Hemant J. Purohit · Vipin Chandra Kalia
Ravi Prabhakar More *Editors*

Soft Computing for Biological Systems

Springer

Soft Computing for Biological Systems

Hemant J. Purohit • Vipin Chandra Kalia
Ravi Prabhakar More
Editors

Soft Computing for Biological Systems

Springer

Editors
Hemant J. Purohit
Environmental Biotechnology
and Genomics Division
CSIR-National Environmental
Engineering Research Institute
(NEERI)
Nagpur, Maharashtra, India

Vipin Chandra Kalia
Microbial Biotechnology and Genomics
CSIR-Institute of Genomics and Integrative
Biology (IGIB)
Delhi University Campus
Delhi, India

Ravi Prabhakar More
ADBS, Lab 18, Neural Stem Cell Program
TIFR-National Centre for Biological Sciences
Bangalore, Karnataka, India

ISBN 978-981-10-7454-7 ISBN 978-981-10-7455-4 (eBook)
https://doi.org/10.1007/978-981-10-7455-4

Library of Congress Control Number: 2018931480

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. part of Springer Nature.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Dedicated to our mentors
and
inspiration – the respected Mr. Dashrath
Manjhi:
The Mountain Man

Preface

The biological systems and their functions are driven by information stored in the genetic material, the DNA, and their expression is driven by different factors. The active units of these DNA sequences are genes, which also interact with each other to define a condition-specific expression. The soft computing approaches recognize the different patterns in DNA sequence and assign them biological relevance with available information. At times these patterns not only help in the classification of but also predict functionally active domains. These approaches are equally helpful in predicting protein-protein interaction. To understand any stressed scenario, there is need to predict gene networks by applying tools which can suggest differential gene expressions. The issue extends these tools in a wide range of models from bacteria to human cancers. We wish to present the status of diverse possibilities and our views and opinions to finally provide mankind with novel, innovative, and long-lasting strategies, in the book entitled *Soft Computing for Biological Systems*. The book provides insights into bioinformatics tools for neural networks, metagenomics data analysis, genetic barcoding, machine learning, and diagnostic predictions. The well-illustrated articles written by the experts in the area provide information on thrust scientific R&D areas and their future perspectives for the prospective researchers and graduate students – future of the scientific society. This book has reached its completion primarily due to the sincere efforts of the dedicated academic experts – to share their vision and wisdom. This collection of chapters has been presented in a manner which can benefit the curious minds of the society. We are indebted to all the people, whose invaluable contributions brought this book to fruition.

Delhi, India — Vipin Chandra Kalia

Contents

About the Editors

Dr. Vipin Chandra Kalia is presently working as emeritus scientist. He has been the chief scientist and the deputy director at Microbial Biotechnology and Genomics, CSIR-Institute of Genomics and Integrative Biology, Delhi. He is a professor in AcSIR who obtained his M.Sc. and Ph.D. in genetics from Indian Agricultural Research Institute, New Delhi. He has been elected as (1) fellow of the National Academy of Sciences (FNASc), (2) fellow of the National Academy of Agricultural Sciences (FNAAS), and (3) fellow of the Association of Microbiologists of India (FAMSc). His main areas of research are microbial biodiversity, bioenergy, biopolymers, genomics, microbial evolution, quorum sensing, quorum quenching, drug discovery, and antimicrobials. He has published more than 100 papers in scientific journals such as (1) *Nature Biotechnology*, (2) *Biotechnology Advances*, (3) *Trends in Biotechnology*, (4) *Annual Review of Microbiology*, (5) *Critical Reviews in Microbiology*, (6) *Bioresource Technology*, (7) *PLOS ONE*, (8) *BMC Genomics*, (9) *International Journal of Hydrogen Energy*, and (10) *Gene*. He has authored 14 book chapters. His works have been cited 4100 times with an h-index of 35 and an i10-index of 76 (http://scholar.google.co.in/citations?hl=en&user=XaUw-VIAAAAJ). He has edited seven books: (i) *Quorum Sensing vs Quorum Quenching: A Battle with No End in Sight* (2015), (http://link.springer.com/book/10.1007/978-81-322-1982-8), (ii) *Microbial Factories: Biofuels, Waste Treatment – Vol. 1* (2015) (http://link.springer.com/book/10.1007%2F978-81-322-2598-0), (iii) *Microbial Factories: Biodiversity, Biopolymers, Bioactive Molecules – Vol. 2* (2015) (http://link.springer.com/book/10.1007%2F978-81-322-2595-9), (iv) *Waste Biomass Management: A Holistic Approach* (2017) (http://www.springer.com/in/book/9783319495941), (v) *Drug Resistance in Bacteria, Fungi, Malaria, and Cancer* – Editors: Arora, Gunjan, Sajid, Andaleeb, Kalia, Vipin Chandra (Eds.) (http://www.springer.com/in/book/9783319486826), (vi) *Microbial Applications* – Vol. 1 Kalia, V. (Ed), Kumar, P. (Ed) (2017) (http://www.springer.com/in/book/9783319526652), and (vii) *Microbial Applications – Vol. 2* Kalia, V. (Ed) (2017) (http://www.springer.com/in/book/9783319526683). He is presently the editor-in-chief of the *Indian Journal of Microbiology* (2013–2021) and editor of (1) *Journal of Microbiology and Biotechnology* (Korea), (2) *International Scholarly Res. Network Renewable Energy*, (3) *Dataset Papers in Microbiology*, and (4) *PLOS ONE*. He is a life member of the following scientific societies: (1) the Society of

Biological Chemists of India; (2) Society for Plant Biochemistry and Biotechnology, India; (3) Association of Microbiologists of India; (4) Indian Science Congress Association; (5) BioEnergy Council of India; and (6) Biotech Research Society of India (BRSI). He can be contacted at vckalia@igib.res.in; vc_kalia@yahoo.co.in

Dr. Hemant J. Purohit is head of Environmental Biotechnology and Genomics Division, National Environmental Engineering Research Institute (CSIR), Nagpur. He is also a professor in AcSIR (Academy of Scientific and Innovative Research), New Delhi. He completed his PhD from Nagpur University in 1986. He has been involved in designing a strategy for capturing microbial diversity by interfacing culturable and DNA fingerprinting tools; developing genomics-based monitoring tools for EIA and bio-remediation process; studying stress-dependent microbial response using dynamic gene expression and its application in bioprocess optimization; developing better insights into microbial capacities for utilization of organics through genome sequence analysis, etc. He has been project coordinator for a number of high-value projects. He has 225 publications to his credit. His Google scholar citations is 4711 (as of June 5, 2017), and he has an h-index of 38 and i10-index of 111. He has supervised 25 PhD students and more than 100 MSC student dissertations. He is a highly distinguished scientist. He is a recipient of a number of prestigious awards, including Fogarty International Exchange Program Fellowship; Commonwealth Fellowship, Department of Biochemistry, University of Hull, UK; CSIR Research Fellowships (JRF and SRF); etc.

Dr. Ravi Prabhakar More is a bioinformatician at the National Centre for Biological Sciences in Bangalore, Karnataka. He completed his Ph.D. from Swami Ramanand Teerth Marathwada University (SRTMU) and CSIR-NEERI, India, in 2015. He has developed signature (regular expression)-based DNA BarID and matK-QR classifier software for the identification of bacteria and plant species. He has been involved in next-generation sequencing (NGS), whole-genome sequencing (WGS), and exome sequencing (WES) data analysis for the identification of genes responsible for human brain disorders and in developing automated bioinformatics pipeline by using Perl and shell scripts on high-performance computing (HPC) for NGS data mining; he has worked on insect transcriptomics and phylogenetics, bacterial genomics, metagenomics, and DNA barcoding. He has published 11 scientific papers in reputed journals. He has worked at the National Institute for Basic Biology, Okazaki, Japan, and developed combined supervised approach (naïve Bayesian and homology) for detecting horizontal gene transfer in microbes. He is a recipient of awards of "Outstanding Project Personnel" for contribution in R&D activity, CSIR-NEERI, Nagpur, India, in the year 2014.

