Vijai Singh
Ajay Kumar *Editors*

Advances in Bioinformatics

Springer

Advances in Bioinformatics

Vijai Singh • Ajay Kumar
Editors

Advances in Bioinformatics

Springer

Editors

Vijai Singh
Department of Biosciences
Indrashil University
Mehsana, Gujarat, India

Ajay Kumar
Biotechnology
Rama University
Kanpur, Uttar Pradesh, India

ISBN 978-981-33-6190-4 ISBN 978-981-33-6191-1 (eBook)
https://doi.org/10.1007/978-981-33-6191-1

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

Foreword

I am happy to write the foreword to the *Advances in Bioinformatics*, a timely volume on the rapidly growing field.

Bioinformatics was born when biological data was generated in numbers that necessitated its management mainly in terms of storage, analysis, output, and communication. To begin with, bioinformatics had to mainly deal with protein structure studies. Over time, there has been a surge of data, from parts to pathways to multicellular contexts. Databases have become the norm in biology and modeling has made significant inroads into the experimental labs.

This book covers basic and advanced aspects of bioinformatics in terms of tools, data mining, analytics, computational evolutionary biology, computational vaccine, and drug design. It also covers proteomics, metabolomics, DNA sequencing and NGS to genome analysis, biological computation, neural network analysis, big data analysis, soft computing, and artificial intelligence. All chapters are written by eminent scientists who have well-established research in bioinformatics.

I am delighted to observe the valuable efforts of Dr. Vijai Singh and Dr. Ajay Kumar, who have worked hard to bring out an excellent volume with the support of Springer Nature.

This book offers a valuable source of information for not only beginners in bioinformatics but also for students, researchers, scientists, clinicians, practitioners, policymakers, stakeholders who are interested in harnessing the potential of bioinformatics from fundamental science to applications.

School of Biotechnology,
Jawaharlal Nehru University,
New Delhi, India

Pawan K. Dhar

Preface

Bioinformatics is a rapidly growing area of biology and has gained significant scientific and public attention. It is currently used in all domains of biological sciences research and has accelerated the research work. It combines principles of biology, computer science, information technology, mathematics, and statistics to analyze and interpret biological data. This book covers the momentousness of bioinformatics assistance and knowledge, thereby highlighting its role in the advancement of modern science. It plays major roles in the development of biological science, without which the researchers' community cannot extricate, assess, or analyze any type of large-scale paradigm whether it is genomics, proteomics, or transcriptomics which ensures to tackle biological problems.

A wide range of topics from basic to advanced level of bioinformatics have been covered. This book covers introduction, tools, data mining and analysis, computational evolutionary biology, protein analysis, computational vaccine, and drug design. It also covers computational genomics, proteomics, metabolomics, DNA sequencing and NGS, microRNA, gene to genome analysis, biological computation, neural network analysis, artificial intelligence, big data analysis, soft computing, and many other relevant topics.

We believe that this book covers great range of topic in different aspects of bioinformatics. This book offers an excellent and informative text on bioinformatics, benefitted by simple to understand and easy to read format. This book uses a rich literary text of excellent depth, clarity, and coverage. It highlights a number of aspects of bioinformatics in a way that can help future investigators, researchers, students, and stakeholders to perform their research with greater ease. This book provides an excellent basis from which scientific knowledge can grow, widen, and accelerate bioinformatics research in many areas.

Mehsana, Gujarat, India Vijai Singh
Kanpur, Uttar Pradesh, India Ajay Kumar

Acknowledgement

I would like to express my sincere gratitude and deep appreciation to Dr. J.S. Yadav, Director (Research), Indrashil University, India, for extending his outstanding support and motivation to complete this book. I would like to give many thanks to co-editor Dr. Ajay Kumar Singh of this book who gave me outstanding personal and professional support as well as inspiration to finish this book.

I am delighted to thank all the authors for their excellent contributions to this book. I would like to thank Dr. Bhavik Sawhney (Associate Editor—Biomedicine) and Ms. Veena Perumal (Production Editor) from Springer for their excellent management of this project.

I would like to give thank Prof. Chaitanya G. Joshi, Prof. Rakesh Rawal, Prof. Bharat Maitreya. Prof. Pawan K. Dhar, Dr. Poonam Bhargava, Dr. Madhvi Joshi, Prof. S. R. Dave, Dr. Bhabatosh Das, Dr. Pablo Carbonell, Dr. Javin P. Oza, Dr. Rupesh Maurya, Dr. Shams Yazdani, Dr. Indra Mani, Dr. Satya Prakash, Dr. Vimal C. Pandey, Dr. Dinh Toi Chu, Dr. Suresh Ramakrishna, Dr. Prof. Nisheeth Agrawal, and those whose names do not feature here but have directly or indirectly contributed in shaping this project. I greatly appreciate the support of my students Mr. Nisarg Gohil, Mr. Khushal Khambhati, and Ms. Gargi Bhattacharjee, whose discussion and comments helped to shape this book.

I wish to express my gratitude to my beloved wife Pritee Singh for her endless support, patience, and inspiration. Lots of affection for my kids Aaradhya and Ayush who missed me during this project. I would like to warmly thank the faculty and staff of Indrashil University for providing a great working environment.

I am aware that even despite our best efforts, the first version always comes with some error that may have crept in the compilation. I would be delighted to receive feedback from readers to further improve the future book. Last but not least, my sincere thanks to GOD for his supreme POWER for endowing me to live with joy and victory in the shape of this book.

Vijai Singh

This book is an outcome of utter studies and literature survey. I express my profound gratitude to family, friends, and colleagues who always have good words for me inculcating a source of strength and inspiration for taking this venture.

I am especially grateful to Dr. Prateek Singh (Director, Rama University Kanpur, India), Dr. Pranav Singh (Director PR, Rama University), Dr. Hari Om Sharan (Dean, Rama University), and Dr. Sandeep Shukla (Deputy Registrar, Rama University) who never stopped me to take this challenge for developing ideas and supported me to develop a strong foundation.

I also take this opportunity to thank my worthy and learned colleagues in the Department of Biotechnology, Faculty of Engineering and Technology, Rama University, whose knowledge and experience have eased and developed the confidence to take up my work and finally taking to finish the note.

I would also like to mention sincere cooperation from Er. Monisa Anwer and Er. Fariya Khan, who always extended their helping hands for discussions and cooperation. And last but not least, I feel deeply and highly obliged to my beloved wife Shraddha and my children Nishit and Anshika who not only motivated me to author this book but also showed patience when deprived them of my much-needed attention during this course of time so that my book could see the light of the day. It would not be out of place to seek the blessings of my elder brother Er. Prem Chandra who always had been a guiding figure in my passion.

I am also thankful to Almighty God who has given me the wisdom to edit this book. Finally, I dedicate this book to my parents for their countless blessings to accomplish the target.

Ajay Kumar

Contents

About the Editors

Vijai Singh is an Associate Professor and Dean (Research and Innovation) at School of Sciences, Indrashil University, Rajpur, Mehsana, Gujarat, India. He was an Assistant Professor in the Department of Biological Sciences and Biotechnology at the Institute of Advanced Research, Gandhinagar, India and also an Assistant Professor in the Department of Biotechnology at the Invertis University, Bareilly, India. Prior to that, he was a Postdoctoral Fellow in the Synthetic Biology Group at the Institute of Systems and Synthetic Biology, Paris, France and School of Energy and Chemical Engineering at the Ulsan National Institute of Science and Technology, Ulsan, South Korea. He received his Ph.D. in Biotechnology (2009) from the National Bureau of Fish Genetic Resources, Uttar Pradesh Technical University, Lucknow, India with a research focus on the development of molecular and immunoassays for diagnosis of *Aeromonas hydrophila*. His research interests are focused on building novel biosynthetic pathways for production of medically and industrially important biomolecules. Additionally, his laboratory is working on CRISPR-Cas9 tools for genome editing. He has more than 8 years of research and teaching experience in synthetic biology, metabolic engineering, bioinformatics, microbiology, and industrial microbiology. He has published 76 articles, 31 chapters, 9 books and 3 patents. He serves as an associate editor, editorial board member, and reviewer of several peer-reviewed journals. He is also a member of the Board of Study and Academic Council of Indrashil University and is the Member Secretary of the Institutional Biosafety Committee (IBSC) at the same university.

Ajay Kumar is the Professor and Head of the Department of Biotechnology at the Faculty of Engineering and Technology at Rama University, Kanpur, Uttar Pradesh. He has 15 years of post-PhD experience in academics and research. He has held several key positions in different universities and engineering institutes. He received M. Tech. (Biotechnology in 2001) at Institute of Engineering and Technology, Lucknow (UP) and PhD (2006) from the ICAR—Central Institute for Research on Goats, Mathura. His research interests include computational vaccine and drug development, genomics and proteomics, and fermentation technology. He has published more than 90 research papers in international/national Journals. He authored many books and chapters. He serves as an editor and reviewer of several peer-reviewed journals. He is a member of professional body such as the International Association of Engineers (IAENG) and INSA. He has rendered consultancy services in the field of vaccine research.

An Introduction and Applications of Bioinformatics

1

Henny Patel, Dhruti Bhatt, Shreya Shakhreliya, Navya L. Lam, Rupesh Maurya, and Vijai Singh

Abstract

In the past few decades, bioinformatics has been extensively explored in many areas of biological sciences. It combines the principles of biology, computer science, mathematics, physics, and statistics to analyze and interpret biological data. It uses computation power, algorithm and software for extracting knowledge from biological data for analysis, prediction, imaging, and visualization purpose. In this chapter, we highlight recent developments in this field and how the potential of bioinformatics has harnessed in multiple disciplines.

Keywords

Bioinformatics · Data analysis · Algorithm · Drug discovery · Data · Modelling

1.1 Introduction

A constant need has arisen to develop and establish a technology to gain knowledge and to meliorate human and animal life. Over the past few decades, development of number of breakthrough technologies including DNA sequencing (Sanger et al. 1977; Gohil et al. 2019), genome sequencing, proteomics, genome annotation

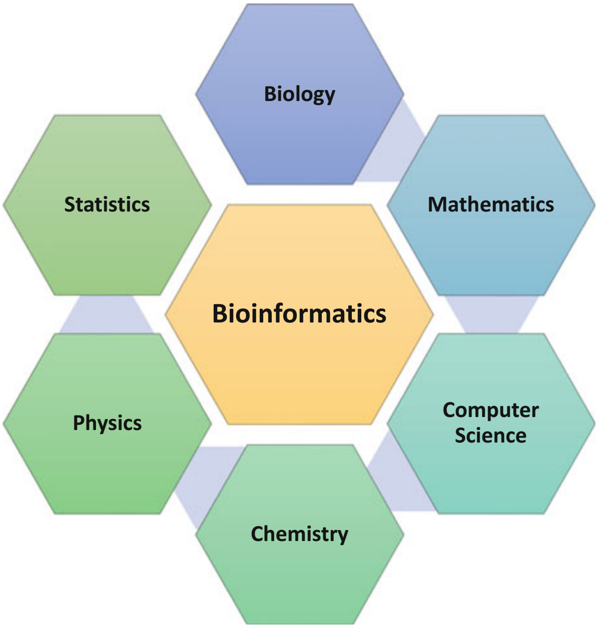
H. Patel · S. Shakhreliya · R. Maurya · V. Singh (✉)
Department of Biosciences, School of Science, Indrashil University, Rajpur, Mehsana, Gujarat, India
e-mail: vijai.singh@indrashiluniversity.edu.in

D. Bhatt
School of Science, The University of Auckland, Auckland, New Zealand

N. L. Lam
The J. David Gladstone Institutes, San Francisco, CA, USA

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_1

Fig. 1.1 Biological combines the principles of different disciplines



and assembly (Gibson et al. 2010), protein sequencing, genome synthesis (Gibson et al. 2010), genome editing (Cong et al. 2013; Singh et al. 2017, 2018a; Gohil et al., 2021; Bhattacharjee et al. 2020), and many such technologies has been made immensely to enhance the quality of research. The advent of bioinformatics has further made it possible to accelerate and improve wet laboratory experiments greatly. Bioinformatics is a combination of biology, computer science, mathematics, statistics, physics, and engineering principles that helps in analyzing and interpreting biological data (especially large and complex) with the assistance of software tools (Fig. 1.1).

Bioinformatics involves biological investigation that utilizes computer programming and algorithms as part of its techniques. Additionally, in the field of genomics, there is repetitive use of particular analysis "pipelines." Scientists commonly use bioinformatics tools for identifying candidate genes and single-nucleotide polymorphisms (SNPs) (Cargill et al. 1999; Bhattacharjee et al. 2019) for studying characteristics of population, different adaptations, genetic disorders, or desirable properties (mainly agricultural species). Bioinformatics also contributes in understanding of organizational principles of nucleic acids and protein sequences, also called proteomics (Graves and Haystead 2002).

In the several areas of biology, bioinformatics has already proven its significance. In experimental molecular biology, image and signal processing (bioinformatics methodologies) aids in extracting significant results from large amounts of raw data. Bioinformatics has its importance in the field of genetics where it assists in sequencing and annotating genomes and their visualized mutations. It contributes to text mining of biological writing and supports in developing biological and gene ontologies for organizing and quizzing biological data. Gene and protein expression and regulation is now possible due to development of bioinformatics tools. Also, it assists in comparison, analysis, and interpretation of genetic and genomic data (especially studying evolutionary characteristics of molecular biology). Combinatorially, it supports in analyzing and classifying biological pathways and networks that have significance in systems biology. It helps in simulation and modelling of DNA, RNA, proteins, and bimolecular interactions in the field of structural biology. This chapter includes basics of bioinformatics and its use in different fields that overall contributes in clear understanding of biological sciences.

