003; Shao et al. 2006; York et al. 2009), ecology (Hilbert and Ostendorf 2001; Adriaenssens et al. 2004), agriculture (Huang 2009), etc. Initial applications of artificial neural networks in plant science include optimization studies. Earlier studies used neural networks with image analysis to identify live or dead plant cells (Fukuda et al. 1991), analyze developmental stages of somatic embryos (Uozumi et al. 1993), etc. Some researchers (Honda et al. 1997) used hybrid fuzzy neural model to predict the length of shoots regenerated from rice callus to be transferred from the growth medium to sugar-free medium for acclimatization. Radius, length, width, roundness, area, and perimeter of the digital images of somatic embryo taken from CCD camera were used as input data for the network. The results of neural network analysis were compared to that of multiple regression analysis. Neural network predicted shoot lengths with 95 % accuracy with an average error of 1.3 mm. In another work, fuzzy neural network was used for modeling the production of Ginjo sake by fermentation in bioreactor (Hanai et al. 1997).

The use of neural network technology in plant and agriculture biology has increased in the last 20 years (Huang 2009). Neural networks have been used for predicting crop yield and modeling pest control treatments according to environmental conditions. Optimization of pesticide concentration and periods of treatment improves the economy of production and minimizes toxic residual levels of agricultural products (Jiménez et al. 2008). Other authors developed neural network models to predict crop yield for crops like corn (Kaul et al. 2005), sugar beet (Kehagias et al. 1998), soybean (Kaul et al. 2005) and winter wheat for different cultivation conditions and distribution of crops in different climate distributions. These works provide important information about the effect of climate change on vegetation of different areas and thus help in conservation of these areas (Hilbert and Ostendorf 2001).

Neural networks have been used in combination with a bioelectric recognition assay to detect plant viruses (Frossyniotis et al. 2008). The authors used biosensors to detect electric signals from plant cells suspended in a gel matrix. The responses of plant cells on interaction with viruses were recorded and used for training of a neural network, and a classification model of cultured cell was prepared. In another study, a multilayer neural network was used with genetic algorithm for the detection of plant virus (Glezakos et al. 2010).

Neural networks have been used in tissue culture on determining the effect of different parameters like carbon source, pH, etc. on plant growth (Prasad and Gupta 2006). A growth model was

Table 1 Some applications of artificial neural network in plant biology

S. No.	Application	Network architecture	Plant species	Database	References
1.	Growth modeling of alfalfa shoots	Feed-forward neural network with Kalman filter	*Medicago sativa*	Dry weight, leaf number, and root initiation stage	Tani et al. (1992)
2.	Classification of embryo types from non-embryos and predicting embryo-derived plantlet formation	Feed-forward	*Apium graveolens*	Area, length to width ratio, circularity, and distance dispersion of plant cell cultures	Uozumi et al. (1993)
3.	Biomass estimation of cell cultures	Feed-forward neural network with gradient descent training method	*Daucus carota*	Sucrose, glucose, and fructose level of medium	Albiol et al. (1995)
4.	Simulation of temperature distribution in culture vessel	Fuzzy neural network with back-propagation algorithm		Spatial temperature distribution of culture vessel	Suroso et al. (1996)
5.	Identification of live and dead plant cells	Three-layered neural network		Size and color of cells	Fukuda et al. (1991)
6.	Clustering of regenerated plantlets into groups	Adaptive resonance theory		Mean brightness values, maximum pixel count, and gray level of maximum pixel count in RBG regions	Zhang et al. (1999)
7.	Predicting shoot length of regenerated callus	Three-layered fuzzy neural network with Kalman filter	*Oryza sativa*	Radius, length, width, roundness, area, and perimeter of the somatic embryo images	Honda et al. (1997)
8.	Detection of plant viruses	Feed-forward back-propagation network with BFGS quasi-Newton optimization algorithm		Electric response from plant cells	Frossyniotis et al. (2008)
9.	Model in vitro rhizogenesis and subsequent acclimatization	Feed-forward neural network	*Vitis vinifera*		Gago et al. (2010c)
10.	Classification of sweet potato embryos	Back-propagation neural network	*Ipomoea batatas*	Embryo area, length, and symmetry; polar coordinates of an embryo's perimeter with respect to its centroid	Molto and Harrell (1993)
11.	Classification of somatic embryos into normal and abnormal	Feed-forward neural network	*Daucus carota*	Morphological features	Ruan et al. (1997)

(continued)

Table 1 (continued)

S. No.	Application	Network architecture	Plant species	Database	References
12.	Prediction of culture condition for optimal productivity	Feed-forward back-propagation neural network	*Glycyrrhiza glabra*	Inoculum size, fresh wt, density, culture temperature, pH, and time of inoculation	Mehrotra et al. (2008)
13.	Growth modeling of hairy roots in nutrient mist reactor	Feed-forward and cascade forward and recurrent neural networks	*Artemisia annua*	Mist On/Off cycle time, initial packing density, media volume, initial sucrose concentration in media, and culture time	Osama et al. (2013)
14.	Sorting of regenerated plants on the basis of their photometrical behavior	Adaptive resonance theory	*Gladiolus*	Photometrical behavior of their leaves in red, blue, and green color regimes	Mahendra et al. (2004)
15.	Prediction of culture conditions for maximum biomass growth	Generalized regression network with radial basis function	*Glycyrrhiza glabra*	Inoculum density, medium pH, sucrose conc., media volume	Prakash et al. (2010)

developed for the study in effect for CO_2 and sucrose content on in vitro shoots of alfalfa (Tani et al. 1992). Gago et al. (2010c) developed a neural network model for in vitro rhizogenesis and acclimatization of micropropogated *Vitis vinifera* L. plants. Effect of cultivars, IBA concentration, and exposure time of plant to IBA on acclimated plant were studied. It was found that they have significant effect on root numbers, number of nodes, and height of the acclimatized plantlets, with exposure time being a more prominent factor. Study was done in different cultivars and the model did good predictions for all cultivars.