Current Scenario on Application of Computational Tools in Biological Systems

1

Hemant J. Purohit, Hitesh Tikariha, and Vipin Chandra Kalia

Abstract

The uncertainties and complexities of biological system challenge analytical approach and process of understanding. The wet lab experiments supported by soft algorithms find a way to resolve these scenarios. In the last decade, the biological analytical approach has found tremendous shift in data generation and analysis capacities. From sequencing of DNA and RNA to prediction of 3D structure and function of protein, there are a wide array of soft tools to make the job of exploring a system lot easier. This development eases our understanding of gene networks, plasticity and pattern of gene expression at gene to epigenomic level. In this book, we attempted to document selected areas of biological system and their advances, which will be frontier areas.

Keywords

Databases · Epigenome · Gene networks · Omic tools · Plasticity · Signatures

1.1 Introduction

The biological research has seen rapid progress through the use of computational tools for understanding physiological events. However, with the advent of next generation sequencing, there has been an explosive generation of data at different

H. J. Purohit (✉) · H. Tikariha
Environmental Biotechnology and Genomics Division, CSIR-National Environmental Engineering Research Institute (NEERI), Nagpur, Maharashtra, India
e-mail: hj_purohit@neeri.res.in; hemantdrd@hotmail.com

V. C. Kalia
Microbial Biotechnology and Genomics, CSIR-Institute of Genomics and Integrative Biology (IGIB), Delhi University Campus, Delhi, India
e-mail: vckalia@igib.in; vc_kalia@yahoo.co.in

H. J. Purohit et al. (eds.), *Soft Computing for Biological Systems*,
https://doi.org/10.1007/978-981-10-7455-4_1

levels of cellular organisation. A deeper understanding of protein expression profiles further supported this phenomenon. This has brought the data generated by biological systems into the domain of the big data analysis. The soft computing and artificial intelligence have become a prerequisite for the field of biological research to unfold system phenomenon. Bioinformatics tools have now become an essential hand for every section of biological data not only for handling and processing but also for validating the wet laboratory experiments. The omics era actually has now started emerging out of its lag phase. The progress of every laboratory is based on how intelligently they are harnessing the analytical tools for shaping the log phase trend of their physiological understanding. Keeping all these ingredients in the mind, this book opens up the current recipes of biological data de-codification. It is an attempt to focus on a few key areas and define their present status. The different areas challenge the readers to exploit a diversity of tools for applications in biological systems.

1.2 Protein Structure Prediction and Interaction

From the protocol of protein assay to chromatography and finally to NMR, now time has brought the reliability and rapidity in understanding the same information by in silico protein structure and function prediction (Tikariha et al. 2016). Even for an unknown protein, the implementation starts with the determination of its primary sequence. There has been an intense shift in the simple prediction of the secondary and tertiary structure of a protein from geometrical-based programming to new machine learning algorithms. Spencer et al. (2015) have given a vivid detail on ab initio protein secondary and tertiary prediction with the help of deep learning network. The concept eliminates the need for large protein structure database with known predicted proteins. This concept along with the incorporation of dihedral angle, torsion angle, solvent accessible surface area, positions and interactions of hydrogen bonds data can make the structure prediction a piece of cake (Heffernan et al. 2015). Even protein sequence and PDB database are on the rise, which will add to our knowledge on protein folding. This database can help in training the programs, which can help them in predicting the folding pattern and hence proposing the structure of an unknown protein. Thus, to get a vivid insight, one of the chapters gives an idea about the application of machine learning advancement in protein structure prediction.

The prediction of protein 3D structure is followed by the challenge of unearthing its interaction with other molecules, such as DNA and mRNA, or even with another protein. This part of the study holds immense potential for application in cellular pharmacology and drug discovery. Majorly there are three methods to study protein-protein interactions (PPIs) such as (1) residue coupling, (2) prediction of binding surface patch and (3) assembly prediction (Keskin et al. 2016). Based on this information, dozens of tools to analyse interfacial changes and calculate residues physicochemical changes have been developed. The database is also being constructed where one can look for curated PPIs such as CORUM, HIPPIE, IntAct, SPIKE, etc. Exploration of an interaction of peptide chain is also on the rise, and there is a huge market build-up on using peptide as a therapeutic agent (Nevola

and Giralt 2015). Research avenues are also being built for modulation of PPI using small inhibitory molecules (Arkin et al. 2014). Computational tools can modulate the dynamics of protein structure and its behaviour in a particular solvent system. Integration of this data with the thermodynamics of molecule interactions and characteristics of amino acid residues involved in the interactions can simulate the interactive behaviour of two protein molecules. Protein interactions with smaller molecules can also assist in deciphering the signalling cascade; and so by understanding this, one can precisely regulate/modulate the machinery inside a cell. One can also detect crucial amino acid residue involved in PPI. An evolutionary biologist can also seek for changes in protein-protein interactions, which can be responsible for metabolic and phenotypic changes (Bartlett et al. 2016). Seeking this past trend and huge market potential, this chapter aims to provide a deep insight on PPIs.

Drug designing strategies are driven by the protein interactions with various other molecules, in which the target locations are very specific and are with minimum free energy levels. Data mining and drug discovery have been rising in the field of pharmacology in recent years (Lavecchia 2015). Machine learning systems mimic from nature's own cellular system, where a molecule can play multiple roles, and now this drives drug designing forward to poly-pharmacology (Lavecchia and Cerchia 2016). In silico analyses are carried out in designing multi-target drug relying on a huge database of ligands and protein 3D structure, docking dynamics and pattern-based designing of the molecule. This not only ensures the drug design but also its delivery to the site of action that is also a major concern for its effectiveness. The whole effectiveness of drug relies on the intracellular transporter system (Nigam 2015). Here algorithms on basis of nature of molecule, transporter protein and interaction between them can predict how well the drug can find its way into the cell and carry out their action. We have discussed the potential of drug transporters system in one of the chapters.

1.3 Emerging Areas in Tool Development

With the advent of sequencing technologies, there has been a progressive rise in computational tools (Kalia 2015; Koul et al. 2015; Yu et al. 2015; Ambardar et al. 2016; Koul and Kalia 2016; Kalia et al. 2017; Kumar et al. 2015, 2017; Meza-Lucas et al. 2016). From pairing and assembling, the sequence reads to their annotation as genes with different algorithms are becoming faster and accurate. In this, the foremost approach is to design a robust multiple sequence alignment (MSA) program. MSA is a key step for functional annotation, phylogenetic studies and a necessity for comparative genomics and metagenomics (Pooja et al. 2015). Most of the MSA tools such as CLUSTAL, MUSCLE, K-align and a lot more are based on de novo assembly and pairwise alignment by tree construction. These programs are good at handling a small set of sequences, but they become redundant while handling thousands of dataset. For overcoming this deficiency, new tools have been designed such as HAlign, a fast multiple similar DNA/RNA sequence

alignment (Zou et al. 2015), and PASTA, ultralarge MSA (Mirarab et al. 2015) which can resolve this issue. Even tools like GUIDANCE2 are introduced to detect unreliable alignment regions in MSA (Sela et al. 2015). The hardware driver limitations are also being resolved, which can be seen in the development of GPU named CUDA ClustalW v1.0, and these will accelerate the computation of large datasets (Hung et al. 2015). We have dealt in detail the development in the domain of MSA and its application in the sequence alignment.