1.2 Applications of Bioinformatics

In the past decade, bioinformatics and its applications have widely used in number of areas for better understanding and have helped to accelerate research in many areas including DNA sequencing, gene and genome analysis, evolutionary biology, immunoinformatics, gene expression, proteomics, and many others (Fig. 1.2).

1.2.1 DNA Sequence and Analysis

Fred Sanger discovered the bacteriophage ΦX174 and analyzed it (Sanger et al. 1977). This was a major discovery and subsequently number of genes were sequenced. DNA sequences were used for identification of genes and organisms. It

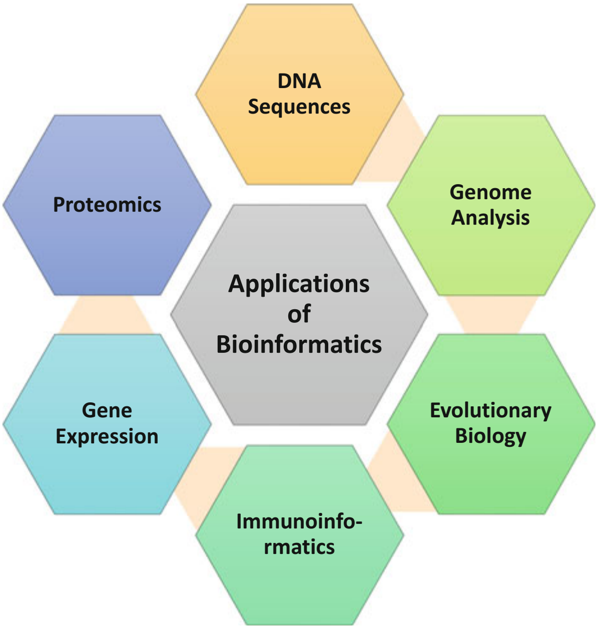


Fig. 1.2 Applications of bioinformatics in various disciplines

was later started being used for construction of evolutionary relationship among species. DNA sequences are also analyzed by a number of tools and software. In 2008, a software program called Basic Local Alignment Search Tool (BLAST) was used to scan sequences—from more than 260,000 species, including more than 190 billion nucleotides (Benson et al. 2008). Still a major part of biodiversity is completely left unexplored. New DNA sequencing approaches are being discovered for accelerating research.

For analyzing sequences, one can obtain data from Genbank. It is a data storage bank that has publicly available DNA sequences. Similarity, DNA sequences can be checked and verified using BLAST program (Singh et al. 2018b). Development of algorithms has assisted with base calling for several experimental approaches to DNA sequencing.

1.2.2 Genome Sequencing

Many different techniques of DNA sequencing produce short, small fragments of unique sequences that need to be specifically combined to achieve almost complete gene or genome action sequences. Since the advent of the process, *Haemophilus influenzae* was the first ever bacterial genome to be sequenced by the Institute for Genomic Research (TIGR, USA) producing sequences of several thousand small fragments of DNA depending on the sequencing technology, ranging from 35 to 900 nucleotides (Fleischmann et al. 1995). The final ends of each of these overlapping fragments can be used to replicate the whole genome when exactly and regularly aligned around genome assembly program.

Approach of merging fragments can be very difficult with substantially larger genomes, and in such conditions shotgun sequence of events yield sequence data quickly and reliably (Ekblom and Wolf 2014). Human genome has also been sequenced and analyzed which has immensely helped to design personalized medicines and has aided identification of number of drug targets (Venter et al. 2001). The sequences were big and required several days of CPU time to combine fragments on large-memory multiprocessor computers. Practically all genomes are sequenced nowadays, shotgun sequencing is the process of preference and sequence alignment optimizations are really an essential aspect of bioinformatics science.

1.2.3 Genome Annotation and Analysis

After the complete genome was sequenced, a major challenge was its annotation and analysis. Annotation tends to be a method for the recognition of certain genes and other biological features throughout the DNA sequence in the genomics context. It is important to automate this process as most genomes are massive to actually annotate first hand, not to mention the need to continuously annotate and provide as many genomes as technically practicable, but the sequence rate sometimes stops during operations and it somehow tends to be a big bottleneck. The fact that specific genes in us have recognizable beginning and stopping regions makes annotation relatively possible, because the exact sequence present in these autonomous regions may differ greatly across all genes.

In 1995, the entire team eventually published the first thorough explanation of a systematic genome annotation method at TIGR, which comprised of the first full genomic sequence and analysis of the whole genome of a fully free microorganism *H. influenzae* (Fleischmann et al. 1995). Instead of making original practical and available allocations, Owen White developed a software framework to classify the gene character encoding of almost all proteins, transfer RNAs, ribosomal RNAs (for this and other sites). Most of the current genome annotation systems work properly, but the resources available for genomic DNA analysis have been constantly evolving and developing, including the GeneMark program, which is trained and used to classify protein coding genes in *H. influenzae*.

A new effort set up by the National Human Genome Research Institute in the United States has arisen to fulfill the goals of the Human Genome Project, which has remained to be followed since its completion in 2003. Encyclopedia of DNA Elements (ENCODE) project, a collaborative data collection employs innovative DNA sequencing methods and genomic tiling systems for the functional components of the human genome. This technology at the same time produces vast volumes of data at a significantly reasonable cost per base at precisely the same precision template.

In the last few decades, knowledge regarding earth biodiversity has significantly increased. Due to hard work of taxonomists and technology, respective data of living species have tremendously increased. Earth biogenome project targeting sequence catalogs and characterizes the genomic data of all eukaryotic life on earth (Lewin et al. 2018). The 1000 genome project (Buchanan et al. 2012), 10,000 plant genome projects (Li and Harkess 2018), sheep genome (International Sheep Genomics Consortium 2010), *Bostaurus* genome (Elsik et al. 2016) are informatics resources for farming, drug, and health industries.

Several databases such as Ensemble genome browser (Stalker et al. 2004) for quality annotation of gene sequences, the CATH Database for protein structure and function relationships (Orengo et al. 1999), National Center for Biotechnology Information (NCBI) (Sherry et al. 2001; Barrett et al. 2012; Pruitt et al. 2005), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2017; Gohil et al. 2017) are the major resources of genomic and systemic functional information as well as for helping to better understand the molecular processes.

1.3 Computational Evolutionary Biology

The field of computational evolutionary biology is the unification of evolution biology and informatics that enables researchers to continue increasing understanding of evolution in natural and artificial systems. Evolution is an iterated, population-based, heritable variation, selection and mutation. Evolutionary computational biology is now highly popular in the chemical, medical industry and bioinformatics (Mitchell and Taylor 1999).

This informatics enables the life scientists to

- Differentiate whole genomic changes that allow to study diverse evolutionary events, for example gene amplification (Andersson and Hughes 2009) factors responsible for bacterial speciation and lateral/horizontal gene transfer.
- Map the evolution changes in a great number of species by tracing the change in genomics, instead of endeavor the taxonomy, biochemical and physiological changes in organism.
- Over time, it develops detailed computerized models of population genetics to insure the end result of the system and to store and track the statistics data of a number of different organisms (Carvajal-Rodríguez 2010).

Computational biology and sequencing technologies led to comprehensive advancement in understanding phylogeny (Pagel 2006). Gene regulation has a role in speciation and adaptation (Romero et al. 2012) of organisms. Phylogenetic networks (Huson and Bryant 2006), biodiversity collections (Graham et al. 2004), anatomy illustrations (Ghosh 2015) all play an essential role in evolutionary informatics. Phylogenetic networks can be used for understanding systematics. Evolutionary informatics can be used for weighting, partitioning, and combining characters (Chippindale and Wiens 1994), use of polymorphic characters (Wiens 1995) and confidence intervals for regression equations in phylogeny (Garland and Ives 2000). Important components of algorithms for evolutionary programming are representation, parent selection (Eiben et al. 1999), crossover operators (Spears 1995), mutation operators (Chellapilla 1998), survival selection (Eiben and Smith 2015), and termination condition. Evolutionary algorithms are influenced by procreation or breeding, change in a DNA sequence, recombination and selection (Mani and Mani 2017).

Tools such as molecular evolutionary genetics analysis (MEGA6.0) (Tamura et al. 2013) for construction of phylogenetic, parsimony (PAUP) (Swofford 1993), ClustalW (Thompson et al. 1994) for phylogenetic analysis, prediction confidence intervals for regression equations (Garland and Ives 2000) have been used widely. Apart from that, MrBayes, MAFFT, SweepFinder, JAVA (BEAST), etc. (Darriba et al. 2018) are also important software that are used for evolutionary informatics. Software such as KEEL (Alcalá-Fdez et al. 2009), Python are frequently used in evolutionary studies. In order to precisely understand and analyze molecular mechanisms and function, more advanced tools based on artificial intelligence, machine learning and deep learning development should be stressed upon and implemented.

1.4 Comparative Genomics

In the 1970–1980s, comparing the viral genomes became the starting point in the comparative genomics (Koonin and Galperin 2002). The basics of comparative genomics is to compare genomic features such as nucleotide sequences, genome size, genes, orthology analysis, regulation of gene expression and other genomic structural changes of two different organisms (Xia 2013). Similarity is the key point in the comparative genomics (Primrose and Twyman 2009). The curve of DNA evolution acts as an evolutionary proceeding at a different organizational level. At first level it acts on species nucleotides through point mutations and at second level, rapid speciation occurs due to chromosomal duplication, genomic transfer, mutation, transposition that lead to genome processes of endobiont, polyploidy, hybridization and many (Brown 2002; Chen and Ni 2006).

Tools used for comparing genome are UCSC browser (large scale sequencing references and assembling genome draft), Ensembl (eukaryotic and vertebrates species genome databases), Map view (data sequencing and genome mapping), VISTA (visualizing results based on DNA alignments), and BlueJay Genome Browser (visualization for multi-scale genomic changes) (Soh et al. 2012). With mathematical models, statistical (Bayesian analysis) and algorithms (Markov chain Monte Carlo algorithms), developers troubleshoot the complexity of changing or evolving genomic sequences. Major studies are contingent upon removal of DNA sequences homology to direct protein family's sequences (Carter et al. 2002).

1.4.1 Pan-genomics

Pangenome is the complete set of genes of all strains of a species. Tettelin et al. (2005) introduced the concept of pan-genomics that was eventually established in bioinformatics. However, at the beginning, it involved strains of species that have close relations. It was divided into two parts: first, the core genome, which is a set of genes unique for every genome under study (housekeeping genes) and, second is the dispensable or flexible genome that does not show its presence in all but only in one or several strains. Characterization of bacterial species' pan-genome is possible by using a bioinformatics tool such as Bacterial Pan Genome Analysis (BPGA). BPGA is an ultra-fast software package that helps comprehensive pan-genome analysis of microorganisms (Chaudhari et al. 2016).

1.4.2 Genetics of Disease

Studying complex diseases has become easy since the introduction of next-generation sequencing. Nowadays, one can easily access adequate sequence data for mapping the genes of infertility (Aston 2014), breast cancer (Véron et al. 2014), or Alzheimer's disease (Tosto and Reitz 2013). Studies of genome-wide correlation are a valuable method to identify the mutations liable behind certain complicated diseases (Londin et al. 2013). With the help of these studies, thousands of associated DNA variants can be researched that have relations with alike diseases and traits (Hindorff et al. 2009). In addition, one of the most important applications is the potential for genes to be used in prognosis, diagnosis, or accurate treatment. Several researches have addressed positive methods of selecting the genes to be used and the issues as well as disadvantages to use genes to determine the existence or prognosis of diseases (Bejar 2014).

1.4.3 Analysis of Gene Mutations in Cancer

In diseases such as cancer, there is a complex rearrangement in genomes of unhealthy cells. In order to locate completely undiscovered point mutations in several genes involved in cancer, massive sequencing attempts are required. In order to handle more sequence generated data, bioinformaticians have started to generate advanced automated tools and develop new algorithms and software to correlate the sequencing findings to the increasing array of human genome sequences and germline polymorphisms. For identification of chromosomal gains/losses and single-nucleotide polymorphism (SNP) sets to pinpoint recognized point mutations, modern physical identification methods such as oligonucleotide microarrays are used. These detection techniques concurrently test multiple hundreds to thousands sites in the genome and produce terabytes of data per study while employed in high-throughput measuring of thousands of samples. There are more opportunities available for bioinformaticians because there is availability of large amounts of new data. Available data is frequently known to possess substantial variability or noise. To address this, a hidden Markov model and methods of change-point analysis have been introduced to predict actual variations in the number of copies (Morris and Baladandayuthapani 2017).

Cancer genomics can shift dramatically with the advancements in next-generation sequencing technology. Bioinformaticians can afford and sequence multiple cancer genomes in very less time by taking help from these methodologies and software. The study of cancer-driven mutations in the genome could provide a more versatile method for identifying cancer forms (Hye-Jung et al. 2014).

1.5 Bioinformatics in Gene and Protein Expression Analysis

1.5.1 Analysis of Gene Expression

The gene expression is regulated by measuring mRNA level with different noise-prone techniques such as DNA microarrays or biochip—assembly of DNA microscopic spots substrate to a solid surface for hybridization of two DNA strands; expressed cDNA sequence tag sequencing (EST)—short sub-sequence of cloned cDNA; serial analysis of gene expression tag sequencing (SAGE)—output of analysis is to list out short tags sequence and the number of time it occurs; massively parallel signature sequencing (MPSS)—to identify and quantify mRNA transcripts; and RNA-seq or whole transcriptome shotgun sequencing(WTSS). The next-generation sequencing is used to identify the presence and quantity of RNA in the sample for analyzing change in cellular transcriptome and other complex *in situ* hybridization (Wang et al. 2009; Kukurba and Montgomery 2015).