Albiol et al. (1995) compared a deterministic mathematical model with a neural network model and found that neural network modeling was cost effective and time efficient and required smaller data set. Neural networks have been used in classification and pattern recognition in plant tissue culture (Prasad and Gupta 2006). Zhang et al. (1999) used neural networks with image analysis for selection of embryos in embryonic tissue culture of Douglas fir. In another study, machine vision was used for distinguishing between white and bright-yellow callus formed in sugarcane callus culture. The technique was successfully used to obtain regeneration frequency of callus culture (Honda et al. 1999). ANNs have been used for sorting regenerated plantlets according to trichometric features of leaves. The sorting of plantlets was done using adaptive resonance theory (ART) having unsupervised architecture (Mahendra et al. 2004). This approach provided a means of selecting plants suitable for *ex vitro* transfer and helps in quality control of commercial micropropagation.

Hairy roots are a good source of plant secondary metabolites. These roots show genetic and biosynthetic stability and require no exogenous growth hormone. However, culture of these hairy roots for high secondary metabolite associated with better biomass production requires optimization of several physical and chemical parameters that affect the growth and productivity of these roots. A feed-forward back-propagation neural network model was developed (Mehrotra et al. 2008) for prediction of in vitro culture conditions for hairy root growth. The model used inoculum size, fresh weight, density, culture temperature, pH, and time of inoculation as input parameters and final fresh weight of roots as final parameters. The trained neural network model was able to predict the final biomass for a partic-

ular culture condition efficiently. Later, Prakash et al. (2010) developed a regression and feed-forward neural network model for prediction of optimal culture conditions for prediction of hairy root maximum biomass yield. It was found that both networks predicted culture conditions efficiently; however, regression neural network was more accurate.

Scale-up of hairy roots in bioreactors is a very difficult task. During their growth hairy roots form clumps which causes heat and mass transfer restrictions. In order to solve these problems, agitators and aerators have to be used in bioreactors, but these equipment cause shear stress on roots, which results in injury and callus formation. Neural networks have also been used for modeling of bioreactors for hairy root growth. Neural network model (Osama et al. 2013) was developed for modeling of hairy root growth in a nutrient mist reactor. The significant features for culture parameters, viz., inoculum size, mist ON time, mist OFF time, initial packing density, media volume, initial sucrose concentration in media, and time of culture, were considered as input of the network. The final biomass of hairy roots on dry weight basis was taken as network output. Three network architectures, viz., feed-forward, cascade forward, and recurrent, were tested and all these networks were found efficient with recurrent neural network being the most accurate.

The knowledge derived through ANNs can be easily increased by training the model by adding to database new inputs (salt concentration, type of medium, other plant hormone, etc.) and/or outputs (plantlet weight, chlorophyll and carotene content, stomata analysis, etc.).

Support Vector Machine

A support vector machine is an abstract machine that uses supervised learning to solve classification problems. These are relatively new, general formulation for learning machines. These techniques learn to assign labels to objects. To control the generalization ability of a learning machine, one has to control two different factors: error rate on the training data and capacity of learning machine as measured by its Vapnik–Chervonenkis dimension, which is a non-negative integer that measures the expressive power for the family of classification functions realized by the learning machine (Haykin 2003). A special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin. Hence, they are also known as maximum margin classifiers. A classification task usually involves with training and testing data, which consists of some data instances. Each instance in the training set contains one target value called class labels and several "attributes" called features. The goal of SVM is to produce the model that predicts the target value of data instances in the testing set, which are given only the attributes.

The training data provided for training contains n ($i = 1, 2, 3 \ldots n$) input vectors denoted by x_i, with each of vectors paired with corresponding labels and are denoted by y_i. The labels in the training data lie in two classes, and for classification of these data in different classes, support vector machine uses an oriented hyperplane. This hyperplane separates the two classes of data points on either side of it. The data points on one side of the hyperplane are labeled positive and on the other side negative. The directed hyperplane is defined by the maximally distant hyperplane from the data points on both of its sides. Thus, the points closest to the hyperplane on each side are the most influential for defining its position, and therefore, these points are called support vectors. The perpendicular distance from support vectors and hyperplane is called margin.

A nonlinear transformation function $\Phi(.)$ is defined to map the input space to a higher dimensional feature space. The oriented separating hyperplane is given as $w.\ \Phi(x) + b = 0$ where $w.\ \Phi(x)$ is the product of data points and weights that determine their orientation and b is the bias or offset of hyperplane from the origin. If for a point x_i, $w.\ \Phi(x_i) + b = 1$, it lies on one side of the hyperplane, and if $w.\ \Phi(x_i) + b = -1$, the point lies on the other side. The support vector machine uses the optimization theory to minimize an objective function which is half the distance