Next-generation sequencing has brought the computational biology to a new level. The databases for DNA, mRNA and proteins are growing geometrically. The repositories such as NCBI, EMBL, IMG, MG-RAST, SILVA and RCSB PDB are among the most exhaustively used databases. The Web has a large number of repositories and analysis pipeline for each separate domain such as CRCDA for cancer, Cas-Analyzer, Omics Pipe, etc. (Fisch et al. 2015; Thangam and Gopal 2015; Park et al. 2017). With NGS, the cost is going down, and there has been a tremendous amount of metagenome data generation. It is thus demanding new tools for accurate and reliable processing of this huge datasets. The attention has now been laid on the interpretation of this data rather than functional and taxonomical categorisation. Machine learning techniques, deep neural network generation and highly sophisticated statistical analysis are being used to understand this data. Integrated approach has been wired to connect all the analysis pipeline. A programming language such as Pearl, Python, R and Ruby are extensively used for investigation of NGS data. Nowadays, the Python and R have become two hands for interpretation of complex biological system and aid in connecting new links between the large and different datasets. The demands put regular pressure on program developers for updating the algorithms; recently QIIME pipeline was updated with the incorporation of PhyloToAST, which boosts its species-level classification and gives more elaborative visual evaluation (Dabdoub et al. 2016).

Transcriptome analysis has even brought the sequencing and analysis of miRNA, piRNA and IncRNAs possible, which is a big deal for disease diagnostic especially in the case of cancer. Extraction of secondary data from sequenced and annotated primary data is now becoming a remarkable strategy. Genome construction from metagenome is a new technique developed recently employing the process of binning, coverage, reassembly and curation (Sangwan et al. 2016). Tools like CheckM are devised to check the quality of reconstructed genomes (Parks et al. 2015). Apart from the reconstruction of the genome, the scheme is being designed to understand community-level talks, gene transfer and resistance development (More et al. 2014; Kapley et al. 2015). Thus to make the reader aware of this vast area, a chapter has been dedicated to bioinformatics tools for NGS data analysis.

Genomic tools are not limited to sequence identification or characterisation but can be implemented as pattern search algorithms to generate signatures, which can be utilised as biomarkers for diagnostic purposes (Porwal et al. 2009; Bhushan et al. 2015). A genomic biomarker can be used as both prognostic and predictive biomarker. Due to its high sensitivity and high specificity, the medical industry is looking for the discovery of such biomarker for every type of diseases (Kalia and

Kumar 2015; Kalia et al. 2015, 2016; Kekre et al. 2015; Kumar et al. 2016; Lee et al. 2016; Puri et al. 2016). Cancer is one of the deadly diseases and hard to diagnose at initial stages and opens a wide door for exploration of the genomic marker. A whole bunch of biomarkers discovered till date for head and neck cancer have been presented in a recent review (Kang et al. 2015). The promising nature of biomarker application has provoked us to include a chapter on the use of genomic biomarker in the case of oral cancer.

1.4 Gene Networks and Plasticity

Cells represent a collection of very well-coordinated and synchronised interactions and movement of every molecule residing in it. This is due to inherited intelligence cell carries for regulating expression of genes for every desirable event. Understanding this network of genes and how they regulate various machineries of cells by modulating itself is a challengeable task. Exploration of gene network involves the study of their expression pattern. The biological phenomenon evolved over a period of time, with one gene, one expression and identified physiology to a now collection of genes but even with the most sophisticated tools not completely understood till date. Gene Expression Atlas, Gene Expression Commons, CODEX and many more single gene expression databases are being created of which BloodSpot is the recent one which provides the tree-based relationship between different gene expression profiles present in the database (Bagger et al. 2016). As genes are differentially expressed in diverse conditions, it provides the plasticity to the gene networking; the wide range of data need to be generated to predict even an interaction of a single gene that behaves as a node in a network. A database on gene plasticity named ImmuSort is already being released, which provides an electric sorting system for immune cells (Wang et al. 2015). Thus from different expression profiles of a single gene to linking its connection with the expression profile of another gene requires a network-based analysis and a mammoth database.

Artificial neural networks are a set of models designed to classify and predict the outcome from a provided data; hence they are widely used algorithms in gene network prediction. Feedforward neural network, radial basis function network, modular neural network and physical neural network are the general types of neural network that are routinely applied in such analyses. Implying the data within a given set of conditions, the network is designed to calculate the expression behaviour for a set of genes. We have discussed the beneficial role of above study in diagnostic prediction by the aid of gene expression profile and artificial neural network.

The array of the genetic circuit in the cells can be grouped into various categories and modules specialised to carry out a specific task. Carving out this module of gene network could render the task easier for decoding the process associated with it. We are mostly interested in a specific set of the gene network, which we can modulate in a way that achieves a specific task such as understanding the response

of a signal cascade when osmotic stress is faced by the cell. Exploration of each module can give an idea of complete genetic web collectively working in the cell. Realising the core importance of this idea, we have added a chapter on soft computing approaches to extract biologically significant gene network modules that presents how through computational convergence one can study such network module and function carried out by each module separately.

Not only understanding of gene network is essential but we should also know how we can create a network. Mapping of a network relies on data used for its creation which in our case would be gene expression profiles. This requires a series of expression data of every single gene than stacking them upon one another in time series or imposed variant conditions and in the last layering and connecting the links between each gene involved in the network. Either the supervised or unsupervised model can be used for creating a network. In a recent paper, authors describe the use of both the approaches for the creation of gene regulatory network (Huynh-Thu and Sanguinetti 2015). Single cell network synthesis toolkit has been used to identify an interconnected network of 20 transcription factors in human blood cell (Moignard et al. 2015). So to get acquainted about such emerging topic, we have incorporated a chapter which deals with the construction of gene network.

1.5 Epigenome: Emerging Area

All the current techniques target the pattern of the four nucleotides, thereby predicting its functions, but in the case of eukaryotes, this scenario changes. The methylation pattern under the epigenetic tag governs which gene will get expressed and which does not. Epigenomic research targets such molecule which can alter the expression pattern of the genes in a chromosome. The epigenetic study has two core areas – DNA methylation and histone modification. Methylation of DNA is usually on CpG islands and follows a particular pattern used to deduce the expression profile of gene under study. Techniques like methylation array detect DNA methylation, whereas ChIP sequencing determines a modification in histone. Both the tools have helped in generating an epigenetic map of the human chromosome. The epigenomic study is particularly interesting as it delivers the regulatory expression channel of gene thereby influencing the phenotypic expression. The cross-links through which epigenetic action is controlled by environmental factors are also a great issue of interest. Lots of epigenome-wide-associated studies have linked diets, smoking, stress, etc. to changes in genotypic and phenotypic variation in human (Lee et al. 2015; Provençal and Binder 2015). Realising such rising trend in the area of epigenomic, we have included a chapter on Module-Based Knowledge Discovery for Multiple-Cytosine-Variant Methylation Profile.

1.6 Expanding the Domain of Computational Statistical Analysis

With the expansion of biological data, lots of statistical tools have been developed to sort, group, analyse and predict the outcome from the data. Statistic combined with appropriate programming language results in more analytical approach and visually enhanced result. Along with the application of computational tools, various modelling techniques are also being integrated to understand the pest population dynamics (Whish et al. 2015; Gilioli et al. 2016). With so much focus on the application of computational statistic in the field of biotechnology, we introduce our reader to the domain of agriculture with such analysis. This chapter describes the role of various computing tools and techniques based on background statistical analysis for studying pest population dynamics.