In the field of computational biology, statistical tools are used to differentiate signal and noise for output of gene expression (Grau et al. 2006). These techniques are used to identify the change in molecular impression of a disease and it has the potential utility to lead to drug discovery for clinical treatment (Bai et al. 2013). These tools are used to identify the amplification in genes in patients by comparing techniques such as microarrays, which differentiate the information of non-cancerous epithelial cells from details of cancerous cells in order to regulate proper gene transcripts in tumor cells.

1.5.2 Analysis of Protein Expression

A description of proteins present in an organism can be given by protein microarray analysis and high-throughput (HT) mass spectrometry (MS). Bioinformatics seems to be very involved in understanding the importance of the protein microarray and HT-MS material of the suggested technique presents almost the same challenge as the microarray for targeting mRNA, and also includes the difficulty in relating vast volumes of huge protein sequence library data to the expected weights. Through association proteomics seen as satellite information focussed on immunocytochemistry or tissue nanomaterials, cellular protein specialization can be performed in a type of tissue background (Hall et al. 2007).

1.5.3 Analysis of Gene Regulation

Bioinformatics also explores the analysis of gene regulation. Regulation of gene expression includes activation or repression of various cellular mechanisms by extracellular signals such as hormones or metabolites or by concentration gradient for the formation of certain proteins at a particular time. It can be controlled by factor affecting the gene activation. It is important to identify and understand the sequence motif (amino acid sequence) in DNA around the coding sequence by promoter analysis. These motif sequences affect specific transcribing region that forms mRNA. Gene expression is also regulated by enhancer (cis-acting) region that is bound by activator proteins and influences 3-D chromatin looping for interaction of enhancer and target gene (Pennacchio et al. 2013). With bioinformatics technique, chromosome conformation capture experiment is used to analyze the interaction between enhancer sequence and target gene sequences.

Nevertheless, gene regulation is concluded by the gene expression data to compare microarray gene data of various species to form hypotheses of gene complexity. Under *in vivo* condition, microarray expression data and cluster analysis are used to identify regulatory regions that are enclosed through transcription factors for analysis and understanding dynamics of gene regulation (Fogel and Corne 2002). For example, in unicellular organisms, one can compare the cell division cycle with various stress factors (starvation, temperature, heat shock, and many more). For the determination of co-expressed gene, clustering algorithms are used with expression data. Various clustering algorithms are involved in genome clustering such as k-means clustering, hierarchical cluster analysis (HCA), self-organizing feature map (SOFM), and consensus clustering (cluster ensembles).

1.6 Structural Bioinformatics

Structural bioinformatics is the field of bioinformatics that analyzes and identifies the 3-D structure of protein, RNA, and DNA. The major emphasis of structural bioinformatics is to form a new technique that settles with macromolecules data to resolve the issue of researchers and to create better understanding of biological molecules. It mainly inscribes structural interactions among space coordinates. Prediction of structure of protein is the foremost application in bioinformatics (Gu and Bourne 2011). Some structural databases of the protein structure are as follows: protein data bank (PDB)—Macromolecular Structures Resource Group, nucleic acid database—Nucleic Acid Database (NDB), critical assessment of protein structure prediction (CASP)—Prediction Center, protCID—Protein Common Interface Database, electron density server (EDS)—EDS, Uppsala Electron Density Server, and some others (Luscombe et al. 2001). Understanding of structure helps to understand protein function and disease conditions (Sudha et al. 2014).

In the field of bioinformatics, homology is used for determining the role of specific genes and which sequences of protein is vital for structural formation and interaction with other molecules, such as, if sequence of one of the gene “A” is known and is homologous to other gene “B”, then it can be assumed that gene A can share function of gene B. Homology modelling technique uses this data to assume the protein structure by known homologous protein. An example of this is of the homology between hemoglobin in humans and leghemoglobin in legumes which have the same function, i.e., to carry oxygen to cells. Both have different sequences of amino acids but their structure is same, which therefore mirrors their same function and common ancestor. Another application of structural bioinformatics is target selection, trial mapping, or analysis of X-ray crystallography (Ilari and Savino 2008) and NMR spectroscopy data (Sugimoto et al. 2012), and for virtual screening of models in drug discovery (Chou 2004).

1.7 Immunoinformatics for Vaccine Design

A field of science that studies both immunogenetics and immunology data with the help of bioinformatics tools is known as immunoinformatics. This helps to identify and analyze epitopes (antigenic part of protein) in protein for development of antigen or vaccine candidates. A number of online tools are available for identification of B-cell and T-cell epitopes. BepiPred-2.0 (http://www.cbs.dtu.dk/services/BepiPred/) (Jespersen et al. 2017) and LBtope (http://crdd.osdd.net/raghava//lbtope/) (Singh et al. 2013) were developed for identification of B-cell epitopes from highest antigenic region of protein based on random forest algorithms through epitopes annotated from antigen-antibody protein structure. T-cell epitopes are also important for development of vaccine candidates and can be predicted using NetCTL 1.2 server (http://www.cbs.dtu.dk/services/NetCTL/), ProPred (Singh and Raghava 2001) and ProPred 1 (Singh and Raghava 2003). Several monovalent and multivalent epitopes have been predicted and more *in vitro* validation is required to experimentally prove the efficiency of predicted epitopes to bring vaccines into market.

1.8 Conclusions and Future Perspective

Bioinformatics is a rapidly growing field and is also contributing in accelerating many associated fields of biological sciences. It has shown tremendous potential and its ability can be harnessed for better understanding of molecular mechanisms, identifying a new gene/protein, small RNA, generating a 3-D model, screening of drug, predicting vaccine etc. Due to rapid development in tools, software, genomics, proteomics, metabolomics data, it is now possible to predict functions of biological molecules more accurately. Researchers are now able to predict *in silico* personalized medicine and its uses for accurately treating patients. With more recent developments in artificial intelligence, machine learning, and deep learning, this ability can be further enhanced in the near future to predict better health.

Acknowledgements The authors would like to give thanks to Ms. Gargi Bhattacharjee and Mr. Nisarg Gohil for support and proofreading of this chapter.

References

Alcalá-Fdez J, Sanchez L, Garcia S et al (2009) KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Comput 13:307–318

Andersson DI, Hughes D (2009) Gene amplification and adaptive evolution in bacteria. Annu Rev Genet 43:167–195

Aston KI (2014) Genetic susceptibility to male infertility: news from genome-wide association studies. Andrology 2:315–321

Bai JP, Alekseyenko AV, Statnikov A et al (2013) Strategic applications of gene expression: from drug discovery/development to bedside. AAPS J 15:427–437

Barrett T, Wilhite SE, Ledoux P et al (2012) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995

Bejar R (2014) Clinical and genetic predictors of prognosis in myelodysplastic syndromes. Haematologica 99:956–964

Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2008) GenBank. Nucleic Acids Res 36:D25–D30

Bhattacharjee G, Khambhati K, Gohil N et al (2019) Exploiting the potential of DNA fingerprinting in forensic science. In: Shukla RK, Pandya A (eds) Introduction of forensic nanotechnology as future armour. Nova Publishers Inc., New York, pp 145–185

Bhattacharjee G, Mani I, Gohil N et al (2020) CRISPR technology for genome editing. In: Faintuch J, Faintuch S (eds) Precision medicine for investigators, practitioners and providers. Academic Press, London, pp 59–69

Brown TA (2002) Genomes, 2nd edn. Wiley-Liss, Oxford

Buchanan CC, Torstenson ES, Bush WS et al (2012) A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. J Am Med Inf Assoc 19:289–294

Cargill M, Altshuler D, Ireland J et al (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238

Carter NP, Fiegler H, Piper J (2002) Comparative analysis of comparative genomic hybridization microarray technologies: report of a workshop sponsored by the Wellcome Trust. Cytometry 49:43–48

Carvajal-Rodríguez A (2010) Simulation of genes and genomes forward in time. Curr Genomics 11:58–61

Chaudhari NM, Gupta VK, Dutta C (2016) BPGA-an ultra-fast pan-genome analysis pipeline. Sci Rep 6:24373

Chellapilla K (1998) Combining mutation operators in evolutionary programming. IEEE Trans Evol Comput 2:91–96

Chen ZJ, Ni Z (2006) Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. Bioessays 28:240–252

Chippindale PT, Wiens JJ (1994) Weighting, partitioning, and combining characters in phylogenetic analysis. Syst Biol 43:278-287

Chou KC (2004) Structural bioinformatics and its impact to biomedical science and drug discovery. Curr Med Chem 11:2105–2134

Cong L, Ran FA, Cox D et al (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339:819–823

Darriba D, Flouri T, Stamatakis A (2018) The state of software for evolutionary biology. Mol Biol Evol 35:1037–1046

Eiben AE, Smith JE (2015) What is an evolutionary algorithm? In: Eiben AE, Smith JE (eds) Introduction to evolutionary computing. Springer, Berlin, Heidelberg, pp 25–48

Eiben ÁE, Hinterding R, Michalewicz Z (1999) Parameter control in evolutionary algorithms. IEEE Trans Evol Comput 3:124–141

Ekblom R, Wolf JB (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042

Elsik CG, Unni DR, Diesh CM et al (2016) Bovine Genome Database: new tools for gleaning function from the Bostaurus genome. Nucleic Acids Res 44:D834–D839

Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fogel GB, Corne DW (2002) Evolutionary computation in bioinformatics. Elsevier

Garland T Jr, Ives AR (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am Nat 155:346–364

Ghosh SK (2015) Evolution of illustrations in anatomy: a study from the classical period in Europe to modern times. Anat Sci Educ 8:175–188

Gibson DG, Glass JI, Lartigue C et al (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329:52–56

Gohil N, Bhattacharjee G, Lam NL et al (2021) CRISPR-Cas systems: challenges and future prospects. In: Singh V (eds) Progress in Molecular Biology and Translational Science, Volume 180. https://doi.org/10.1016/bs.pmbts.2021.01.008

Gohil N, Panchasara H, Patel S et al (2017) Book review: recent advances in yeast metabolic engineering. Front Bioeng Biotechnol 5:71

Gohil N, Panchasara H, Patel S et al (2019) Molecular biology techniques for the identification and genotyping of microorganisms. In: Tripathi V, Kumar P, Tripathi P, Kishore A (eds) Microbial genomics in sustainable agroecosystems. Springer, Singapore, pp 203–226

Graham CH, Ferrier S, Huettman F et al (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol Evol 19:497–503

Grau J, Ben-Gal I, Posch S et al (2006) VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees. Nucleic Acids Res 34:W529–W533

Graves PR, Haystead TA (2002) Molecular biologist's guide to proteomics. Microbiol Mol Biol Rev 66:39–63

Gu J, Bourne PE (2011) Structural bioinformatics, 2nd edn. Wiley, Hoboken

Hall DA, Ptacek J, Snyder M (2007) Protein microarray technology. Mech Ageing Dev 128:161–167

Hindorff LA, Sethupathy P, Junkins HA et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci 106:9362–9367

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267

Hye-Jung EC, Jaswinder K, Martin K et al (2014) Second-generation sequencing for cancer genome analysis. In: Dellaire G, Berman JN, Robert JA (eds) Cancer genomics. Academic Press, Boston, USA, pp 13–30

Ilari A, Savino C (2008) Protein structure determination by x-ray crystallography. In: Keith JM (ed) Bioinformatics. Humana Press, pp 63–87

International Sheep Genomics Consortium, Archibald AL, Cockett NE, Dalrymple BP et al (2010) The sheep genome reference sequence: a work in progress. Anim Genet 41:449–453

Jespersen MC, Peters B, Nielsen M et al (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res 45:W24–W29

Kanehisa M, Furumichi M, Tanabe M et al (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353–D361

Koonin E, Galperin MY (2002) Sequence—evolution—function: computational approaches in comparative genomics

Kukurba KR, Montgomery SB (2015) RNA sequencing and analysis. Cold Spring Harb Protoc 2015:951–969

Lewin HA, Robinson GE, Kress WJ et al (2018) Earth BioGenome Project: sequencing life for the future of life. Proc Natl Acad Sci 115:4325–4333

Li FW, Harkess A (2018) A guide to sequence your favorite plant genomes. Appl Plant Sci 6:e1030

Londin E, Yadav P, Surrey S et al (2013) Use of linkage analysis, genome-wide association studies, and next-generation sequencing in the identification of disease-causing mutations. Pharmacogenomics. Methods Mol Biol 1015:127–146

Luscombe NM, Greenbaum D, Gerstein M (2001) What is bioinformatics? An introduction and overview. Yearb Med Inf 1:83–100

Mani N, Mani A (2017) Design of cellular quantum-inspired evolutionary algorithms with random topologies. In: Bhattacharya S, Maulik U, Dutta P (eds) Quantum inspired computational intelligence, pp 111–146

Mitchell M, Taylor CE (1999) Evolutionary computation: an overview. Ann Rev Ecol Syst 30:593–616

Morris JS, Baladandayuthapani V (2017) Statistical contributions to bioinformatics: design, modeling, structure learning, and integration. Stat Model 17:245–289

Orengo CA, Pearl FMG, Bray JE et al (1999) The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res 27:275–279

Pagel M (2006) Evolution, bioinformatics and evolutionary bioinformatics online 2

Pennacchio LA, Bickmore W, Dean A et al (2013) Enhancers: five essential questions. Nat Rev Genet 14:288–295

Primrose SB, Twyman R (2009) Principles of genome analysis and genomics. Wiley

Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501–D504

Romero IG, Ruvinsky I, Gilad Y (2012) Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet 13:505–516

Sanger F, Air GM, Barrell BG et al (1977) Nucleotide sequence of i X 174 DNA. Nature 265:687–695