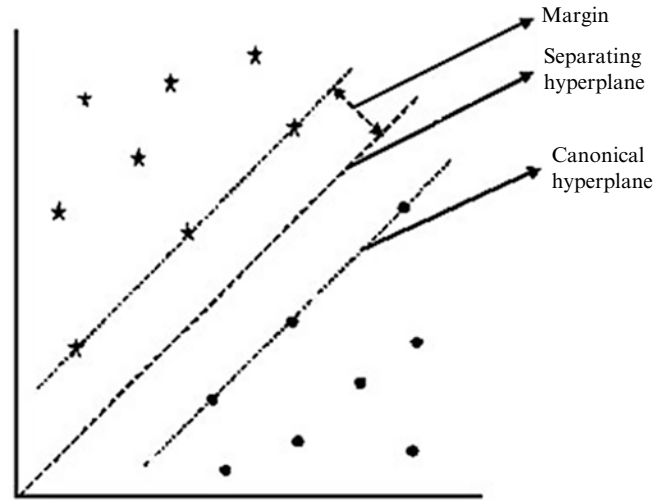

Fig. 8 Classification using support vector machine

between two canonical hyperplanes (hyperplanes passing through support vectors) (Fig. 8).

As mentioned above, most of the traditional neural network models seek to minimize the training error by implementing the empirical risk minimization principle, whereas the support vector machines implement the structural risk minimization principle which attempts to minimize an upper bound on the generalization error by striking a right balance between training error and capacity of machine. Support vector machine also provides guaranteed global optimal solution (Haykin 2003).

Applications of Support Vector Machine in Plant Biology

Biological applications of support vector machines involve classifying objects such as protein and DNA sequences and microarray expression profiles. Performance of support vector machine is mostly similar to or better than traditional machine learning approaches (Hua and Sun 2001). At present there is no algorithm for finding the optimum network architecture, i.e., ideal number of hidden layers, best activation function, etc. This is usually done by trial and error method and is time consuming and often less effective (Shigidi and Garcia 2003). Generally, neural network structures are developed according to past experiences; this requires considerable skills, and efficiency of these networks depends largely on their training (Gonzalez 2000). Another common approach is simply selecting an arbitrarily large number of neurons as models (Xiang et al. 2005). However, a large number of nodes may lead to poor generalization and large computational requirement (Archer and Wang 1993). Support vector machines can thus be an alternative for more accurate classifications.

Protein interactions play a very significant role in any of the biological operation. Prediction of these interactions is a point of key focus for researchers. However, very little has been achieved in this area due to expensive and time-consuming experimental approaches. A support vector machine-based model was developed (Lin et al. 2009) to predict potential *Arabidopsis* (*A. thaliana*) protein interactions based on a variety of indirect evidences. The potential interactions were predicted based on 14 features derived from four types of indirect evidence (coexpression, domain interaction, colocalization and shared annotations). The confidence of predicted interaction was estimated to be 46.87 % and these interactions were expected to cover 29.02 % of the entire interactome. The model successfully recognized 28.91 % of new interactions, similar to its expected sensitivity (29.02 %).

Knowledge of locations in protein expression is important for better understanding of defined cellular processes at organellar and cellular levels. A complete map of a plant proteome is clearly a major goal for plant research community in terms of determining the function and regulation of each encoded protein. An integrative support vector machine-based localization predictor called AtSubP was developed (Kaundal et al. 2010) which was based on the combinatorial presence of diverse protein features, viz., amino acid composition, sequence-order effects, terminal information, position-specific scoring matrix, and similarity search-based, position-specific, iterated Basic Local Alignment Search Tool information. The model predicted seven subcellular compartments through fivefold cross-validation test and achieved an overall sensitivity of 91 % with high-confidence precision and Matthew's correlation coefficient values of 90.9 % and 0.89, respectively.

Eichner et al. (2011) used a support vector-based model for identification of alternate splicing in *A. thaliana* by detecting intron retention and exon skip from tiling arrays. The model used existing EST and cDNA sequences for supervised training. The method developed in this work expands the scarce repertoire of analysis tools for identification of alternative mRNA splicing from whole-genome tiling arrays.

Other authors used image analysis with support vector machine for detection and tracking of plant cell division by in vivo imaging. Cell division in plants takes place mostly in meristems which contain stem cells that give rise to all cell types by regular cell division. However, the control mechanism of cell division is not understood properly and is the center of interest of developmental biologists doing in vivo research in plant cell division (Marcuzzo et al. 2008b). However, for automated machine, vision images have to be partitioned into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Some authors developed support vector machine-based model for classification of plant root cells of *A. thaliana* (Marcuzzo et al. 2008a, 2009). The images were segmented using watershed algorithm and result was improved by merging adjacent regions. The selection of individual cells was obtained using a support vector machine (SVM) classifier, based on a cell descriptor constructed from the shape and edge strength of the cells' contour.

Recently, support vector machines have been used in agricultural research. Support vector machine-based models have been used for classification of crop types, seed pollen grains, etc. Crop type classification is an important application of remote sensing technology, and since the advent of remote sensing technologies, several studies on crop type classification have been published. A support vector machine-based model was developed (Karimi et al. 2006) for detection of weed and nitrogen stress in corn. The results of support vector machine-based classification were compared to that of neural network, and SVM-based classifier was found more accurate. These models use imaging spectroscopy for studying hyperspectral images. Extraction of end-members from these remote-sensed images is a difficult task. A number of algorithms based on notion of spectral mixture modeling have been proposed to accomplish the complex task of finding appropriate end-members for spectral unmixing in hyperspectral data (Martinez et al. 2006). A support vector machine-based end-member extraction (SVM-BEE) model was developed (Filippi and Archibald 2009) for hyperspectral agricultural mapping. This model accurately and rapidly yields a computed representation of hyperspectral data that can accommodate multiple distributions. The efficacies of SVM-BEE, N-FINDR and SMACC algorithms in extracting end-members from real, predominantly agricultural scene were compared. SVM-BEE estimated vegetation and other end-members for all classes in the image, which N-FINDR and SMACC failed to do. Shi et al. (2012) used SVM for classifying agricultural data from public agricultural data set and concluded that SVM outperformed other popular algorithms, like naive Bayes and artificial neural network, in terms of the F_1 measure. Different machine learning algorithms and traditional maximum likelihood algorithm were com-

pared for classification of crops (Nitze et al. 2012). Researchers compared support vector machine, artificial neural network, random forest and maximum likelihood algorithms for classification of ten different crop types. Support vector machine was found to exhibit better results than the other algorithms.