1.7 Pattern Recognition/Barcoding/Diagnostics

Identification of species and determining its role in an environment are crucial step in ecological discernment. DNA barcoding is one of the emerging genomic tools to tackle this problem. Based on consensus pattern of a sequence, it aids in the identification of species (Kalia and Kumar 2015; Kalia et al. 2015, 2016; Kekre et al. 2015; Kumar et al. 2016; Lee and Rho 2016). DNA barcoding applies to all domain of life for their classification. A great deal of DNA barcode application till date has been broadly reviewed recently (Kress et al. 2015). Barcoding also allows revealing the diversity pattern of flora and fauna thereby producing a species map for niche/habitat (More and Purohit 2016; More et al. 2016).

The great nature of DNA barcoding is that it is applicable to every organism with minor modification. The initial step in barcoding is deducing the signature sequence in the species. After the identification, it can be used to tag every other species which have an exact signature (Porwal et al. 2009; Kalia et al. 2011; Bhushan et al. 2013). Thus the nature and location of code vary from species to species and organism to organism. The DNA barcode has a great role in conservative biology as it can help in tracking the species of interest. To open up the reader more about the application of DNA barcoding, we have discussed thoroughly the fish DNA barcoding as a model. This chapter also covers up the various bioinformatics tools and techniques deployed in generating a DNA barcode for a given species.

Earlier when DNA barcode was introduced, it was limited to eukaryotic organisms due to high mutation rate in prokaryotes and absence of mitochondrial or plastid DNA, which have rich consensus region. But now this scenario is changing, and bacterial DNA barcode is being introduced in recent years along with the introduction of meta-barcoding. Recent publications on marine benthic meta-barcoding have already laid down this trend (Leray and Knowlton 2015). In upcoming years we can expect the rise of meta-barcoding along with the metagenomics. For providing a complete package of tools and software used in bacteria DNA barcoding and analysis, reader can refer to a later chapter.

Pattern- and network-based computational analyses are not only limited to the microorganism or medical biology, but it has an expanded horizon in plant biology too. Earlier it was concentrated to the regime of plant classical genetics and breeding but gradually arose with plant genomics. The surge in plant genomics can be seen with the recent introduction of PLAZA 3.0 which is a server assisting in comparative plant genomics (Proost et al. 2015). Genomic analysis has already been extended to the study of metabolite-based quantitative trait loci. Identification of metabolites is one of the highlighted areas in plant metabolomics. Luo in 2015 discussed the genome-wide association studies based on metabolite. Not only genetic trait but analysis of phenotypic trait in plant biology is a keen area. The various repositories have been created to store phenotypic data for a selected plant species, e.g. MaizeGDB, Ephesis databases, etc. This has laid down the incorporation of microarray, metabolomics, sequencing and proteomics data in a single platform for understanding the link between phenotypic expression, genetic makeup and environmental factors. This has arisen the need for handling ample amount of data synchronising it with metadata (Krajewski et al. 2015). Modelling framework is also being applied in plant biology for better resolution of its cellular event (Boudon et al. 2015). Observing a high trend in the application of computational tools in the subject of plant biology, a vivid description of the integration of computational approach in plant biology and also its field application has been discussed in this book.

With metadata, biological systems are challenging the scientific community with its complexity. Covering different emerging disciplines in biology where computational approach is essential or playing an essential role has been discussed in this book, which will surely give the reader a new paradigm in their analytical processes.

References

Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J (2016) High throughput sequencing: an overview of sequencing chemistry. Indian J Microbiol 56:394–404. https://doi.org/10.1007/s12088-016-0606-4

Arkin MR, Tang Y, Wells JA (2014) Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. Chem Biol 21:1102–1114. https://doi.org/10.1016/j.chembiol.2014.09.001

Bagger FO, Sasivarevic D, Sohi SH, Laursen LG, Pundhir S, Sønderby CK, Winther O, Rapin N, Porse BT (2016) BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. Nucleic Acids Res 44(D1):D917–D924. https://doi.org/10.1093/nar/gkv1101

Bartlett M, Thompson B, Brabazon H, Del Gizzi R, Zhang T, Whipple C (2016) Evolutionary dynamics of floral homeotic transcription factor protein–protein interactions. Mol Biol Evol 33:1486–1501. https://doi.org/10.1093/molbev/msw031

Bhushan A, Joshi J, Shankar P, Kushwah J, Raju SC, Purohit HJ, Kalia VC (2013) Development of genomic tools for the identification of certain *Pseudomonas* up to species level. Indian J Microbiol 53:253–263. https://doi.org/10.1007/s12088-013-0412-1

Bhushan A, Mukherjee T, Joshi J, Shankar P, Kalia VC (2015) Insights into the origin of *Clostridium botulinum* strains: evolution of distinct restriction endonuclease sites in *rrs* (16S rRNA gene). Indian J Microbiol 55:140–150. https://doi.org/10.1007/s12088-015-0514-z

Boudon F, Chopard J, Ali O, Gilles B, Hamant O, Boudaoud A, Traas J, Godin C (2015) A computational framework for 3D mechanical modeling of plant morphogenesis with cellular resolution. PLoS Comput Biol 11:e1003950. https://doi.org/10.1371/journal.pcbi.1003950

Dabdoub SM, Fellows ML, Paropkari AD, Mason MR, Huja SS, Tsigarida AA, Kumar PS (2016) PhyloToAST: bioinformatics tools for species-level analysis and visualization of complex microbial datasets. Sci Rep 6. https://doi.org/10.1038/srep29123

Fisch KM, Meißner T, Gioia L, Ducom JC, Carland TM, Loguercio S, Su AI (2015) Omics pipe: a community-based framework for reproducible multi-omics data analysis. Bioinformatics 31:1724–1728. https://doi.org/10.1093/bioinformatics/btv061

Gilioli G, Pasquali S, Marchesini E (2016) A modelling framework for pest population dynamics and management: an application to the grape berry moth. Ecol Model 320:348–357

Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Yang Y, Zhou Y (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci Rep 5:11476. https://doi.org/10.1038/srep11476

Hung CL, Lin YS, Lin CY, Chung YC, Chung YF (2015) CUDA ClustalW: an efficient parallel algorithm for progressive multiple sequence alignment on multi-GPUs. Comput Biol Chem 58:62–68. https://doi.org/10.1016/j.compbiolchem.2015.05.004

Huynh-Thu VA, Sanguinetti G (2015) Combining tree-based and dynamical systems for the inference of gene regulatory networks. Bioinformatics 31:1614–1622. https://doi.org/10.1093/bioinformatics/btu863

Kalia VC (2015) Let's explore the latent features of genes to identify bacteria. J Mol Genet Med 9: e105. https://doi.org/10.4172/1747-0862.1000E105

Kalia VC, Kumar P (2015) Genome wide search for biomarkers to diagnose *Yersinia* infections. Indian J Microbiol 55:366–374. https://doi.org/10.1007/s12088-015-0552-6

Kalia VC, Mukherjee T, Bhushan A, Joshi J, Shankar P, Huma N (2011) Analysis of the unexplored features of *rrs* (16S rDNA) of the genus *Clostridium*. BMC Genomics 12:18. https://doi.org/10.1186/1471-2164-12-18

Kalia VC, Kumar P, Kumar R, Mishra A, Koul S (2015) Genome wide analysis for rapid identification of *Vibrio* species. Indian J Microbiol 55:375–383. https://doi.org/10.1007/s12088-015-0553-5

Kalia VC, Kumar R, Kumar P, Koul S (2016) A genome-wide profiling strategy as an aid for searching unique identification biomarkers for *Streptococcus*. Indian J Microbiol 56:46–58. https://doi.org/10.1007/s12088-015-0561-5