Sherry ST, Ward MH, Kholodov M et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Singh H, Raghava GP (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17:1236–1237

Singh H, Raghava GP (2003) ProPred1: prediction of promiscuous MHC Class-I binding sites. Bioinformatics 19:1009–1014

Singh H, Ansari HR, Raghava GPS (2013) Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS One 8:e62216

Singh V, Braddick D, Dhar PK (2017) Exploring the potential of genome editing CRISPR-Cas9 technology. Gene 599:1–18

Singh V, Gohil N, Ramírez García R et al (2018a) Recent advances in CRISPR-Cas9 genome editing technology for biological and biomedical investigations. J Cell Biochem 119:81–94

Singh V, Gohil N, Ramírez-García R (2018b) New insight into the control of peptic ulcer by targeting the histamine H2 receptor. J Cell Biochem 119:2003–2011

Soh J, Gordon PM, Sensen CW (2012) The Bluejay genome browser. Curr Protoc Bioinf 37:10–19

Spears WM (1995) Adapting crossover in evolutionary algorithms. In: Evolutionary programming, pp 367–384

Stalker J, Gibbins B, Meidl P et al (2004) The Ensembl Web site: mechanics of a genome browser. Genome Res 14:951–955

Sudha G, Nussinov R, Srinivasan N (2014) An overview of recent advances in structural bioinformatics of protein–protein interactions and a guide to their principles. Prog Biophy Mol Biol 116:141–150

Sugimoto M, Kawakami M, Robert M et al (2012) Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. Curr Bioinf 7:96–108

Swofford DL (1993) PAUP: Phylogenetic Analysis Using Parsimony, version 3.1.1. Laboratory of Molecular Systematics, Smithsonian Institution, Washington D.C

Tamura K, Stecher G, Peterson D et al (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729

Tettelin H, Masignani V, Cieslewicz MJ et al (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci 102:13950–13955

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tosto G, Reitz C (2013) Genome-wide association studies in Alzheimer's disease: a review. Curr Neurol Neurosci Rep 13:381

Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291:1304–1351

Véron A, Blein S, Cox DG (2014) Genome-wide association studies and the clinic: a focus on breast cancer. Biomarkers Med 8:287–296

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wiens JJ (1995) Polymorphic characters in phylogenetic systematics. Syst Biol 44:482–500

Xia X (2013) What is comparative genomics? In: Comparative genomics. Springer, Berlin, Heidelberg, pp 1–20

Bioinformatics Tools and Software

2

Aeshna Gupta, Disha Gangotia, and Indra Mani

Abstract

Bioinformatics or computational biology is a rapidly evolving field of science. Genome sequences of various organisms are routinely being deposited in biological databases owing to the availability of next generation sequencing (NGS). With an increasing amount of biological data being produced through different research projects, interpretation and analysis becomes very much essential to carry out meaningful functional and structural studies of nucleic acids and protein sequences. Various tools and software have been designed to perform such complex analysis. This chapter focuses on the different tools and software used in bioinformatics for the purpose of sequence submission, sequence retrieval, structure submission, sequence analysis, and structure prediction. The chapter highlights sequence submission tools like BanqIt, SPIN, WEBIN, Sequin, Sakura, structure submission tools like ADIT, pdb_extract, etc., and sequence retrieval tools such as SRS, Entrez, Getentry. Further, tools for sequence analysis like BLAST, CLUSTALW/X, and structure prediction tools such as SWISS-MODEL, Modeller, JPred, 3D-Jigsaw, and ModBase have been discussed in detail.

Keywords

Bioinformatics · Tools · Software · BLAST · Clustal W/X · SWISS-MODEL

A. Gupta · I. Mani (✉)
Department of Microbiology, Gargi College, University of Delhi, New Delhi, India
e-mail: indra.mani@gargi.du.ac.in

D. Gangotia
University College Dublin, Dublin, Ireland

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_2

2.1 Introduction

A large amount of “omic” data has been produced by the onset of immense technological advances in science. Comprehension of the massive amount of this sequence and structure data being produced at numerous levels of biological systems is the principle task (Pevsner 2015). This is where “Bioinformatics” comes into play. It is an interdisciplinary field that can be referred to as the use of computational algorithms to assemble, evaluate, comprehend, visualize, and archive data associated with biomolecules (Luscombe et al. 2001; Pevsner 2015). Various fields of modern biology like genomics, transcriptomics, proteomics, genetics, and evolution are incorporated within bioinformatics (Kumar and Chordia 2017). Applications of bioinformatics range from sequencing of genomes, prediction of gene and its function to protein analysis like prediction of protein structure and function, phylogenetic studies, designing drugs and vaccines, identification of organisms, and for supporting and advancing research in the area of biotechnology. Ultimately, bioinformatics facilitates the discovery of new biological insights (Kumar and Chordia 2017). Due to a rapid increase of biological data in the form of sequence, structure, pathway, and interactions, biological science has become data-rich science.

2.2 Importance of Bioinformatics

The post-genomics revolution period has witnessed a large amount of data that relates to the analysis of DNA, RNA, and proteins, the complex networks and ecosystems in which living organisms engage, and the crucially important metadata—which puts “omics” data in context. The 2020 coronavirus pandemic proves the importance of rapid data analysis and interpretation in controlling the spread through data being shared quickly and openly. This further throws light on the significance of bioinformatics in data sharing and analysis (Peter Bickerton 2020). In a given sample, metagenomics (culture-independent molecular approach) is the process of sequencing DNA from the genomes of all species and is one of the common methods for the study of the structure and function of the microbiome population. Researchers are discovering new metagenome-encoded genes, many of which may be of biotechnological or pharmaceutical concern. Such kind of analysis requires complex bioinformatics tools to analyze the data (Roumpeka et al. 2017). Several methodologies are used to deduce different levels of microbiome knowledge. These techniques include analysis of the 16S ribosomal RNA (16S rRNA) gene, analysis of the whole genome shotgun (WGS; metagenome), and analysis of the whole transcriptome shotgun known as metatranscriptome (Niu et al. 2018). Some bioinformatics tools used for analyzing 16S rRNA are QIIME, UPARSE, MOTHUR, DADA2, and minimum entropy decomposition (MED) (Niu et al. 2018). Metagenomic analysis tools include MetaPhlAn2 (Truong et al. 2016), Kraken (Wood and Salzberg 2014), CLARK (Ounit et al. 2015), FOCUS (Silva et al. 2014), SUPER-FOCUS (Silva et al. 2016), and MG-RAST (Meyer et al. 2008).

Metagenomics and metatranscriptomics have played a very significant role with the increased limitations in understanding the mechanisms of an individual microbe on a large scale and the difficulties associated with culturing individual microorganisms. Metatranscriptomic research also gives valuable insight into gene function by analyzing microbiota levels of gene expression or additional insight into gene expression profiles and even regulatory mechanisms, which can contribute effectively to drug development and human health (Bashiardes et al. 2016). Metatranscriptomics research methods include HUMAnN2, MetaTrans, SAMSA, and Leimena-2013 (the pipeline does not have a clear name). Moreover, for the study of metatranscriptome data sets, another method, MG-RAST, can be used.

Metaproteomics helps researchers to classify on a wide scale the entire protein complement of complex microbiomes (Wilmes and Bond 2004). The primary objective of metaproteomics is to examine the critical functions of multiple species within an environmental or host ecosystem that sustains the metabolic activity of the microbial population. Tröscher et al. used metaproteomics to analyze intestinal microbiome samples from five separate porcine gastrointestinal tract regions to gain functional information on bacterial groups with a combined host-specific protein analysis (Tröscher-Mußotter et al. 2019). A web-based bioinformatics platform for disseminating metaproteomics workflows and software is the Galaxy-P platform (Blank et al. 2018). The modular data analysis framework offers numerous processing steps related to the data analysis of metaproteomics, including the generation of databases, peptide spectrum matching, taxonomic annotation, and functional analysis (Seifert and Muth 2019). Using gene ontology (GO) terminology, MetaGOmics is an online tool that automates the quantitative functional analysis of metaproteome results. The functional overview of a metaproteomics workflow is performed by the tool (Riffle et al. 2017). Metabolomics is the biological perception of environmental factors causing changes to metabolic pathways. By analyzing thousands of small molecules in cells, tissues, organs, or biological fluids, it captures global biochemical events accompanied by the application of computer techniques to identify metabolite biomarkers. List of some metabolic pathway databases and visualization tools includes KaPPA-View, KEGG (Kyoto Encyclopedia of Genes and Genomes), HumanCyc, MetaCyc, and MetaMapp (Kusonmano et al. 2016). These different omics, such as metagenomics, metatranscriptomics, metaproteomics, and metabolomics, are an enormous source of biological data. For annotation and curation of these data, there are various tools available.

2.3 Tools Used in Bioinformatics

2.3.1 Sequence Submission Tools

Availability of biological sequences in the public domain is essential for their universal access. This enables the research community to conduct searches and analyses of similarity/homology on the current nucleotide and protein sequence data. Following are some widely used sequence submission tools used to submit sequence data to major biological databases like NCBI (National Centre for Biotechnology Information), EMBL (European Molecular Biology Laboratory), and DDBJ (DNA Data Bank of Japan).

2.3.1.1 Banqlt

BankIt is a web-based tool to submit sequence data to NCBI-GenBank (https://www.ncbi.nlm.nih.gov/WebSub/). It very well may be utilized to submit a single sequence, a few unrelated sequences (with different features and/or source information), or even a large set of sequences (Fetchko and Kitts 2011).

2.3.1.2 SPIN

SPIN (https://www.ebi.ac.uk/swissprot/Submissions/spin/) is a web interface provided by UniProt (Universal Protein Resource) for submission of protein sequences for which data at the protein level is present. Any protein sequence that is determined either through Edman degradation or through mass spectrometry results interpreted manually can be submitted. Sequence data obtained from peptide mass fingerprinting or some other procedure of mass spectrometry based on database searches or any nucleotide sequence that is translated is not recognized. Therefore, the newly sequenced proteins are accessible to the research community and scientists can obtain UniProt accession numbers via this service that can be used in literature (Pichler et al. 2018).

2.3.1.3 WEBIN

WEBIN (https:/www.ebi.ac.uk/ena/submit/sra/#home) is EMBL's recommended web-based submission platform for nucleotide sequences and biological annotation data (biology associated with sequence data). This tool enables the submission of single, multiple, or very large numbers of sequences (bulk sequences) (Stoesser et al. 2002).

2.3.1.4 Sequin

Sequin (https:/www.ncbi.nlm.nih.gov/Sequin/) is a standalone software to submit and update sequence data to the GenBank, EMBL, or DDBJ sequence database. It is devised by the NCBI. It is ideal for processing simple submissions containing a single sequence of short mRNA. Complicated entries containing long sequences with or without gaps, various annotations, or phylogenetic and population studies can likewise be submitted with the assistance of this software (Stoesser et al. 2002).

2.3.1.5 SAKURA

SAKURA is a data submission system developed by DDBJ. It is a World Wide Web (WWW) interface-oriented system. Submitters have an opportunity to "pause and resume" their work while using this tool, in which the typed-in information is temporarily kept on the server for one month, unless the session is intentionally terminated by the submitter. When contrasted with E-mail entries, errors are more thoroughly checked by SAKURA. Three types of errors are classified by this system, mandatory, illegal, or semantic, and will issue error and warning messages whenever appropriate (Yamamoto et al. 1996).

2.3.2 Sequence Retrieval Tools

Retrieval of data is as important as submission and one of the main objectives of any database is to provide the users with the required information. Any database contains immense amounts of information, retrieving which is also a critical task depending on the right use of search strings.

2.3.2.1 Entrez

Entrez (http:/www.ncbi.nlm.nih.gov/Entrez/) is a text-based search and retrieval platform implemented by NCBI that offers interconnected links to nucleotide and protein sequence information, information on gene and genome mapping, structural data, biomedical literature, etc. It constitutes over 20 databases including the nucleotide sequence data from GenBank that includes information from EMBL and DDBJ and complete protein sequence data from PIR (Protein Information Resource)-International, PRF (Protein Research Foundation), Swiss-Prot, PDB (Protein Data Bank), and database documents containing biological sequence and 3-D structural data, or abstracts from the scientific literature can be retrieved using simple Boolean queries (type of search allowing users to combine keywords or phrases with operators such as “and,” “not,” and “or” to further produce more relevant results) (Schuler et al. 1996). A single, well-defined object (e.g. a particular protein sequence or PubMed citation) is recognized by a unique ID (UID) (Ostell 2002). Results may be viewed in various formats like flat-file, FASTA, XML, etc. Entire genomes or chromosomes, as well as biological annotation on individual sequences can be visualized via a graphical interface.

However, documents identified in this manner are not endpoints in themselves. Instead, they serve as entry points for further exploration with the help of hypertext links. For example, cross-reference between a sequence and the corresponding research article in which it was reported, or between a protein sequence and the sequence of the gene encoding it, is possible when using Entrez (Schuler et al. 1996). The extent of such hypertext links can also be expanded to include external services, such as organism-specific genome databases through a service called LinkOut (Sayers et al. 2009). Henceforth, connections between different data that may propose future analyses or help in understanding of the available information can be deduced through Entrez.

2.3.2.2 SRS

Sequence retrieval system (SRS), developed at the European Bioinformatics Institute (EBI) at Hinxton, UK, is a homogeneous interface to approximately 80 biological databases. It is suitable for flat-file databases, such as the EMBL nucleotide sequence database or the Swiss-Prot database of protein sequences. It comprises databases of sequence and protein 3-D structure data, information about metabolic pathways, transcription factors, application results (like BLAST), genomes, mutations, etc. SRS data retrieval is usually limited to searching within particular data fields for the existence of key terms. It does not address complicated queries involving numerical data or computations (Etzold and Argos 1993).