Hidden Markov Model

Andrey Markov gave a concept of a mathematical system that undergoes change from one state to another, between finite numbers of possible states. In this process, next state depends upon the current state and not on sequence of events that preceded it. This phenomenon is called Markov property. Markov model is a probabilistic graphical model considering Markov property. The simplest Markov model is a Markov chain which is like any random process with Markov property. In Markov chain model, the system is assumed to be autonomous and the transition states are fully observable. However, in a hidden Markov model (HMM), the states of transition are not fully observable but output depending upon those states is observable. These models are named because of their two properties. First, it assumes that states Y_t are not observable or hidden from the observer, and second, it assumes that the states of transition and output follow the Markov property, i.e., state Y_t at time t does not depend upon the previous states and also the output A_t does not depend upon the state Y_t (Fig. 9).

These models are used to analyze different types of time series problems in different areas like speech recognition (Juang and Rabiner 1991), ion channel recording (Venkataramanan and Sigworth 2002), optical character recognition (Agazzi and Kuo 1993), computational biology (Krogh et al. 1994), etc.

Hidden Markov model is a tool for estimating probability distributions of a sequence of observations over a time series, finite time series t. The observations can be represented as discrete alphabets, integers, real-valued numbers, etc. In biological applications, the HMM observations are generally discrete alphabets (the 20-letter amino acid, 4-letter nucleotide, 64-letter codon triplet, etc.). A HMM used for solving real-life problems contains many transition states and hidden variables connected to each other. Generally, in biological problems, a unidirectional HMM is used which is also called left–right model. In this type of model, the direction of transition of state is from left to right, and it prevents any transition to a state if transition from that state to another state has taken place (machine learning bioinformatics). There are three basic questions one can ask immediately for a HMM: the likelihood question (how likely is the output for HMM), decoding question (probable sequence of transition states followed by the model to reach the output), and the learning question (how to revise values of transitions and emissions from the given information given that they are not known with complete certainty).

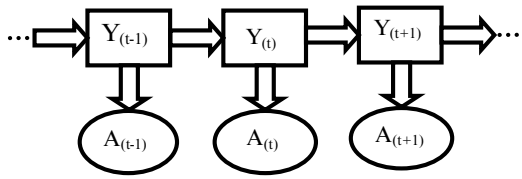

Fig. 9 Simple hidden Markov model

Applications of Hidden Markov Model in Plant Biology

HMM is widely used in speech recognition, natural language modeling, and on-line handwriting recognition. HMM is widely being used in a variety of biological problems like gene finding, secondary structure prediction, gene annotation, etc.

With development in automated whole-genome sequencing, complete genome sequences are becoming more and more abundant. The first and most important task after getting a new genome is to find a protein coding sequence. One of the most successful gene finders was GeneMark (Borodovsky and McIninch 1993; Borodovsky et al. 1994, 1995), which in its first version was based on frame-dependent nonhomogeneous

Markov models. The accuracy of gene finder depends on various factors, the most important being training. A HMM-based gene finder named SNAP was developed and is easily adaptable to a number of organisms (Korf 2004). This gene finder was used to evaluate genomes of *A. thaliana* and *Oryza sativa.* Earlier, three HMM-based gene finders—Exonomy, Unveil and GlimmerM—were presented (Majoros et al. 2003). These gene finders were trained with *Oryza sativa* and *A. thaliana* along with other organisms. Later, a HMM profile was developed (Feng and Xue 2006) to search the proteome of *O. sativa* L. ssp. japonica for serine carboxypeptidase (SCP) and serine carboxypeptidase-like (SCPL) protein. A total of 71 SCPs and SCPLs were found in rice.

Alternative splicing contributes to genome complexity and proteome diversity. Thus, study of alternative splicing sites in a genome is a topic of great interest. Experimental research in this field is both costly and time consuming. A hidden Markov model was used for genome-wide detection and analysis of alternative splicing for nucleotide binding site/leucine-rich repeat sequences in rice (Gu and Guo 2007). HMM-based searches were performed for nucleotide binding site and leucine-rich repeat (NBS-LRR) domain. Out of 875 NBS-LRR sequences obtained from The Institute for Genomic Research (TIGR), 119 (13.6 %) sequences had alternative splicing. Conversely, 71 intron retention events, 20 exon skipping events, 16 alternative termination events, 25 alternative initiation events, 12 alternative 5′ splicing events and 16 alternative 3′ splicing events were identified.