Kalia VC, Kumar R, Koul S (2017) In silico analytical tools for phylogenetic and functional bacterial genomics. In: Arora G, Sajid A, Kalia VC (eds) Drug resistance in bacteria, fungi, malaria and cancer. Springer, Cham, pp 339–355. https://doi.org/10.1007/978-3-319-48683-3_15. ISBN: 978-3-319-48682-6

Kang H, Kiess A, Chung CH (2015) Emerging biomarkers in head and neck cancer in the era of genomics. Nat Rev Clin Oncol 12:11–26. https://doi.org/10.1038/nrclinonc.2014.192

Kapley A, Liu R, Jadeja NB, Zhang Y, Yang M, Purohit HJ (2015) Shifts in microbial community and its correlation with degradative efficiency in a wastewater treatment plant. Appl Biochem Biotechnol 176:2131–2143. https://doi.org/10.1007/s12010-015-1703-2

Kekre A, Bhushan A, Kumar P, Kalia VC (2015) Genome wide analysis for searching novel markers to rapidly identify *Clostridium* strains. Indian J Microbiol 55:250–257. https://doi.org/10.1007/s12088-015-0535-7

Keskin O, Tuncbag N, Gursoy A (2016) Predicting protein–protein interactions from the molecular to the proteome level. Chem Rev 116:4884–4909. https://doi.org/10.1021/acs.chemrev.5b00683

Koul S, Kalia VC (2016) Comparative genomics reveals biomarkers to identify *Lactobacillus* species. Indian J Microbiol 56:253–263. https://doi.org/10.1007/s12088-016-0605-5

Koul S, Kumar P, Kalia VC (2015) A unique genome wide approach to search novel markers for rapid identification of bacterial pathogens. J Mol Genet Med 9:194. https://doi.org/10.4172/1747-0862.1000194

Krajewski P, Chen D, Ćwiek H, van Dijk AD, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP, van Oeveren J, Pommier C, Scholz U, van Schriek M, Usadel B, Weise S (2015) Towards recommendations for metadata and data handling in plant phenotyping. J Exp Bot 66:5417–5427. https://doi.org/10.1093/jxb/erv271

Kress WJ, García-Robledo C, Uriarte M, Erickson DL (2015) DNA barcodes for ecology, evolution, and conservation. Trends Ecol Evol 30:25–35

Kumar A, Mohanty NN, Chacko N, Yogisharadhya R, Shivachandra SB (2015) Structural features of a highly conserved Omp16 protein of *Pasteurella multocida* strains and comparison with related peptidoglycan-associated lipoproteins (PAL). Indian J Microbiol 55:50–56. https://doi.org/10.1007/s12088-014-04896-1

Kumar R, Koul S, Kumar P, Kalia VC (2016) Searching biomarkers in the sequenced genomes of *Staphylococcus* for their rapid identification. Indian J Microbiol 56:64–71. https://doi.org/10.1007/s12088-016-0565-9

Kumar R, Koul S, Kalia VC (2017) Exploiting bacterial genomes to develop biomarkers for identification. In: Arora G, Sajid A, Kalia VC (eds) Drug resistance in bacteria, fungi, malaria and cancer. Springer, Cham, pp 357–370. https://doi.org/10.1007/978-3-319-48683-3_16. ISBN: 978–3–319-48682-6

Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 20:318–331. https://doi.org/10.1016/j.drudis.2014.10.012

Lavecchia A, Cerchia C (2016) *In silico* methods to address polypharmacology: current status, applications and future perspectives. Drug Discov Today 21:288–298. https://doi.org/10.1016/j.drudis.2015.12.007

Lee S, Rho JY (2016) Development of a specific diagnostic system for detecting *Turnip Yellow Mosaic Virus* from Chinese cabbage in Korea. Indian J Microbiol 56:103–107. https://doi.org/10.1007/s12088-015-0557-1

Lee KW, Richmond R, Hu P, French L, Shin J, Bourdon C, Reischl E, Waldenberger M, Zeilinger S, Gaunt T, McArdle W, Ring S, Woodward G, Bouchard L, Gaudet D, Smith GD, Relton C, Paus T, Pausova Z (2015) Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. Environ Health Perspect 123:193. https://doi.org/10.1289/ehp.1408614.

Lee S, Kim CS, Shin YG, Kim JH, Kim YS, Jheong WH (2016) Development of nested PCR-based specific markers for detection of peach rosette mosaic virus in plant quarantine. Indian J Microbiol 56:108–111. https://doi.org/10.1007/s12088-015-0548-2

Leray M, Knowlton N (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proc Natl Acad Sci 112:2076–2081. https://doi.org/10.1073/pnas.1424997112

Luo J (2015) Metabolite-based genome-wide association studies in plants. Curr Opin Plant Biol 24:31–38. https://doi.org/10.1016/j.pbi.2015.01.006

Meza-Lucas A, Pérez-Villagómez M, Martínez-López JP, García-Rodea R, Martínez-Castelán MG, Escobar-Gutiérrez A, de la Rosa-Arana JL, Villanueva-Zamudio A (2016) Comparison of DOT-ELISA and Standard-ELISA for detection of the *Vibrio cholerae* toxin in culture supernatants of bacteria isolated from human and environmental samples. Indian J Microbiol 56:379–382. https://doi.org/10.1007/s12088-016-0596-2

Mirarab S, Nguyen N, Guo S, Wang LS, Kim J, Warnow T (2015) PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J Comp Biol 22:377–386. https://doi.org/10.1089/cmb.2014.0156

Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, Nishikawa S, Piterman N, Kouskoff V, Theis FJ, Fisher J, Göttgens B (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol 33:269–276. https://doi.org/10.1038/nbt.3154

More RP, Purohit HJ (2016) The identification of discriminating patterns from 16S rRNA gene to generate signature for *Bacillus* genus. J Comp Biol 23:651–661. https://doi.org/10.1089/cmb.2016.0002

More RP, Mitra S, Raju SC, Kapley A, Purohit HJ (2014) Mining and assessment of catabolic pathways in the metagenome of a common effluent treatment plant to induce the degradative capacity of biomass. Bioresour Technol 153:137–146. https://doi.org/10.1016/j.biortech.2013.11.065

More RP, Mane RP, Purohit HJ (2016) matK-QR classifier: a patterns based approach for plant species identification. BioData Min 9:39. https://doi.org/10.1186/s13040-016-0120-6

Nevola L, Giralt E (2015) Modulating protein–protein interactions: the potential of peptides. Chem Commun 51:3302–3315. https://doi.org/10.1039/c4cc08565e

Nigam SK (2015) What do drug transporters really do? Nat Rev Drug Discov 14:29–44. https://doi.org/10.1038/nrd4461

Park J, Lim K, Kim JS, Bae S (2017) Cas-analyzer: an online tool for assessing genome editing results using NGS data. Bioinformatics 33:286–288. https://doi.org/10.1093/bioinformatics/btw561

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114

Pooja S, Pushpanathan M, Jayashree S, Gunasekaran P, Rajendhran J (2015) Identification of periplasmic a-amlyase from cow dung metagenome by product induced gene expression profiling (Pigex). Indian J Microbiol 55:57–65. https://doi.org/10.1007/s12088-014-0487-3

Porwal S, Lal S, Cheema S, Kalia VC (2009) Phylogeny in aid of the present and novel microbial lineages: diversity in *Bacillus*. PLoS One 4:e4438. https://doi.org/10.1371/journal.pone.0004438

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K (2015) PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res 43 (D1):D974–D981. https://doi.org/10.1093/nar/gku986

Provençal N, Binder EB (2015) The effects of early life stress on the epigenome: from the womb to adulthood and even before. Exp Neurol 268:10–20. https://doi.org/10.1016/j.expneurol.2014.09.001