2.3.2.3 Getentry

Getentry (http://getentry.ddbj.nig.ac.jp/top-e.html) is used to retrieve data from various databases of DDBJ. Unique identifiers required for retrieval through Getentry can be accession number, gene name, etc.

2.3.3 Structure Submission Tools

The worldwide Protein Data Bank (wwPDB) is a partnership of servers for the collation, maintenance, and dissemination of macromolecular structure data. It contains structures of biological macromolecules determined by NMR (nuclear magnetic resonance), X-ray or neutron diffraction, and cryo-electron microscopy (Abriata 2017). The current wwPDB members include the Research Collaboratory for Structural Bioinformatics PDB (RCSB PDB) in the USA, the PDB in Europe (PDBe), PDB Japan (PDBj), and the Biological Magnetic Resonance Bank (BMRB, University of Wisconsin in the USA). Entries in PDB include structures of isolated proteins, nucleic acids, their complexes with each other as well as with other small molecules like cofactors, substrate analogues, regulators, inhibitors, etc. (Abriata 2017). For automated and precise structure deposition, various tools have been designed by PDB. The deposition process consists of assembling and entering data (coordinates and structure-factor files, source and sequence of the macromolecules in the structure, citations) and finally submitting it to the PDB (Yang et al. 2004). Following are some of the tools for deposition of structure data to PDB:

2.3.3.1 ADIT

ADIT "Auto Dep Input Tool" (http:/deposit.pdb.org/adit) is an integrated structural data collection, editing, evaluation, and deposition software system for the PDB. Three tasks can be performed during an ADIT session: (a) a data-format pre-check in which the format of the coordinate data file is checked to ensure that it conforms with either PDB or mmCIF (macromolecular Crystallographic Information File) format; (b) validation, which requires verifying the accuracy of data with known standards and generating a report; (c) actual deposition of a structure. All categories in ADIT should be completed correctly during the deposition process and reviewed before submission. A PDB ID is allocated to the entry upon the structure's successful deposition (Yang et al. 2004).

2.3.3.2 pdb_extract

Pdb extract is an application that, at each level of the process of structure determination, can retrieve information from the output of standard crystallographic programs. In order to construct two mmCIF data files, one with structure factors and the other with structure details, along with its coordinates, files containing the necessary information are merged. These two mmCIF data files have now become fully prepared for deposition and authentication. Three versions of pdb_extract are available: a web interface (http://pdb-extract.rutgers.edu), a standalone application, and part of the CCP4 package (Collaborative Computational Project No. 4 - Software for Macromolecular X-Ray Crystallography) (Yang et al. 2004).

2.3.3.3 AutoDep

AutoDep (https://www.ebi.ac.uk/pdbe/deposition) is a web-based tool developed by EBI (European Bioinformatics Institute) for the submission of X-ray crystallography and NMR spectroscopy structures and data to the PDB and BMRB.

2.3.3.4 EMDep

EMDep (https://www.ebi.ac.uk/pdbe-emdep/emdep/) is a web-based tool developed by EBI that enables the submission of data to the Electron Microscopy Data Bank (EMDB), an archive of high-resolution 3D cryo-electron microscopy data. EMDB contains 3D maps (volumes), masks, images, and bibliographic citations, as well as processed primary data. The deposition system allows users to deposit 3D maps to EMDB and associated coordinate data to the PDB.

2.3.3.5 OneDep

A common portal for deposition of atomic coordinates and related experimental data derived from the three currently accepted structure determination techniques to the PDB archive exists. It is known as the OneDep system (http://deposit.wwpdb.org). It was conceived by the wwPDB partners as a multinational collaboration. Depending on the geographical location of the depositor, the structure is allocated to one of the wwPDB sites for processing: RCSB PDB, PDBe, or PDBj (Young et al. 2017).

2.3.4 Sequence Analysis Tools

With the use of sequence alignment, the structural and functional aspects of a novel sequence can be easily predicted. Higher the sequence closeness, more prominent is the opportunity that they share comparable structure or function. The sequence alignments can be of two sorts, i.e., looking at two (pairwise) or more sequences (numerous) for a progression of characters. Alignment of more than three proteins/ nucleotides sequences refers to multiple sequence alignment (MSA). The genes which are similar are the ones that may be conserved among different species (Troy et al. 2001).

2.3.4.1 BLAST

BLAST stands for basic local alignment search tool developed by Stephen Altschul of NCBI in 1990 (Altschul et al. 1990). It is one of the most commonly used programs for sequence analysis based on pairwise sequence alignment. It carries out alignment as well as provides statistical information about the alignment.

Pairwise sequence alignment helps to identify regions which are similar between the two biological sequences. The similarities obtained maybe indicative of the functional, structural, and evolutionary relationships. The algorithm of BLAST is based on heuristic word method. This technique works by discovering short stretches of indistinguishable or almost indistinguishable matches of letters in the two sequences. It is based on the premise that at least one word (short stretch of characters) should be common in two linked sequences. Once the word matches have been identified, extending the similarity regions from these words lead to formation of a longer alignment. In addition, finding a high sequence similarity region followed by joining neighboring high scoring regions leads to the development of a full alignment (Xiong 2006). Michael J Conway showed that BLAST analysis of the cDNA pool of cell line of *Carassius auratus*, commonly known as crucian carp and head kidney tissue of *Ctenopharyngodon idella*, commonly known as grass carp, indicates the belongingness of the sequence to SARS-like coronaviruses and the evolutionary divergence of sequences in other species. Therefore, it could be likely that SARS-like coronaviruses in aquatic habitat regions are a widespread environmental pathogen (Conway 2020).

Types of BLAST

Different variants of BLAST program have been developed, namely blastn, blastp, blastx, tblastn, and tblastx. These programs are based on the form of query sequences that may be protein or nucleotide sequences, accessible via https:/blast.ncbi.nlm.nih.gov/Blast.cgi. Table 2.1 represents the query and subject sequences of different types of BLAST programs. Blastn has a nucleotide sequence query with nucleotide sequence database. For protein sequence queries, Blastp scans the protein sequence database. Blastx has a query of the nucleotide sequence, which is translated into all six reading frames and displays the subject sequence as the translated protein sequence. tblastn has protein sequences as query against nucleotide database.

It is ideal for searching protein homologs, which are encoded in newly sequenced genomes. The tblastx uses nucleotide sequences as query and searches for a collection of nucleotide sequence databases having all translated sequences (Berkley Library, University of California, 2020, https://guides.lib.berkeley.edu/ncbi). In addition to this, BLAST offers programs for special purposes as well, ex. bl2seq, immunoglobulin BLAST, and VecScreen. In view of its speed, high selectivity, and adaptability, BLAST is frequently the best option among other sequence similarity search programs, and more importantly, this method forms the basis for genome annotation.

Table 2.1 List of different BLAST programs and their sequence searches

Type of BLAST	Query sequence	Subject sequence
BLASTn	Nucleotide	Nucleotide
BLASTp	Protein	Protein
BLASTx	Nucleotide	Protein
TBLASTn	Protein	Nucleotide
TBLASTx	Nucleotide	Nucleotide

Statistical Significance

A list of matching sequences ranked on the basis of statistical significance is given by the BLAST output window. The scores provided help to distinguish between sequences that are evolutionarily related and unrelated. One of the most important statistical indicators in BLAST is the *E-value or expectation value*. As the name suggests, it is possible that a random chance would cause the resulting alignments from a database search. This provides details on the probability that a given sequence match is merely occurring by chance. Therefore, lesser the E-value, there is lower likelihood of the database match to occur by random chance and thus, the match is more significant (Koonin and Galperin 2003). Another statistical parameter for BLAST includes *Bit Score*. Bit score represents a prospective level for sequence comparison that is independent of size of the database and query length. The bit score, expressed as "S," is a standardized score communicated in bits that lets you gauge the size of the search space you would need to glance through before you would hope to discover a score on a par with or superior to this one by some coincidence (Fassler and Cooper 2011).

BLAST Output Format

A graphical presentation, a matching list, and the alignment portion are part of the BLAST output window. The graphical representation gives a quick representation of the degree of similarities between sequences and consists of colored horizontal bars (Fig. 2.1). Each color corresponds to the degree of similarity between the sequences, such as red for closely related sequences, green and blue are moderately related, and black is unrelated or novel. Each bar has a hyperlink to the pairwise sequence alignment in the alignment section. The matching list consists of the BLAST hits arranged in the order of decreasing score and increasing E-value. It also displays the accession number, title, percentage identity, bit score, and E-value for each hit (Fig. 2.2).

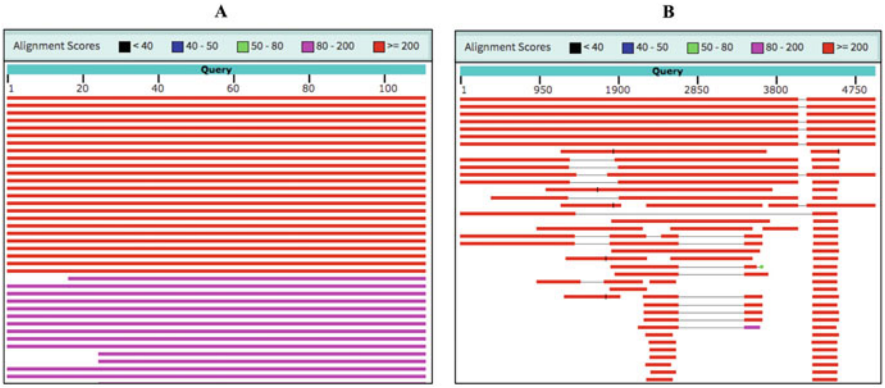
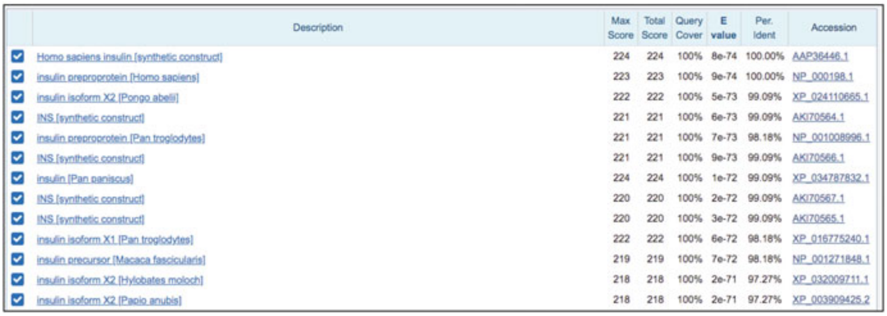
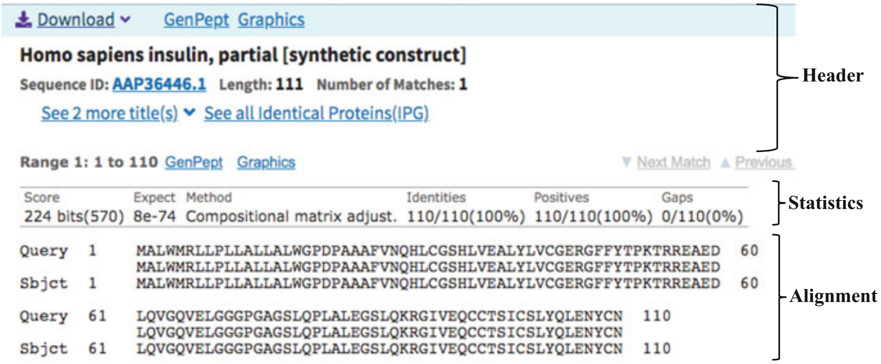


Fig. 2.1 (**a**) Graphical presentation of blastp for human insulin (AAA59172.1) and (**b**) blastn for human insulin (AH002844.2)



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Homo sapiens insulin [synthetic construct]	224	224	100%	8e-74	100.00%	AAP36446.1
☑	insulin preproprotein [Homo sapiens]	223	223	100%	9e-74	100.00%	NP_000198.1
☑	insulin isoform X2 [Pongo abelii]	222	222	100%	5e-73	99.09%	XP_024110665.1
☑	INS [synthetic construct]	221	221	100%	6e-73	99.09%	AKI70564.1
☑	insulin preproprotein [Pan troglodytes]	221	221	100%	7e-73	98.18%	NP_001008996.1
☑	INS [synthetic construct]	221	221	100%	9e-73	99.09%	AKI70566.1
☑	insulin [Pan paniscus]	224	224	100%	1e-72	99.09%	XP_034787832.1
☑	INS [synthetic construct]	220	220	100%	2e-72	99.09%	AKI70567.1
☑	INS [synthetic construct]	220	220	100%	3e-72	99.09%	AKI70565.1
☑	insulin isoform X1 [Pan troglodytes]	222	222	100%	6e-72	98.18%	XP_016775240.1
☑	insulin precursor [Macaca fascicularis]	219	219	100%	7e-72	98.18%	NP_001271848.1
☑	insulin isoform X2 [Hylobates moloch]	218	218	100%	2e-71	97.27%	XP_032009711.1
☑	insulin isoform X2 [Papio anubis]	218	218	100%	2e-71	97.27%	XP_003909425.2

Fig. 2.2 Matching list or description for human insulin (AAA59172.1): Displaying E-value, accession no., score, identity %



Homo sapiens insulin, partial [synthetic construct]

Sequence ID: AAP36446.1 Length: 111 Number of Matches: 1

See 2 more title(s) See all Identical Proteins(IPG)