In another study, HMM was used in a neural network-based combinatorial model for prediction of optimal culture condition for maximum biomass yields in *Rauwolfia serpentina* hairy root cultures. Neural network approaches can be evaluated through spatial variations; there is no proper resolution for temporal variations. Nonlinear biological responses are affected by both spatial as well as temporal differences. Therefore, a stochastic approach where time-based differences are taken as random variables to evaluate the whole bioprocess should be considered (Mehrotra et al. 2013). In this study, five HMMs were derived for five test culture conditions and connected to the input layer of three-layered feed-forward neural network. The results of combinatorial ANN-HMM model simulation was compared with ANN model and it was observed that only 2.99 % deviation from experimental result was recorded from combinatorial model against 44 % recorded from ANN model.

Genetic Algorithms

Genetic algorithms were invented in the late 1960s by John Holland and colleagues and students at the University of Michigan in between 1960s and 1970s. Furthermore, John Holland was the first researcher who not only challenged but succeeded to put computational evolution and development on a firm hypothetical footing. Until this, theoretical foundation, based on the ideas of schemas, was the basis of almost all succeeding theoretical works on genetic algorithms. In the past few years, there have been extensive interactions among various researchers studying different computational methods, and existing boundaries between genetic algorithms, evolutionary programming, evolution strategies,, and related approaches have broken down to some extent. At present, the interpretations of genetic algorithm have changed to a very far meaning from John Holland's original conception. Genetic algorithms are an adaptive search heuristic algorithm based on the evolutionary ideas of genetic and natural selection. The basic idea of genetic algorithms is designed to simulate various processes in natural system, which is essential for evolution, specifically those processes which follow the doctrines laid down by Charles Darwin of fittest survival. They represent a sharp exploitation of random search space to solve any particular problem. This search heuristic is consistently used to generate useful solutions to search problems and optimization. GAs are the higher class of evolutionary algorithms, which is used to generate the solutions to optimization problems using the techniques evolved by natural evolution such as mutation, inheritance, crossover, and

selection. Genetic algorithms not only provide alternate method for problem solving but also outperforms consistently other traditional methods. Many problems of real world involve optimal parameters which might become cumbersome for traditional methods but ideally suited to genetic algorithms. Because of its outstanding performance in optimization, genetic algorithms have been wrongly regarded as a function optimizer. There are many ways of viewing the genetic algorithms and perhaps the idea that most users come to make use of genetic algorithms for a problem solver is a restrictive view.

In contrast with evolutionary programming and evolution strategies, John Holland's original aim was to establish and design algorithms for solving specific problems, but to study the phenomenon of adaptation occurring in nature and to develop some ways to import the mechanism of natural adaptation into computer systems to run simulation and study the effect. John Holland's 1975 book *Adaptation in Natural and Artificial Systems* presented the genetic algorithm as a generalization of natural evolution and also gave theoretical structure for adaptation under the genetic algorithm. Genetic algorithm is a popular method for moving one inhabitant of chromosomes or individual of population (which is a computer program or algorithm capable of providing a potential solution of the problem) to a new inhabitant using a kind of natural selection with the influx of genetics and stimulated operators of crossover, inversion, and mutation. The individuals or the computer programs used in genetic algorithms are basically a set of rules arranged in the form of trees. These tree-like structures also called parse tree can be mutated and recombined with new variants. Thus, the machine evolves to find a solution of a complex problem using simple equations. These equations give different outputs from different inputs characteristic for different classes (Kell et al. 2001). The individuals or rules (algorithm/programs) from the population which are chosen by selection operator are allowed to reproduce, and a fitter individual is more likely to produce further offspring than the less fit one. For any defined problem to be solved, the simplest of genetic algorithm works as follows:

(a) Start with a randomly generated population of computer programs/algorithms.
(b) Evaluate the fitness f(x) of each individual in the population.
(c) Select individuals from the present population, the probability of selection being an increasing function of fitness. Selection of individuals is done with replacement, which means any particular chromosome can be selected more than once for it to become a parent one.
(d) Modify the individuals by mutation, recombination, or crossover.
    1. With the crossover rate or its probability, cross over the pair at an assorted selected pair to form two offspring. If no crossover takes place, form two offspring that are exact copies of respective parents. The thing, which is here to be noted, is that crossover rate is simply the probability of two-parent crossing over in a single point. There is some multi-point crossover of the genetic algorithm in which the crossover rate is the number of valid points for crossover to take place.
    2. Mutate the resulted offspring at locus with probability P_m, mutation rate or probability, and set the individual in new population.
(e) After this, replace the present population with new population.
(f) Go back to step b.

Each resultant of this process is called "generation"; genetic algorithm (GA) is typically iterated for anything in between 50 and 500 or could be even more generations. The entire set of required generation is called as "run." At the end of run, there is likelihood to get one or more fit chromosomes from the population. Randomness plays a vital role in run; that's why different runs with different number seeds are more likely to produce different behaviors of generations. Researchers on genetic algorithm often scale down observation based on various simulations averaged over many different runs on the same problem.

Applications of Genetic Algorithms in Plant Biology

Major application of genetic algorithms in fermentation technology and plant science has been for optimization of bioprocess. In a study a mobile robot was developed for harvesting fruit automatically (Noguchi and Terao 1997). Genetic algorithms were used to optimize methods for finding small paths in a given space. Neural networks have been for long used for optimization of bioprocess in fermentation technology. In plant biology, hybrid models of neural networks and genetic algorithms have been used for optimization of conditions for storage of fruits (Morimoto et al. 1997; Morimoto and Hashimoto 2000) and for detection of plant virus using biosensors to observe virus reactions (Glezakos et al. 2010). More complex in vitro culture processes such as shoot proliferation, root formation (rhizogenesis), and plantlet acclimatization have been modeled by ANNs and successfully optimized by genetic algorithms in woody fruit plants, such as kiwifruit (Gago et al. 2010a) and grapevine (Gago et al. 2010b).