Puri A, Rai A, Dhanaraj PS, Lal R, Patel DD, Kaicker A, Verma M (2016) An *in silico* approach for identification of the pathogenic species, *Helicobacter pylori* and its relatives. Indian J Microbiol 56:277–286. https://doi.org/10.1007/s12088-016-0575-7

Sangwan N, Xia F, Gilbert JA (2016) Recovering complete and draft population genomes from metagenome datasets. Microbiome 4:8. https://doi.org/10.1186/s40168-016-0154-5

Sela I, Ashkenazy H, Katoh K, Pupko T (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res 43 (W1):W7–W14. https://doi.org/10.1093/nar/gkv318

Spencer M, Eickholt J, Cheng J (2015) A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans Comp Biol Bioinform 12:103–112. https://doi.org/10.1109/TCBB.2014.2343960

Thangam M, Gopal RK (2015) CRCDA – Comprehensive resources for cancer NGS data analysis. Database 2015:bav092. https://doi.org/10.1093/database/bav092

Tikariha H, Pal RR, Qureshi A, Kapley A, Purohit HJ (2016) *In silico* analysis for prediction of degradative capacity of *Pseudomonas putida* SF1. Gene 591:382–392. https://doi.org/10.1016/j.gene.2016.06.028

Wang P, Yang Y, Han W, Ma D (2015) ImmuSort, a database on gene plasticity and electronic sorting for immune cells. Sci Rep 5:10370. https://doi.org/10.1038/srep10370

Whish JP, Herrmann NI, White NA, Moore AD, Kriticos DJ (2015) Integrating pest population models with biophysical crop models to better represent the farming system. Environ Model Softw 72:418–425

Yu S, Peng Y, Zheng Y, Chen W (2015) Comparative genome analysis of *Lactobacillus casei*: insights into genomic diversification for niche expansion. Indian J Microbiol 55:102–107. https://doi.org/10.1007/s12088-014-0496-2

Zou Q, Hu Q, Guo M, Wang G (2015) HAlign: fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. Bioinformatics 31(15):2475–2481. https://doi.org/10.1093/bioinformatics/btv177

Diagnostic Prediction Based on Gene Expression Profiles and Artificial Neural Networks

2

Eugene Lin and Shih-Jen Tsai

Abstract

Recent advances in scientific research point out that diagnostic prediction represents a novel paradigm because of the decreased expense and the expanded productivity of multi-omics technologies such as gene expression profiling. In order to evaluate a mammoth amount of biomarkers produced by high-throughput technologies, machine learning and predictive approaches such as artificial neural network (ANN) algorithms have widely been utilized to assess disease mechanisms and intervention outcomes. In this chapter, we first illustrated ANN algorithms for establishing biomarkers in diagnostic prediction studies. We then surveyed a variety of diagnostic prediction applications for numerous diseases and treatments with consideration of ANN algorithms and gene expression profiling. Finally, we outlined their limitations and future directions. Future work in diagnostic prediction studies promises to lead to innovative ideas related to disease prevention and drug responsiveness in light of multi-omics technologies as well as machine learning and predictive algorithms.

Keywords

Gene expression · Artificial neural networks · Machine learning · Chemotherapy · Schizophrenia

E. Lin
Institute of Biomedical Sciences, China Medical University, Taichung, Taiwan

Department of Electrical Engineering, University of Washington, Seattle, WA, USA

TickleFish Systems Corporation, Seattle, WA, USA

S.-J. Tsai (✉)
Department of Psychiatry, Taipei Veterans General Hospital, Taipei, Taiwan

Division of Psychiatry, National Yang-Ming University, Taipei, Taiwan
e-mail: tsai610913@gmail.com

H. J. Purohit et al. (eds.), *Soft Computing for Biological Systems*,
https://doi.org/10.1007/978-981-10-7455-4_2

2.1 Introduction

In this chapter, we briefly describe some key emerging diagnostic prediction studies for various diseases and treatments of significance for public health with consideration of gene expression profiles and machine learning algorithms such as artificial neural network (ANN) models (Lin and Tsai 2011). This review is not intended as a comprehensive survey of all possible diagnostics applications studied in the literature.

First, we described machine learning and predictive algorithms such as ANN models that have been widely used in the research community for pinpointing biomarkers as well as for associating with diseases and drug responses in the diagnostic prediction studies. Furthermore, we surveyed some potential biomarkers that were investigated in the diagnostic prediction studies using gene expression profiles and ANN algorithms and were reported to be linked with disease status or drug efficacy. Moreover, we highlighted the limitations and future outlook regarding the diagnostic prediction studies in terms of gene expression profiles as well as machine learning and predictive algorithms. In future work, replication studies with extensive and independent cohorts will be indispensable in order to establish the characteristics of the potential biomarkers identified in the diagnostic prediction studies in disease diagnosis as well as treatment response (Lin 2012; Lin and Tsai 2012).

2.2 Machine Learning and Artificial Neural Networks

Machine learning and predictive methods contain computer algorithms which are able to naturally perceive complicated patterns based on empirical data (Kononenko 2001; Lane et al. 2012; Lin and Tsai 2016c). The objective of machine learning and predictive algorithms is to facilitate computer algorithms to gain from data of the past or present and then make decisions or predictions for unrecognized forthcoming circumstances by utilizing that knowledge (Landset et al. 2015; Lin and Tsai 2016c). In the general terms, the workflow (as shown in Fig. 2.1) for a machine learning and predictive algorithm incorporates three phases including construct the model from pattern inputs, appraise and refine the model, and then establish the model into construction in prediction-making (Landset et al. 2015). In other words, machine learning and predictive algorithms for classification

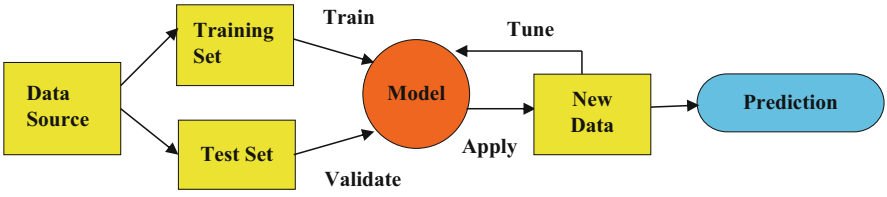

Fig. 2.1 Machine learning workflow

applications such as medical diagnosis or diagnostic prediction are procedures for adopting the best assumption from a set of alternatives that are qualified for a set of observations (Witten and Frank 2005). The strengths of machine learning and predictive algorithms for classification, including nonlinearity, fault tolerance, and real-time operation, make them suitable for complicated applications (Lane et al. 2012).

ANN models, such as multilayer feedforward neural networks, can be frequently utilized to solve complicated applications in classification and predictive modeling due to the fact that ANN algorithms possess the benefits of fault tolerance, nonlinearity, integrality, and real-time operations (Lin et al. 2006; Kung and Hwang 1998). A multilayer feedforward neural network is one category of ANN algorithms where networks between entities construct no directed cycles (Bishop 1995). In other words, a loop or cycle does not exist in the network because the data only relays in an onward order from the input entities, by means of the hidden entities (if any), and then to the output entities.

Moreover, from an algorithmic point of view, the primary operation of this ANN is separated into the learning and retrieving stages (Kung and Hwang 1998). In the learning stage of this ANN, the back-propagation algorithm (Rumelhart et al. 1996) is adopted for the learning scheme. Furthermore, in the retrieving stage, this ANN repeats through all the panels to achieve the retrieval response at the output panel in keeping with the inputs of test patterns. On the other hand, from a structural point of view, this ANN is an iterative and spatial neural network that possesses numerous panels of hidden neuron groups among the input and output neuron panels (Kung and Hwang 1998).