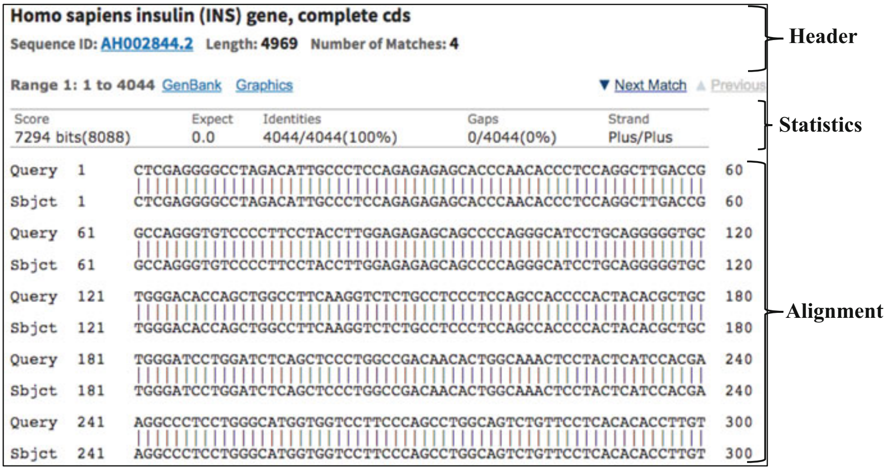
Range 1: 1 to 110 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
224 bits(570)	8e-74	Compositional matrix adjust.	110/110(100%)	110/110(100%)	0/110(0%)

```
Query  1   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60
           MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
Sbjct  1   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60

Query  61  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
           LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
Sbjct  61  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
```

Fig. 2.3 Alignment section of protein BLAST for human insulin (AAA59172.1)

This is trailed by the content depiction, which is additionally separated into three areas specifically header, statistics, and alignment. The description of the database sequence in a single line, along with the reference number of the database hit is displayed by the header section. The statistics result of the output is inclusive of the bit score, E-value, percentage identity, positives, and gaps. The last section or the alignment section of protein BLAST has query sequence on the top, a matching sequence in between, and subject sequence aligned to it at bottom (Fig. 2.3), while the nucleotide BLAST has just the query and subject sequences (Fig. 2.4). When the sequences have matching identical residues they are displayed at their respective positions (blastp) or a vertical line representing matches is present between the two sequences (blastn), while those sequences which are not identical matches have similar physiochemical properties or evolutionary conserved substitutions. For example, in case of amino acids, they are represented by a + sign and represent positive matches (Fig. 2.5). When two dissimilar residues are present, it represents a mismatch. A horizontal line represents a gap which is created in order to get the flanking region to match (Fig. 2.6) and all the low complexity regions are masked



Homo sapiens insulin (INS) gene, complete cds

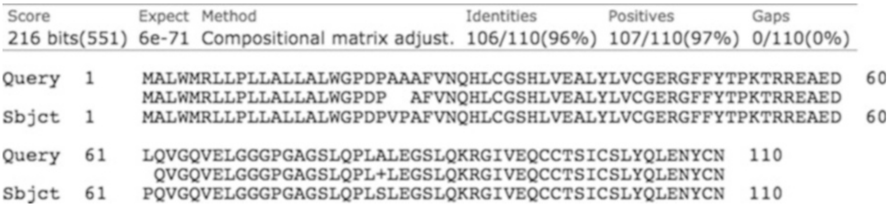
Sequence ID: AH002844.2 Length: 4969 Number of Matches: 4

Range 1: 1 to 4044

Score	Expect	Identities	Gaps	Strand
7294 bits(8088)	0.0	4044/4044(100%)	0/4044(0%)	Plus/Plus

```
Query  1    CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCG  60

Query  61   GCCAGGGTGTCCCCTTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   GCCAGGGTGTCCCCTTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGC  120

Query  121  TGGGACACCAGCTGGCCTTCAAGGTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  TGGGACACCAGCTGGCCTTCAAGGTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGC  180

Query  181  TGGGATCCTGGATCTCAGCTCCCTGGCCGACAACACTGGCAAACTCCTACTCATCCACGA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  TGGGATCCTGGATCTCAGCTCCCTGGCCGACAACACTGGCAAACTCCTACTCATCCACGA  240

Query  241  AGGCCCTCCTGGGCATGGTGGTCCTTCCCAGCCTGGCAGTCTGTTCCTCACACACCTTGT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  AGGCCCTCCTGGGCATGGTGGTCCTTCCCAGCCTGGCAGTCTGTTCCTCACACACCTTGT  300
```

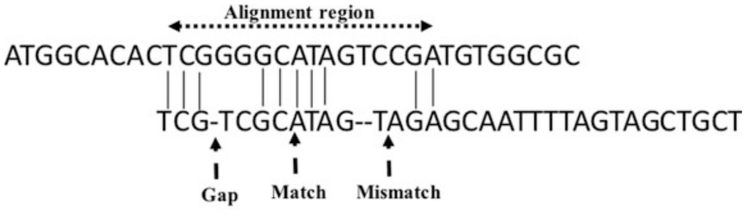
Fig. 2.4 Alignment section of nucleotide BLAST for human insulin (AH002844.2)



Score	Expect	Method	Identities	Positives	Gaps
216 bits(551)	6e-71	Compositional matrix adjust.	106/110(96%)	107/110(97%)	0/110(0%)

```
Query  1    MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60
            MALWMRLLPLLALLALWGPDP   AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
Sbjct  1    MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60

Query  61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
             QVGQVELGGGPGAGSLQPL+LEGSLQKRGIVEQCCTSICSLYQLENYCN
Sbjct  61   PQVGQVELGGGPGAGSLQPLSLEGSLQKRGIVEQCCTSICSLYQLENYCN  110
```

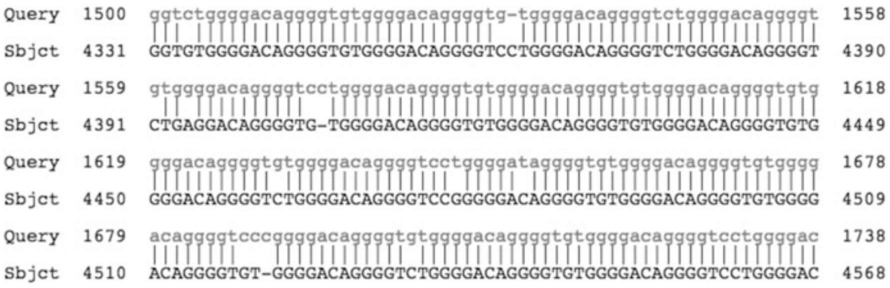
Fig. 2.5 Features of alignment output of a protein sequence


Alignment region
ATGGCACACTCGGGGCATAGTCCGATGTGGCGC
TCG-TCGCATAG--TAGAGCAATTTTAGTAGCTGCT
Gap
Match
Mismatch

$$\boxed{S = \sum(\text{Matches, mismatches}) - \sum(\text{gap penalties})}$$

$$\text{Score} = \text{Max}\ (S)$$

Fig. 2.6 Representing mismatch and gaps in nucleotide alignment



```
Query  1500  ggtctggggacaggggtgtggggacaggggtg-tggggacaggggtctggggacaggggt  1558
             ||| |||||||||||||||||||||||||||  |||||||||||||||||||||||||||
Sbjct  4331  GGTGTGGGGACAGGGGTGTGGGGACAGGGGTCCTGGGGACAGGGGTCTGGGGACAGGGGT  4390

Query  1559  gtggggacaggggtcctggggacaggggtgtggggacaggggtgtggggacaggggtgtg  1618
              || ||||||||||  ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4391  CTGAGGACAGGGGTG-TGGGGACAGGGGTGTGGGGACAGGGGTGTGGGGACAGGGGTGTG  4449

Query  1619  gggacaggggtgtggggacaggggtcctggggataggggtgtggggacaggggtgtgggg  1678
             ||||||||||| ||||||||||||||| ||||| ||||||||||||||||||||||||||
Sbjct  4450  GGGACAGGGGTCTGGGGACAGGGGTCCGGGGGACAGGGGTGTGGGGACAGGGGTGTGGGG  4509

Query  1679  acaggggtcccggggacaggggtgtggggacaggggtgtggggacaggggtcctggggac  1738
             ||||||||   |||||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  4510  ACAGGGGTGT-GGGGACAGGGGTCTGGGGACAGGGGTGTGGGGACAGGGGTCCTGGGGAC  4568
```

Fig. 2.7 Features of alignment output of a nucleotide sequence

with Xs or Ns or displayed as small letters such that it does not interfere with the alignment (Fig. 2.7).

2.3.4.2 CLUSTAL W

A multiple sequence alignment (MSA) software to match homologous nucleotide or protein sequences is CLUSTAL W (Thompson et al. 1994). It can be accessed through www.ebi.ac.uk/clustalw/ on the EMBL-EBI website. CLUSTAL and CLUSTAL V series of programs, which have been developed originally by Des Higgins; Fabian Sievers; David Dineen; Andreas Wilm, gave rise to CLUSTAL W. The "W" in CLUSTAL W stands for "weights" as we now give different weights to sequences and parameters at different positions in alignments. MSA encourages us to recognize the most developmentally conserved regions that are basic in functionality of a specified gene and recognize changes in the function just like its causes at the sequence level. Additionally, data in regard to structure and function of proteins can be acquired which is further useful in examining new domains or motifs having biological importance. The algorithm is based on the argument that groups of sequences are phylogenetically related, i.e., if they can be aligned, there is usually an underlying phylogenetic tree. This approach is commonly referred to as progressive alignment. Figure 2.8 briefly describes the algorithm behind Clustal programs (Aiyar 2000).

Certain basic features of CLUSTAL W include (1) support for more file formats for trees, sequence data sets, and alignments; (2) optional, full dynamic programming alignments for estimating the initial pairwise distances between all the sequences; (3) neighbor-joining TM trees for the initial guide trees, used to guide the progressive alignments; (4) sequence weighting to correct for unequal sampling of sequences at different evolutionary distances; (5) dynamic calculation of sequence- and position-specific gap penalties as the alignment proceeds; (6) the use of different weight matrices for different alignments; and (7) improved facilities for adding new sequences to an existing alignment (Higgins et al. 1996). New options have been included in Clustal W 2.0 and 2.1 (latest), to permit quicker arrangement of huge data sets and to build accuracy of alignment. Moreover, it is capable of handling some very difficult protein alignment problems as well. The Clustal W results then further can be used for creating phylogenetic trees, which

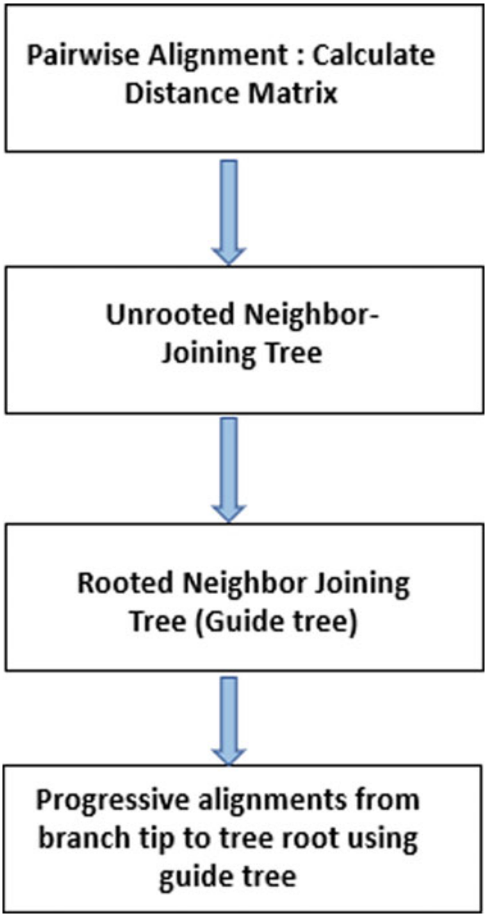

Pairwise Alignment : Calculate Distance Matrix
Unrooted Neighbor-Joining Tree
Rooted Neighbor Joining Tree (Guide tree)
Progressive alignments from branch tip to tree root using guide tree

Fig. 2.8 Flow chart of algorithm behind clustal programs

helps in analyzing the phylogenetic relationships among the query sequences (Larkin et al. 2007).

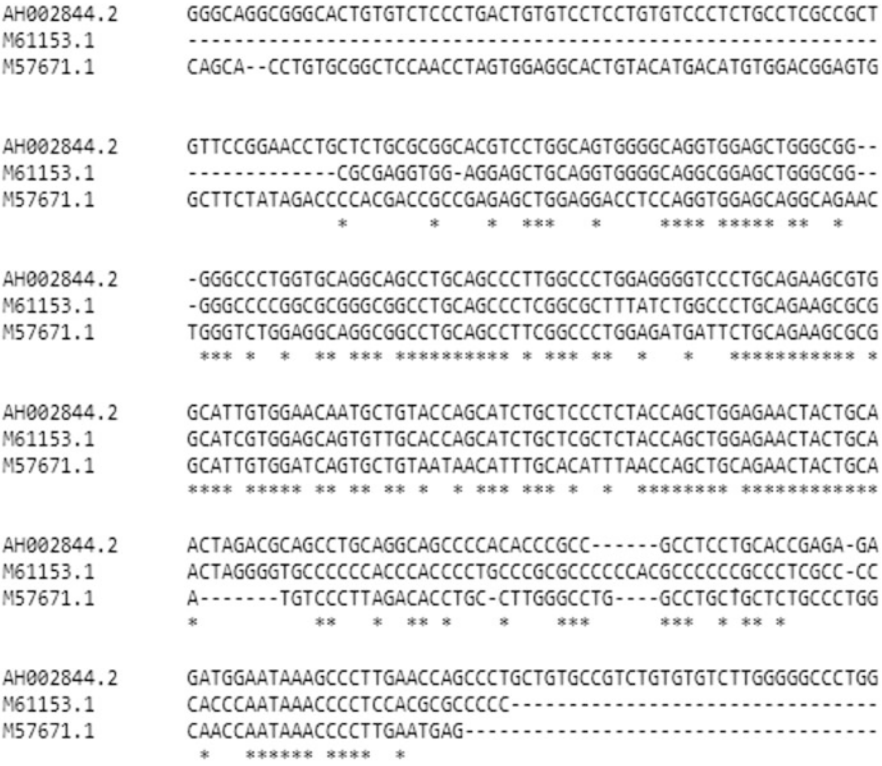
The coronavirus spike protein cytoplasmic tail functions in cellular fusion and subsequent pathogenicity was studied by Sadasivan et al. The research showed that the localization of the spike protein is mediated by the signal sequence present at the cytoplasmic tail, which was characterized by sequence alignment using Clustal W (Sadasivan et al. 2017). In another study, in order to provide basic data for the treatment and prevention of hepatitis C virus (HCV) infection, the existing HCV genotypes in diagnosed cases of infection were identified in Hohhot, China. The sequences compared using NCBI BLAST revealed the reference sequence of maximum similarity and enabled identification of HCV genotypes followed by creation of a homologous relationship tree using MegAlign Clustal W. Finally, the distribution characteristics in HCV genotypes, as well as the relationship between genotypes and host age and sex, were obtained (Lang et al. 2017).