Future Prospects

Biological systems are complex and nondeterministic and depend upon genetic and environmental factors. Due to the advent of modern technologies, we can generate a large amount of biological data. These large data have to be analyzed and interpreted to understand important relations between different factors. Machine learning techniques provide a very good scope to analyze these large biological data, interpret the obtained information, and give deep insight into the biological processes. These technologies can be used to develop models that can explain the relationship between different factors and biological responses, which can further be used to predict future responses in specific situations.

Machine learning techniques are good for determining nonlinear relationships followed in biological systems; thus, these techniques make better models than statistical techniques. Techniques like neural network require less prior data and are more accurate compared to statistical techniques (Gago et al. 2010b). Combinations of these techniques could be used for developing more accurate models which can predict outcome of tissue culture experiments, optimize and control bioprocess operations at large-scale, predict crop yields according to climate changes, etc.. These techniques are easy to understand and a plant biologist can very easily use these techniques by having a very good understanding of mathematical and statistical modeling.

References

Adriaenssens V, De Baets B, Goethals P, De Pauw N (2004) Fuzzy rule-based models for decision support in ecosystem management. Sci Total Environ 319(1):1–12

Agazzi OE, Kuo SS (1993) Hidden Markov model based optical character recognition in the presence of deterministic transformations. Pattern Recogn 26(12):1813–1826

Albiol J, Campmajó C, Casas C, Poch M (1995) Biomass estimation in plant cell cultures: a neural network approach. Biotechnol Prog 11:88–92

Alexander J, Mozer M (1999) Template-based procedures for neural network interpretation. Neural Netw 12:479–498

Andrews R, Diederich J, Tickle AB (1995) Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowl-Based Syst 8:373–389

Archer NP, Wang S (1993) Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems. Decis Sci 24:60–67

Baldi P, Brunak S (2001) Bioinformatics: the machine learning approach. MIT Press, Cambridge, MA

Borodovsky M, McIninch J (1993) GENMARK: parallel gene recognition for both DNA strands. Comput Chem 17:123–133

Borodovsky M, Rudd KE, Koonin EV (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. Nucleic Acids Res 22(2):4756–4767

Borodovsky M, Mclninch JD, Koonin EV, Rudd KE, Médigue C, Danchin A (1995) Detection of new genes in a bacterial genome using Markov models for three gene classes. Nucleic Acids Res 23:3554–3562

Chuong BD, Serafim B (2008) What is the expectation maximization algorithm? Nat Biotechnol 26(8):897–899

Colbourn E (2003) Neural computing: enable intelligent formulations. Pharm Technol Suppl 16–20

Coppola EA, Rana AJ, Poulton MM, Szidarovszky F, Uhl VW (2005) A neural network model for predicting aquifer water level elevations. Ground Water 43(2):231–241

Coruzzi GM, Burga AR, Katari MS, Gutiérrez RA (2009) Systems biology: principles and applications in plant research. Annu Plant Rev 35:3–40

Denton JW (1995) How good are neural networks for casual forecasting? J Bus Forecast Methods Syst 14(2):17–20

D'haeseleer P (2005) How does gene expression clustering work? Nat Biotechnol 23:1499–1502

Eichner J, Zeller G, Laubinger S, Ratsch G (2011) Support vector machines-based identification of alternative splicing in Arabidopsis thaliana from whole-genome tiling arrays. BMC Bioinform 12(1):5

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Feng Y, Xue Q (2006) The serine carboxypeptidase like gene family of rice (Oryza sativa L. ssp. japonica). Funct Integr Genomics 6(1):14–24

Filippi AM, Archibald R (2009) Support vector machine-based endmember extraction. IEEE Trans Geosci Remote Sens 47(3):771–791

Frossyniotis D, Anthopoulos Y, Kintzios S, Moschopoulou G, Yialouris C (2008) Artificial neural network selection for the detection of plant viruses. World J Agric Sci 4(1):114–120

Fukuda T, Shiotani S, Aral F, Takeuchi N, Sasaki K, Kimoshita T (1991) Cell recognition by image processing. 1st report. Recognition of dead or alive plant cells by neural network. Trans Jpn Soc Mech Eng Ser 57:77–84

Gago J, Landín M, Gallego P (2010a) Strengths of artificial neural networks in modeling complex plant processes. Plant Signal Behav 5(6):743–774

Gago J, Landín M, Gallego PP (2010b) Artificial neural networks modeling the *in vitro* rhizogenesis and acclimatization of *Vitis vinifera* L. J Plant Physiol 167:1226–1231

Gago J, Martínez-Núñez L, Landín M, Gallego P (2010c) Artificial neural networks as an alternative to the traditional statistical methodology in plant research. J Plant Physiol 167:23–27

Gallego PP, Gago J, Landín M (2011) Artificial neural networks technology to model and predict plant biology process. In: Suzuki K (ed) Artificial neural networks-methodological advances and biomedical applications. Intech, Rijeka, Carotia, pp 197–216

Gevrey M, Dimopoulos I, Lek S (2003) Review and comparison of methods to study the contribution of variables in artificial neural networks. Ecol Model 160(3):249–264

Glezakos TJ, Moschopoulou G, Tsiligiridis TA, Kintzios S, Yialouris CP (2010) Plant virus identification based on neural networks with evolutionary preprocessing. Comput Electron Agric 70:263–275