The ANN models can be executed using favored machine learning tools such as R (the R Project for Statistical Computing; http://www.r-project.org/) or the Waikato Environment for Knowledge Analysis (WEKA) software (Witten and Frank 2005). However, popular open-source machine learning tools including R and WEKA are not originally constructed and implemented for large-scale data (Landset et al. 2015). To effortlessly design and adopt for big data, there are assorted machine learning tools, such as Mahout (http://mahout.apache.org/), MLlib (https://spark.apache.org/mllib/), H2O (http://h2o.ai/), and SAMOA (https://github.com/samoa-moa/samoa-moa), available to run in a distributed environment (Lin and Tsai 2016c).

2.3 Gene Expression Profile

Noncoding RNAs, such as long noncoding RNAs and small noncoding RNAs, are distinct from their complement messenger RNAs (mRNAs) because the sequence of nucleotides in noncoding RNAs encodes no proteins (Nagano and Fraser 2011; Lin and Tsai 2016a). While long noncoding RNAs represent transcripts with more than 200 nucleotides in length, small noncoding RNAs, such as the microRNAs, are smaller than 200 nucleotides in length.

The microRNAs govern gene expression by regulating mRNA translation, stability, and degradation (Dwivedi 2014; Lin and Tsai 2016a). The characteristics of mRNAs, microRNAs, and long noncoding RNAs in examining disease pathogenesis and in keeping track of response to treatment for human disease are developing rapidly. Future work will be conducted to assess whether gene expression profiling including mRNAs, microRNAs, and long noncoding RNAs may be established as potential biomarkers with respect to human disease and therapeutic responses (Lin and Tsai 2016a).

2.4 Gene Expression Profile Studies with ANN

Table 2.1 summarizes the relevant diagnostic prediction studies by using gene expression profile and ANN models. This is by no means a comprehensive survey of all probable diagnostic prediction studies discovered so far. Nonetheless, a growing body of studies has been investigated when scientists remain to pay much attention to diagnostic prediction research.

2.4.1 Cancer

There were a variety of diagnostic prediction studies for cancer research using ANN models and gene expression profiling. First, Pass et al. (2004) trained a three-layer ANN model based on the expression value of differentially regulated genes and derived a set of 27 genes that distinguishes good-risk and poor-risk surgically

Table 2.1 Diagnostic prediction studies of gene expression profiling for various diseases and treatments of significance using artificial neural networks

Disease/treatment	Results	References
Malignant pleural mesothelioma	Achieved 76% accuracy	Pass et al. (2004)
Neuroblastoma	Achieved 88% accuracy	Wei et al. (2004)
Astrocytic brain tumors	Identified an optimum set of 37 genes	Petalidis et al. (2008)
Breast cancer	Reduced a 70-gene signature to nine genes	Lancashire et al. (2010)
Schizophrenia	Achieved 87.9% accuracy	Takahashi et al. (2010)
Diffuse large B-cell lymphoma	Achieved 93% accuracy	Mehridehnavi and Ziaei (2013)
Luminal A-like breast cancer	Revealed ten microRNAs for further analysis	McDermott et al. (2014)
Childhood sarcomas	Showed strong connection links on certain genes	Tong et al. (2014)
Chemotherapy in non-small cell lung cancer	Achieved 65.71% accuracy	Chen et al. (2015)

treated patients with malignant pleural mesothelioma. A rare and aggressive cancer called malignant pleural mesothelioma usually evolves in the thin row of tissue neighboring the lungs known as the pleura. Of the 27 genes revealed to be significant, 18 have been intensely investigated in the literature, and few have been linked with malignant pleural mesothelioma (Pass et al. 2004).

Secondly, Wei et al. (2004) utilized gene expression profiles from cDNA microarrays to forecast the outcome and extract a minimal gene set in patients with neuroblastoma by using ANN models. Neuroblastoma is the most common cancer in childhood and in infancy. They suggested that the top 24 ANN-ranked clones, which represented 19 unique genes as a minimal gene set, resulted in the minimal classification error. Wei et al. (2004) also indicated that ANN models can predict additional patients according to their survival status based on either all genes or in particular the 19 genes.

Thirdly, Petalidis et al. (2008) assessed whether molecular signatures can define survival prognostic subclasses of astrocytic tumors by using gene expression data from 65 highly annotated tumors and a simple ANN model in the form of a single-layer perceptron. Astrocytic tumors are the most common type of cancer in the brain. They analyzed the ANN model to optimize leave-one-out cross-validation runs, which resulted in an optimum set of 37 genes. Petalidis et al. (2008) selected two genes of special interest, *PEA15* and *ADM*, for further analysis in their study.

In addition, Lancashire et al. (2010) leveraged a previously published dataset of breast cancer and applied an ANN approach to identify an optimal gene expression signature for predicting the outcome of patients with breast cancer. Lancashire et al. (2010) found that only nine genes were needed to forecast metastatic spread with sensitivity of 98% by utilizing an ANN algorithm implemented especially for the optimal biomarker subgroups in gene expression data.

Moreover, Mehridehnavi and Ziaei (2013) utilized ANN models to find the most significant genes and classify patients with diffuse large B-cell lymphoma, which is a cancer of B cells, on the basis of their gene expression profiles. Diffuse large B-cell lymphoma is a form of white blood cell responsible for generating antibodies. Mehridehnavi and Ziaei (2013) used the signal-to-noise ratio as a major approach to reduce the number of genes from 4026 to 2 most significant genes. By using two most significant genes to train the ANN model, their results showed that the training and testing errors were 0% and 7%, respectively (Mehridehnavi and Ziaei 2013).

Furthermore, based on a cDNA microarray dataset, Tong et al. (2014) utilized ANN models to find the potential gene-gene interactions among previously determined biomarkers in children sarcomas, which are a rare kind of cancer arising from transformed cells of mesenchymal origin. Their analysis revealed that seven key genes including *FCGRT*, *FNDC5*, *GATA2*, *HLA-DPB1*, *MT1L*, *OLFM1*, and *TNNT1* had significant associations (Tong et al. 2014).

Finally, McDermott et al. (2014) employed ANN models and microarray profiling to pinpoint circulating microRNAs that were expressed in a differential manner among individuals with luminal A-like breast cancer in comparison to those without luminal A-like breast cancer. They found 76 microRNAs with differential expression in subjects with luminal A-like breast cancer and also identified 10 microRNAs for further analysis using ANN models (McDermott et al. 2014).

2.4.2 Chemotherapy

The use of genetic information and other biomarkers has played a major role in better predicting patients' responses to targeted therapy. For example, adjuvant chemotherapy for non-small cell lung cancer can be used after surgery to put an end to recurrence or metastases. Unfortunately, not every patient is suitable for treatment. Chen et al. (2015) aimed to construct prediction models to recognize who was suitable for adjuvant chemotherapy in subjects with non-small cell lung cancer. Their analysis showed that the best ANN model achieved 65.71% accuracy with two genes such as *DUSP6* and *LCK*.

2.4.3 Schizophrenia

Schizophrenia is a chronic and severe mental disorder that affects social behavior, beliefs, and thinking for a person (Liou et al. 2012; Lin and Tsai 2016b). Takahashi et al. (2010) used an ANN algorithm to assess whether the gene expression signature in whole blood consists of sufficient information to segregate patients with schizophrenia. They singled out 14 probes as predictors for differential diagnosis of schizophrenia with the quality filtering and stepwise forward selection methods. The ANN model was then constructed with the selected probes, and it carried out 91.2% accuracy in the training data and 87.9% accuracy in the testing data (Takahashi et al. 2010).