Output Format

The output has a simple text mode interface with the sequences aligned to each other. "*" represents presence of identical residues or nucleotides in that column while a blank space is indicative of no match and gaps are represented using hyphens "-" (Fig. 2.9). A ":" shows conserved substitutions and "." represents semi-conserved substitutions (Fig. 2.10). The aligned sequences represent regions of similarity and could be indicative of closely related genes or common ancestry.

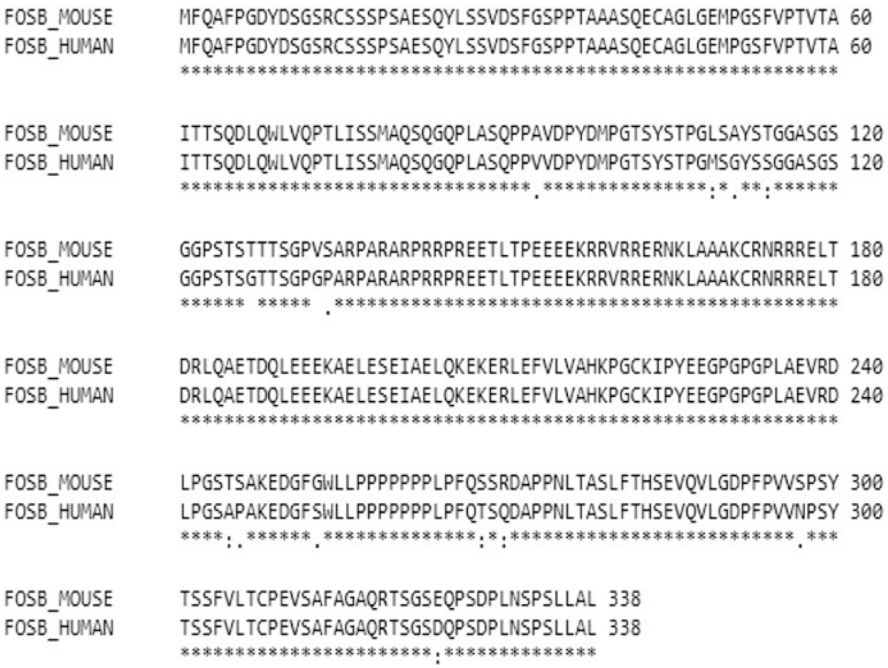
2.3.4.3 CLUSTAL X

CLUSTAL X is a variant of Clustal W that has a graphical user interface developed using the NCBI VIBRANT toolkit (Thompson et al. 1997). The current version of Clustal X named Clustal X 2.1 (Larkin et al. 2007) is available on Linux, Mac, and Windows (http://www.clustal.org/download/clustalx_help.html). The software is intended to (1) introduce multiple alignments, (2) see the effects of methods used for the alignment, and (3) strengthen it if possible. Clustal X has options that are unavailable in Clustal W which help in improving the alignment (such as choosing a part of the alignment with various gap penalties to be realigned while keeping the remainder of the alignment fixed). Clustal X utilizes a similar technique as Clustal W



```
AH002844.2      GGGCAGGCGGGCACTGTGTCTCCCTGACTGTGTCCTCCTGTGTCCCTCTGCCTCGCCGCT
M61153.1        ------------------------------------------------------------
M57671.1        CAGCA--CCTGTGCGGCTCCAACCTAGTGGAGGCACTGTACATGACATGTGGACGGAGTG


AH002844.2      GTTCCGGAACCTGCTCTGCGCGGCACGTCCTGGCAGTGGGGCAGGTGGAGCTGGGCGG--
M61153.1        -------------CGCGAGGTGG-AGGAGCTGCAGGTGGGGCAGGCGGAGCTGGGCGG--
M57671.1        GCTTCTATAGACCCCACGACCGCCGAGAGCTGGAGGACCTCCAGGTGGAGCAGGCAGAAC
                             *       *    *  ***   *     **** ***** **   *

AH002844.2      -GGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTG
M61153.1        -GGGCCCCGGCGCGGGCGGCCTGCAGCCCTCGGCGCTTTATCTGGCCCTGCAGAAGCGCG
M57671.1        TGGGTCTGGAGGCAGGCGGCCTGCAGCCTTCGGCCCTGGAGATGATTCTGCAGAAGCGCG
                 *** *  *  ** *** ********** * *** **  *   *   *********** *

AH002844.2      GCATTGTGGAACAATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCA
M61153.1        GCATCGTGGAGCAGTGTTGCACCAGCATCTGCTCGCTCTACCAGCTGGAGAACTACTGCA
M57671.1        GCATTGTGGATCAGTGCTGTAATAACATTTGCACATTTAACCAGCTGCAGAACTACTGCA
                **** ***** ** ** ** *  * *** *** *  *  ******** ************

AH002844.2      ACTAGACGCAGCCTGCAGGCAGCCCCACACCCGCC------GCCTCCTGCACCGAGA-GA
M61153.1        ACTAGGGGTGCCCCCCACCCACCCCTGCCCGCGCCCCCCACGCCCCCCGCCCTCGCC-CC
M57671.1        A-------TGTCCCTTAGACACCTGC-CTTGGGCCTG----GCCTGCTGCTCTGCCCTGG
                *          **   *  ** *    *    ***      ***  * ** *

AH002844.2      GATGGAATAAAGCCCTTGAACCAGCCCTGCTGTGCCGTCTGTGTGTCTTGGGGGCCCTGG
M61153.1        CACCCAATAAACCCCTCCACGCGCCCCC--------------------------------
M57671.1        CAACCAATAAACCCCTTGAATGAG------------------------------------
                 *   ****** ****  *
```

Fig. 2.9 Output of Clustal W 2.1 showing MSA between *Homo sapiens* insulin (AH002844.2), *Oryctolagus cuniculus* insulin mRNA (M61153.1), and *Octodon degus* insulin mRNA (M57671.1)



```
FOSB_MOUSE      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
FOSB_HUMAN      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
                ************************************************************

FOSB_MOUSE      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 120
FOSB_HUMAN      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 120
                ********************************.***************:*.**:******

FOSB_MOUSE      GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT 180
FOSB_HUMAN      GGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT 180
                ****** ***** .**********************************************

FOSB_MOUSE      DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD 240
FOSB_HUMAN      DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD 240
                ************************************************************

FOSB_MOUSE      LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY 300
FOSB_HUMAN      LPGSAPAKEDGFSWLLPPPPPPPLPFQTSQDAPPNLTASLFTHSEVQVLGDPFPVVNPSY 300
                ****:.******.**************:*:**************************.***

FOSB_MOUSE      TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL 338
FOSB_HUMAN      TSSFVLTCPEVSAFAGAQRTSGSDQPSDPLNSPSLLAL 338
                ***********************:**************
```

Fig. 2.10 Example of output file of Clustal W 1.82 (http://meme-suite.org/doc/clustalw-format.html)

to develop the alignment, i.e., pairwise progressive sequence alignment. In CLUSTAL X, the main display window displays several alignment modes by default, but it can be easily changed. It has two modes which can be selected using the switch directly option: multiple alignment mode and profile alignment mode.

As opposed to CLUSTAL W, CLUSTAL X has the ability to construct a multi-alignment PostScript color output file that may be acceptable for publication or presentation. Another intriguing advantage of utilizing Clustal X over Clustal W is the capacity to outwardly assess the quality of the alignment, specifically, the capacity to feature regions where the alignment is poor. It is possible that one grouping has a short site that shows low degrees of residue similarities to the rest of arrangement. This can be because of developmental procedures, or usually, due to blunders in sequencing. Clustal X is fit for recognizing errors which may go unnoticed and become submitted to GenBank, such as a frameshift error that is corrected by a frameshift algorithm (Aiyar 2000).

In a three tree format, both CLUSTAL W and X will generate output, where trees are presented in ASCII text files that are readable by other programs. The CLUSTAL group, whose yield is descriptive, is one such format, listing all the pair distances between multiple aligned sequences and the number of alignment positions used for each sequence. Similarly, this configuration records the sequences and the branch lengths that are joined at each arrangement level. Another format called PHYLIP

(PHYLogeny Inference Package) displays trees, with branch order, branch lengths, and sequence names, as a progression of nested parentheses.

In a study spanning 3 years (2010–2013, growing seasons from May to July), determination of level of Cucumber mosaic virus (CMV) in Taro was carried out and the maximum likelihood phylogenetic trees of nucleotide and amino acid sequences were generated using Clustal X v1.8. The results showed that these CMV isolates from taro in China came under subgroup I (Wang et al. 2014). In another study conducted in Fujian province to detect the genotypes and the rate of infection for Japanese encephalitis virus (JEV) in mosquitoes, applications such as Clustal X (1.83), MegAlign, GeneDoc 3.2, and MEGA (Molecular Evolutionary Genetic Analysis) 4.0 were used to splice sequence and deduce amino acid sequence and phylogenetic tree differentiation analysis for nucleotides (He et al. 2012).

2.3.5 Structure Prediction Tools

Fruitful model structure requires in any event one tentatively constructed 3-D structure (format) that has a critical sequence similarity to the target sequence. Exploratory structure clarification and comparative modeling supplement each other in the investigation of the protein structure space. Displaying of protein structures generally requires broad aptitude in structural biology and the utilization of exceptionally specialized PC programs for each of the modeling steps (Tramontano et al. 2001). There are different 3-D structure modeling tools available.

2.3.5.1 SWISS-MODEL

SWISS-MODEL (www.expasy.ch/swissmod/SWISS-MODEL.html) is an automated 3D protein structure modeling browser, allowing the user to automatically post a sequence and get a structure. This tool is based on homology modeling or comparative modeling methods, which utilize experimental protein structures for model building of evolutionarily related proteins called as targets. 3D models built by SWISS-MODEL are combined into the INTERPRO database (Mulder et al. 2003). Via programmed alignment or the first approach mode or manual alignment or the optimization mode, the server creates a model. The client presents a succession contribution for displaying in the previous technique and the server aligns the sequence requested with PDB sequences using BLAST. A raw model is constructed after a choice of suitable models. GROMOS carries out structure refinement. Then again, the user can likewise determine or upload structures as templates. The user creates an alignment in the Swiss-PDB Viewer in the Optimize mode and submits it to the model development server (Schwede et al. 2003).

SWISS-MODEL consists of the following components:

I. The SWISS-MODEL pipeline: A suite of automated tools and databases for automatic protein structure simulation (Schwede et al. 2003).
II. A web-based user tool with graphics called the SWISS-MODEL Workspace (Biasini et al. 2014).

III. The SWISS-MODEL Repository: A knowledge database of continuously maintained homology models for many biomedically fascinating proteomes of model organism.

A recent study involving screening of the specific epitopic regions in the spike proteins and selection of their energetic, inhibitory concentration 50 (IC50), MHC II reactivity was done for SARS COV 2 with some of them proving to be a great target for vaccination. The structure quality of spike glycoproteins was verified using SWISS-MODEL, Phyre2, and Pymol. A potential function of glycosylation on epitopic area indicated significant impacts on epitopic acknowledgment which could be useful in developing reasonable immunization regimen against SARS CoV-2 (Banerjee et al. 2020). Rahman *et al.* screened natural compounds using an in silico approach in order to discover potential inhibitors of the host enzyme, transmembrane protease serine 2 (TMPRSS2), in case of SARS COV-2. Viral entry into host cells is mediated by the enzyme, and its inhibition renders the virus unable to bind to the enzyme 2 that converts angiotensin (ACE2). As a consequence, the pathogenesis of SARS-CoV-2 is limited. SWISS-MODEL was used to construct the 3-D structure of TMPRSS2, and validation was done by RAMPAGE (Rahman et al. 2020).

2.3.5.2 Modeller

MODELLER is used to model protein 3D structures for homology (Webb and Sali 2016; Marti-Renom et al. 2000). A sequence arrangement to be demonstrated with known associated structures is given by the user and a model containing all non-hydrogen particles is computed by MODELLER naturally. By fulfilling spatial constraints, the tool updates related protein structure modeling (Sali and Blundell 1993; Fiser et al. 2000). The program models the backbone exercising a restraint technique determined by homology, which relies on multiple sequence alignments to recognize highly conserved residues between target and format proteins. Conserved residues in replication from the layout structures are given high constraints. Less conserved residues are given less or none of the constraints, like loop residues (Xiong 2006). Finally, a 3D model can be generated by satisfying all the restraints as well as possible. In addition, it may perform several different tasks, such as de novo modeling of loops in protein structures, followed by optimizing protein structure models with regard to a flexibly specified objective feature. In addition, tasks such as multiple protein sequence and structure alignment, clustering, sequence database search, protein structure comparison, etc. can easily be performed (https://salilab.org/modeller/). The current release of Modeller is 9.24, which was released on Apr 9th, 2020 and can be accessed through http://bioserv.cbs.cnrs.fr/HTML BIO/frame mod.html.