Gonzalez S (2000) Neural networks for macroeconomic forecasting: a complementary approach to linear regression models, Working paper. Department of Finance, Canada, pp 2000–2007

Gu L, Guo R (2007) Genome-wide detection and analysis of alternative splicing for nucleotide binding site-leucine-rich repeats sequences in rice. J Genet Genomics 34(3):247–257

Hanai T, Katayama A, Honda H, Kobayaski T (1997) Automatic fuzzy modelling for Ginjo sake brewing process using fuzzy neural network. J Chem Eng Jpn 30:94–100

Hayashi Y, Buckley JJ, Czogala E (1993) Fuzzy neural network with fuzzy signals and weights. Int J Intell Syst 8:527–537

Haykin S (2003) Neural networks: a comprehensive foundation, fourth Indian reprint. Pearson Education, Singapore

Hilbert DW, Ostendorf B (2001) The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates. Ecol Model 146:311–327

Hill T, Marquez L, O'Connor M, Remus W (1994) Artificial neural network models for forecasting and decision making. Int J Forecast 10:5–15

Honda H, Takikawa N, Noguchi H, Hanai T, Kobayashi T (1997) Image analysis associated with a fuzzy neural network and estimation of shoot length of regenerated rice callus. J Ferment Bioeng 84:342–347

Honda H, Ito T, Yamada J, Hanai T, Matsuoka M, Kobayashi T (1999) Selection of embryogenic sugarcane callus by image analysis. J Biosci Bioeng 87(5):700–702

Hua S, Sun Z (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 308(2):397–408

Huang Y (2009) Advances in artificial neural networks – methodological development and application. Algorithms 2:973–1007

Jiménez D, Pérez-Uribe A, Satizábal H, Barreto M, Van Damme P, Marco T (2008) A survey of artificial neural network-based modeling in agroecology. In: Prasad B (ed) Soft computing applications in industry, STUDFUZZ. Springer, Berlin/Heidelberg, pp 247–269

Juang B-H, Rabiner LR (1991) Hidden Markov models for speech recognition. Technometrics 33(3): 251–272

Karim MN, Yoshida T, Rivera SL, Saucedo VM, Eikens B, Oh G-S (1997) Global and local neural network models in biotechnology: application to different cultivation processes. J Ferment Bioeng 83(1):1–11

Karimi Y, Prasher S, Patel R, Kim S (2006) Application of support vector machine technology for weed and nitrogen stress detection in corn. Comput Electron Agric 51(1):99–109

Kaul M, Hill RL, Walthall C (2005) Artificial neural networks for corn and soybean yield prediction. Agric Syst 85:1–18

Kaundal R, Saini R, Zhao PX (2010) Combining machine learning and homology-based approaches to accurately predict subcellular localization in Arabidopsis. Plant Physiol 154(1):36–54

Kehagias A, Panagiotou H, Maslaris N, Petridis V, Petrou L (1998) Predictive modular neural networks methods for prediction of sugar beet crop yield. In: IFAC conference on control applications and ergonomics in agriculture, Athens, pp 1–5

Kell DB, Darby RM, Draper J (2001) Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. Plant Physiol 126:943–951

Korf I (2004) Gene finding in novel genomes. BMC Bioinform 5:59

Krenker A, Bešter J, Kos A (2011) Introduction to the artificial neural networks. In: Suzuki K (ed) Artificial neural networks-methodological advances and biomedical applications, Intech, Rijeka, Carotia, pp 3–18

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D (1994) Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol 235(5):1501–1531

Kruschke JK, Movellan JR (1991) Benefits of gain: speeded learning and minimal hidden layers in backpropagation networks. IEEE Trans Syst Man Cybern 21(1):273–280

Landín M, Rowe R, York P (2009) Advantages of neurofuzzy logic against conventional experimental design and statistical analysis in studying and developing direct compression formulations. Eur J Pharm Sci 38:325–331

Lawrence CE, Reilly AA (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins 7:41–51

Lin M, Hu B, Chen L, Sun P, Fan Y, Wu P, Chen X (2009) Computational identification of potential molecular interactions in Arabidopsis. Plant Physiol 151(1):34–46

Mahendra V, Prasad VSS, Dutta Gupta S (2004) Trichromatic sorting of in vitro regenerated plants of gladiolus using adaptive resonance theory. Curr Sci 87:348–353

Majoros WH, Pertea M, Antonescu C, Salzberg SL (2003) GlimmerM, exonomy and unveil: three ab initio eukaryotic genefinders. Nucleic Acids Res 31(13):3601–3604

Mandic DP, Chambers J (2001) Recurrent neural networks for prediction: learning algorithms, architectures and stability. Wiley, Chichester/New York

Marcuzzo M, Quelhas P, Campilho A, Mendonça AM (2008a) Automatic cell segmentation from confocal microscopy images of the Arabidopsis root. ISBI 2008 5th IEEE international symposium on Biomedical Imaging: From Nano to Macro. 712–771

Marcuzzo M, Quelhas P, Campilho A, Mendonça AM, Campilho A (2008b) A hybrid approach for Arabidopsis root cell image segmentation. In: Campilha A, Kamel M (eds) Image analysis and recognition. Springer Berlin Heidelberg, pp 739–749

Marcuzzo M, Quelhas P, Campilho A, Maria Mendonça A, Campilho A (2009) Automated Arabidopsis plant root cell segmentation based on SVM classification and region merging. Comput Biol Med 39:785–793