2.5 Perspectives

Several limitations exist with respect to the aforementioned diagnostic prediction studies. Firstly, studies with limited sample size did not warrant well-defined results (Lin and Lane 2015). Secondly, researchers often investigate all of the available algorithms because the only sure way to find the very best algorithm is to try all of them (Lin and Tsai 2016c; Lin and Lane 2017).

Besides ANN models, there are a variety of machine learning tools we can use to analyze gene expression profiling data in diagnostic prediction studies. Some of the best-known machine learning and predictive algorithms encompass naive Bayes (Domingos and Pazzani 1997), C4.5 decision tree (Quinlan 1993), ANNs (Lin et al. 2006; Kung and Hwang 1998; Bishop 1995; Rumelhart et al. 1996), support vector machine (SVM) (Vapnik 1995), k-means (Lloyd 1982), k-nearest neighbors (kNN) (Altman 1992), and regression (Friedman et al. 2010; Zou and Hastie 2005). These classifiers are usually adopted for comparison owing to the fact that these methods possess a diversity of capacities with distinctively representational models, such as probabilistic models for naive Bayes, decision tree models for the C4.5 algorithm, and regression models for SVM (Hewett and Kijsanayothin 2008).

For instance, Table 2.2 summarizes the relevant diagnostic prediction studies by using gene expression profile and a variety of machine learning models. In order to

Table 2.2 Diagnostic prediction studies of gene expression profiling for various diseases and treatments of significance using a variety of machine learning algorithms

Disease/treatment	Results	References
Breast cancer	Identified 21 most-associated genes	Chou et al. (2013)
Colorectal tumors	Achieved 99% accuracy	Chu et al. (2014)
Colon cancer	Achieved 91% accuracy	Hu et al. (2015)

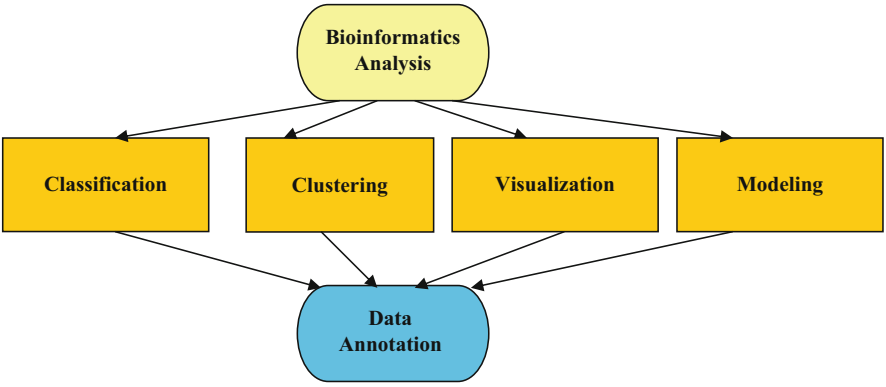

Fig. 2.2 Bioinformatics tools for analyzing and visualizing the relationship between gene expression data and human diseases

predict breast cancer recurrence, Chou et al. (2013) employed gene expression profiling of breast cancer survivability and three methods including logistic regression, decision tree, and ANN models. Their analysis indicated 21 genes closely relevant to breast cancer recurrence (Chou et al. 2013). In addition, in order to screen for the variations in gene expression between colorectal tumors and normal mucosa tissues, Chu et al. (2014) employed four methods, including ANN, prediction analysis of microarray, classification and regression trees (CART), and C5.0 algorithms. Colorectal cancer is a cancer that starts in the colon or rectum. Chu et al. (2014) adopted a two-tier genetic screen to reduce the number of candidate significant genes, and the ANN model achieved the best classification performance, with an average 99% test accuracy. Moreover, based on gene expression data, Hu et al. (2015) classified colon cancer subjects treated with elective standard oncological resection into two groups such as relapse and no relapse by using ANN, Kohonen neural network, and SVM models. The Kohonen neural network model achieved the best classification performance, with an average 91% test accuracy (Hu et al. 2015).

In future work, a bioinformatics pipeline can be used to provide a thorough evaluation and validate whether the findings are replicated in diagnostic prediction studies. Figure 2.2 shows a bioinformatics pipeline for analyzing and visualizing gene expression profiling data in diagnostic prediction studies. Additionally, we could investigate potential biomarkers by using a custom data mining pipeline so that genetic networks would be illustrated at the genome level.

Acknowledgments The authors extend their sincere thanks to Vita Genomics, Inc. and SBIR grants (S099000280249-154) from the Department of Economic Affairs in Taiwan for funding this research.

References

Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46:175–185

Bishop CM (1995) Neural networks for pattern recognition. Clarendon Press, Oxford

Chen YC, Chang YC, Ke WC, Chiu HW (2015) Cancer adjuvant chemotherapy strategic classification by artificial neural network with gene expression data: an example for non-small cell lung cancer. J Biomed Inform 56:1–7. https://doi.org/10.1016/j.jbi.2015.05.006

Chou HL, Yao CT, Su SL, Lee CY, Hu KY, Terng HJ, Shih YW, Chang YT, Lu YF, Chang CW, Wahlqvist ML, Wetter T, Chu CM (2013) Gene expression profiling of breast cancer survivability by pooled cDNA microarray analysis using logistic regression, artificial neural networks and decision trees. BMC Bioinform 14:100. https://doi.org/10.1186/1471-2105-14-100

Chu CM, Yao CT, Chang YT, Chou HL, Chou YC, Chen KH, Terng HJ, Huang CS, Lee CC, Su SL, Liu YC, Lin FG, Wetter T, Chang CW (2014) Gene expression profiling of colorectal tumors and normal mucosa by microarrays meta-analysis using prediction analysis of microarray, artificial neural network, classification, and regression trees. Dis Markers 2014:634123. https://doi.org/10.1155/2014/634123

Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 29:103-137

Dwivedi Y (2014) Emerging role of microRNAs in major depressive disorder: diagnosis and therapeutic implications. Dialogues Clin Neurosci 16:43–61

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22

Hewett R, Kijsanayothin P (2008) Tumor classification ranking from microarray data. BMC Genomics 9:S21. https://doi.org/10.1186/1471-2164-9-S2-S21

Hu HP, Niu ZJ, Bai YP, Tan XH (2015) Cancer classification based on gene expression using neural networks. Genet Mol Res 14:17605–17611. https://doi.org/10.4238/2015.December.21.33

Kononenko I (2001) Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 23:89–109

Kung SY, Hwang JN (1998) Neural networks for intelligent multimedia processing. Proc IEEE 86:1244–1272

Lancashire LJ, Powe DG, Reis-Filho JS, Rakha E, Lemetre C, Weigelt B, Abdel-Fatah TM, Green AR, Mukta R, Blamey R, Paish EC, Rees RC, Ellis IO, Ball GR (2010) A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. Breast Cancer Res Treat 120:83–93. https://doi.org/10.1007/s10549-009-0378-1

Landset S, Khoshgoftaar TM, Richter AN, Hasanin T (2015) A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J Big Data 2:24

Lane HY, Tsai GE, Lin E (2012) Assessing gene-gene interactions in pharmacogenomics. Mol Diagn Ther 16:15–27. https://doi.org/10.2165/11597270-000000000-00000

Lin E (2012) Novel drug therapies and diagnostics for personalized medicine and nanomedicine in genome science, nanoscience, and molecular engineering. Pharm Regul Aff Open Access 1:e116

Lin E, Lane HY (2015) Genome-wide association studies in pharmacogenomics of antidepressants. Pharmacogenomics 16:555–566. https://doi.org/10.2217/pgs.15.5