Martinez PJ, Pérez RM, Plaza A, Aguilar PL, Cantero MC, Plaza J (2006) Endmember extraction algorithms from hyperspectral images. Ann Geophys 49(1):93–101

Mehrotra S, Prakash O, Mishra BN, Dwevedi B (2008) Efficiency of neural networks for prediction of in vitro culture conditions and inoculum properties for optimum productivity. Plant Cell Tissue Organ Cult 95:29–35

Mehrotra S, Prakash O, Khan F, Kukreja A (2013) Efficiency of neural network-based combinatorial model predicting optimal culture conditions for maximum biomass yields in hairy root cultures. Plant Cell Rep 32:309–317

Molto E, Harrell RC (1993) Neural network classification of sweet potato embryos. Optics Agric Forest, Proc SPIE 1836:239–249

Morimoto T, Hashimoto Y (2000) An intelligent control for greenhouse automation, oriented by the concepts of SPA and SFA – an application to a post-harvest process. Comput Electron Agric 29:3–20

Morimoto T, De Baerdemaeker J, Hashimoto Y (1997) An intelligent approach for optimal control of fruit-storage process using neural networks and genetic algorithms. Comput Electron Agric 18:205–224

Mozer MC, Smolensky P (1989) Using relevance to reduce network size automatically. Connect Sci 1:3–16

Nitze I, Schulthess U, Asche H (2012) Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. Proceedings of the 4th GEOBIA Janeiro, Brazil, 35–40

Noguchi N, Terao H (1997) Path planning of an agricultural mobile robot by neural network and genetic algorithm. Comput Electron Agric 18:187–204

Osama K, Somvanshi P, Pandey AK, Mishra BN (2013) Modelling of nutrient mist reactor for hairy root growth using artificial neural network. Eur J Sci Res 97(4):516–526

Prakash O, Mehrotra S, Krishna A, Mishra BN (2010) A neural network approach for the prediction of in vitro culture parameters for maximum biomass yields in hairy root cultures. J Theor Biol 265:579–585

Prasad V, Gupta SD (2006) Applications and potentials of artificial neural networks in plant tissue culture. In: Gupta SD, Ibaraki Y (eds) Plant tissue culture engineering. Springer Netherlands, pp 47–67

Resop JP (2006) A comparison of artificial neural networks and statistical regression with biological

resources applications. University of Maryland, College Park, USA
Russell SJ, Norvig P, Canny JF, Malik JM, Edwards DD (1995) Artificial intelligence: a modern approach. Prentice Hall, Englewood Cliffs
Ruan R, Xu J, Zhang C, Chi C-M, Hu W-S (1997) Classification of plant somatic embryos by using neural network classifiers. Biotechnol Prog 13:741–746
Salas J, Markus M, Tokar A (2000) Streamflow forecasting based on artificial neural networks. In: Govindaraju RS, Roa AR (eds) Artificial neural networks in hydrology. Springer Netherlands, pp 23–51
Shao Q, Rowe RC, York P (2006) Comparison of neurofuzzy logic and neural networks in modelling experimental data of an immediate release tablet formulation. Eur J Pharm Sci 28(5):394–404
Shi L, Duan Q, Ma X, Weng M (2012) The research of support vector machine in agricultural data classification. In: Li D, Chen Y (eds) Computer and computing technologies in agriculture V. Springer Berlin Heidelberg, pp 265–269
Shigidi A, Garcia LA (2003) Parameter estimation in groundwater hydrology using artificial neural networks. J Comput Civ Eng 17:281–289
Stanimirovic PS, Miladinovic MB (2010) Accelerated gradient descent methods with line search. Numer Algorithms 54(4):503–520
Suroso MH, Tani A, Hoami N, Takigawa H, Nishiura Y (1996) Inverse technique for analysis of convective heat transfer over the surface of plant culture vessel. Trans ASAE 39:2277–2282
Tani A, Murase H, Kiyota M, Honami N (1992) Growth simulation of alfalfa cuttings in vitro by Kalman filter neural network. Acta Hort 319:671–676
Tickle AB, Andrews R, Golea M, Diederich J (1998) The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. IEEE Trans Neural Netw 9:1057–1068
Uozumi N, Yoshino T, Shiotani S, Suehara K-I, Arai F, Fukuda T, Kobayashi T (1993) Application of image analysis with neural network for plant somatic embryo culture. J Ferment Bioeng 76:505–509
Venkataramanan L, Sigworth F (2002) Applying hidden Markov models to the analysis of single ion channel activity. Biophys J 82:1930
White H (1992) Artificial neural networks: approximation and learning theory. Blackwell Publishers, Inc., Oxford/Cambridge
Widrow B, Hoff M (1960) Adaptive switching circuits. In: 1960 IRE WESCON convention record, vol 4. IRE, New York, pp 96–104
Xiang C, Ding SQ, Lee TH (2005) Geometrical interpretation and architecture selection of MLP. IEEE Trans Neural Netw 16:84–96
Yang ZR (2006) A novel radial basis function neural network for discriminant analysis. IEEE Trans Neural Netw 17:604–612
Yao X (1999) Evolving artificial neural networks. Proc IEEE 87:1423–1447
Zealand CM, Burn DH, Simonovic SP (1999) Short term streamflow forecasting using artificial neural networks. J Hydrol 214(1):32–48
Zhang C, Timmis R, Hu W-S (1999) A neural network based pattern recognition system for somatic embryos of Douglas fir. Plant Cell Tissue Organ Cult 56:25–35