2.3.5.3 JPred

JPred (in operation since 1998) is a server for prediction of the secondary structure of proteins. The current version is JPred v4 (http://www.compbio.dundee.ac.uk/jpred/). The Jnet calculation uses JPred so as to make more precise predictions. JPred also allows estimates of accessibility to solvents and coiled coil regions known as the Lupas technique, considering the protein secondary structure (http://www.compbio.dundee.ac.uk/jpred/about.shtml). Jnet algorithm of Jpred 3 provides prediction of secondary structure (α-helix, β-strand, and coil) with 81.5% accuracy. Regardless of whether a solitary or multiple protein sequence, Jpred infers alignment profiles which further make predictions of secondary structure and ability to dissolve. The predictions can be represented as colored HTML, plain text, PostScript, PDF, and through the highly versatile Jalview alignment editor to view and apply the results (Cole et al. 2008).

2.3.5.4 3D-Jigsaw

3D-JIGSAW (www.bmm.icnet.uk/servers/3djigsaw/) is an automated framework that uses homology modeling to construct three-dimensional models of proteins. 3D-JIGSAW comes in two modes, automatic or interactive, which allows you to select the templates and correct the alignments before submitting the model (https://bip.weizmann.ac.il/toolbox/structure/3d.html). The database method is the backbone of its loop modeling. There are features in the collaborative mode to modify alignments and pick models, loops, side chains. Conversely, if a submitted protein sequence has a greater identity than 40 percent with known protein structures, the automatic mode has no human involvement and modeling is completed (Xiong 2006).

2.3.5.5 ModBase

MODBASE (database access: http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi) is a database for models of annotated protein structure. ModPipe is an automatic modeling pipeline dependent on the programs like PSI-BLAST and MODELLER, from which these models are derived. Fold assignments and alignments are also a part of the database. Theoretically calculated models form the core of MODBASE, though it may contain significant errors, it does not have experimentally determined structures. Thus, the quality of such models needs to be assessed. In addition to this, knowledge about potential ligand binding sites, SNP (small nucleotide polymorphisms) annotation, and interactions among proteins is central to MODBASE (http://modbase.compbio.ucsf.edu/modbasecgi/display.cgi?server=modbase&type=general).

2.4 Concluding Remarks

The data annotation is very much important; otherwise, enormous information about sequences of nucleic acids is meaningless. It is only possible when we have good algorithm containing tools and software. Data sharing is one of the most critical aspects of biological sciences in current times. Bioinformatics is a multidisciplinary field which deals with computational analysis and sharing of a variety of biological data. High throughput in silico tools and software are the prerequisites for efficient and quick sequence retrieval, analysis, and structure prediction; thus, being instrumental in accelerating research areas of biotechnology, molecular biology, and drug design. Undoubtedly, this requires scientists and research groups with expertise in multiple areas of knowledge. Moreover, these tools and software significantly assist research in various fields of biological science including genomics, transcriptomics, proteomics, metabolomics, systems, and synthetic biology.

Competing Interests There is no competing interest.

References

Abriata LA (2017) Structural database resources for biological macromolecules. Brief Bioinform 18 (4):659–669

Aiyar A (2000) The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment. Methods Mol Biol 132:221–241

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Banerjee A, Santra D, Maiti S (2020) Energetics and IC50 based epitope screening in SARS CoV-2 (COVID 19) spike protein by immunoinformatic analysis implicating for a suitable vaccine development. J Transl Med 18(1):281

Bashiardes S, Zilberman-Schapira G, Elinav E (2016) Use of metatranscriptomics in microbiome research. Bioinform Biol Insights 10:19–25

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G et al (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42 (W1):195–201

Blank C, Easterly C, Gruening B, Johnson J, Kolmeder CA, Kumar P, May D, Mehta S, Mesuere B, Brown Z et al (2018) Disseminating metaproteomic informatics capabilities and knowledge using the Galaxy-P framework. Proteomes 6:1

Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36:W197–W201

Conway MJ (2020) Identification of coronavirus sequences in carp cDNA from Wuhan, China. J Med Virol. https://doi.org/10.1002/jmv.25751

Etzold T, Argos P (1993) SRS-an indexing and retrieval tool for flat file data libraries. Bioinformatics 9(1):49–57

Fassler J, Cooper P (2011) BLAST Glossary: The GenBank Submissions Handbook. National Center for Biotechnology Information (US), Bethesda (MD). https://www.ncbi.nlm.nih.gov/books/NBK62051/

Fetchko M, Kitts A (2011) What is BankIt? The GenBank Submissions Handbook. National Center for Biotechnology Information (US), Bethesda (MD). https://www.ncbi.nlm.nih.gov/books/NBK63590/

Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9:1753–1773

He XX, Wang HY, Fu SH et al (2012) Zhonghua shi yan he lin chuang bing du xue za zhi = Zhonghua shiyan he linchuang bingduxue zazhi. Chin J Exp Clin Virol 26(2):81–83

Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266:383–402

Koonin EV, Galperin MY (2003) Sequence - evolution - function: computational approaches in comparative genomics. Kluwer Academic, Boston. Chapter 4, Principles and Methods of Sequence Analysis. https://www.ncbi.nlm.nih.gov/books/NBK20261/

Kumar A, Chordia N (2017) Role of bioinformatics in biotechnology. Res Rev Biosci 12(1):116

Kusonmano K, Vongsangnak W, Chumnanpuen P (2016) Informatics for metabolomics. Adv Exp Med Biol 939:91–115

Lang J, Sun P, Lu S et al (2017) Genotypes of the Hepatitis C virus in infected patients in Hohhot, China. Bing Du Xue Bao 33(1):61–66

Larkin MA, Blackshields G, Brown NP et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948

Luscombe NM, Greenbaum D, Gerstein M (2001) What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med 40:346–358

Marti-Renom MA, Stuart A, Fiser A, Sánchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325

Meyer F, Paarmann D, D'Souza M et al (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform 9:386

Mulder NJ, Apweiler R, Attwood TK, Bairich A, Barrell D et al (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31:315–318

Niu SY, Yang J, McDermaid A, Zhao J, Kang Y, Ma Q (2018) Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. Brief Bioinform 19(6):1415–1429

Ostell J (2002, Updated 2014) The Entrez Search and Retrieval System. The NCBI handbook, 2nd edn. National Center for Biotechnology Information (US), Bethesda (MD). https://www.ncbi.nlm.nih.gov/books/NBK184582/

Ounit R, Wanamaker S, Close TJ, Lonardi S (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genom 16(1):236

Pevsner J (2015) Bioinformatics and functional genomics, 3rd edn. Wiley, Chichester

Pichler K, Warner K, Magrane M (2018) UniProt Consortium SPIN: submitting sequences determined at protein level to UniProt. Curr Protoc Bioinformatics 62(1):e52

Rahman N, Basharat Z, Yousuf M, Castaldo G, Rastrelli L, Khan H (2020) Virtual screening of natural products against Type II Transmembrane Serine Protease (TMPRSS2), the Priming Agent of Coronavirus 2 (SARS-CoV-2). Molecules 25(10):2271

Riffle M, May DH, Timmins-Schiffman E, Mikan MP, Jaschob D, Noble WS, Nunn BL (2017) MetaGOmics: a web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. Proteomes 6:2

Roumpeka D, Wallace RJ, Escalettes F, Fotheringham I, Watson M (2017) A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. Front Genet 8:1664–8021

Sadasivan J, Singh M, Sarma JD (2017) Cytoplasmic tail of coronavirus spike protein has intracellular targeting signal. J Biosci 42(2):231–244

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815

Sayers EW, Barrett T et al (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 37(Database issue):D5–D15

Schuler GD, Epstein JA, Ohkawa H, Kans JA (1996) Entrez: molecular biology database and retrieval system. Methods Enzymol 266:141–162

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31(13):3381–3385

Seifert J, Muth T (2019) Editorial for special issue: metaproteomics. Proteomes 7(1):9

Silva GG, Cuevas DA, Dutilh BE, Edwards RA (2014) FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ 2:e425

Silva GG, Green KT, Dutilh BE, Edwards RA (2016) SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics 32(3):354–361

Stoesser G, Baker W, van den Broek A et al (2002) The EMBL Nucleotide Sequence Database. Nucleic Acids Res 30(1):21–26

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Tramontano A, Leplae R, Morea V (2001) Analysis and assessment of comparative modeling predictions in CASP4. Proteins 45(Suppl 5):22–38

Tröscher-Mußotter J, Tilocca B, Stefanski V, Seifert J (2019) Analysis of the bacterial and host proteins along and across the porcine gastrointestinal tract. Proteomes 7:4

Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT et al (2001) Sequence-evolution - function: computational approaches in comparative genomics. Chapter 4: principles and methods of sequence analysis. Genetic evidence for Near-Eastern origins of European cattle. Nature 410:1091

Truong DT, Franzosa EA, Tickle TL et al (2016) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 13(1):101

Wang YF, Wang GP, Wang LP, Hong N (2014) First report of cucumber mosaic virus in Taro plants in China. Plant Dis 98(4):574

Webb B, Sali A (2016) Comparative protein structure modeling using modeller. Current protocols in bioinformatics 54. Wiley, pp 5.6.1–5.6.37

Wilmes P, Bond PL (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ Microbiol 6:911–920

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15(3):R46

Xiong J (2006) Essential bioinformatics. Cambridge University Press, New York, United States of America

Yamamoto H, Tamura T, Isono K et al (1996) SAKURA: a new data submission system of DDBJ to meet users' needs in the age of mass production of DNA sequences. Genome Inform 7:204–205

Yang H, Guranovic V, Dutta S, Feng Z, Berman HM, Westbrook JD (2004) Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. Acta Crystallogr D60:1833–1839

Young JY, Westbrook JD et al (2017) OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. Structure 25(3):536–545

Online Resources

https://www.ncbi.nlm.nih.gov/Sequin/

http://www.clustal.org/download/clustalx_help.html

https://www.genome.jp/tools-bin/clustalw

https://www.ncbi.nlm.nih.gov/

https://www.ebi.ac.uk/Tools/psa/

https://guides.lib.berkeley.edu/ncbi

https://salilab.org/modeller/

https://www.ebi.ac.uk/training/online/course/pdbe-quick-tour/submitting-data-pdb-and-emdb

https://www.ddbj.nig.ac.jp/getentry-help-e.html#About_getentry

Peter Bickerton, 2020 Scientific Communications & Outreach Manager, Erlham Institute https://www.earlham.ac.uk/articles/why-bioinformatics-important

Role of Bioinformatics in Biological Sciences 3

Disha Gangotia, Aeshna Gupta, and Indra Mani

Abstract

Bioinformatics is an emerging area of science because technological advances in the field of life sciences have led to the generation of increasingly accumulating large volumes of biological data. The large size of such data will create a huge amount of value but presents numerous challenges of storage, annotation, and curation at the same time. Bioinformatics has allowed the handling of such a large amount of data and has thus paved the way for the rise of "omics" technologies. This chapter focuses on the role of bioinformatics in major "omics" fields, namely genomics, transcriptomics, proteomics, and metabolomics and also highlights upcoming fields like nutrigenomics, chemoinformatics, molecular phylogenetics, systems and synthetic biology, which have progressed due to the beautiful amalgamation of information technology, mathematics, chemistry, and omics sciences. The diverse research areas of bioinformatics like the development of biological databases, genome analysis, 3D structure prediction, drug discovery, clinical applications as well as mathematical modeling of metabolic processes have also been discussed in this chapter.

Keywords

Bioinformatics · Sequence · Structure · Databases · Genomics · Proteomics

D. Gangotia
University College Dublin, Dublin, Ireland

A. Gupta · I. Mani (✉)
Department of Microbiology, Gargi College, University of Delhi, New Delhi, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_3

3.1 Introduction

Bioinformatics is characterized as the use of computational and analytical instruments to identify and analyze biological data. It is an interdisciplinary field that takes advantage of computer science, mathematics, chemistry, physics, and biology. Bioinformatics is important for data management in modern biology and medicine. Other than analyzing the sequence data, it is used for a broad range of crucial tasks, including expression analysis and gene variation, gene and protein structure and prediction and function analysis, gene regulation network prediction and detection, whole cell modeling simulation environments, complex gene regulatory dynamics and network modeling, and presentation (Tsoka and Ouzounis 2000).

In addition, bioinformatics techniques are used by physicians to collect knowledge in the clinic or hospital environment about genetic disorders. An example of the application of new therapeutic advancements is the development of novel designer-targeted drugs such as imatinib mesylate (Gleevec) that hampers with an irregular protein produced in chronic myeloid leukemia. The ability to use bioinformatics techniques to recognize and target particular genetic markers enabled the discovery of this drug (Bayat 2002). The following section contains information about the role of bioinformatics in different areas of biology.

3.2 Role of Bioinformatics

3.2.1 Genomics

Genomics is the study of whole genomes (an organism's complete set of DNA). It allows sequencing, assembling, and analysis of the structure and function of genomes by using a combination of recombinant DNA technology, DNA sequencing techniques, and bioinformatics tools and software. It varies from "genetics" as genetics explores the functioning and structure of the single gene or single gene product at a time, whereas genomics approaches all genes and their interrelationships in order to understand their cumulative impact on the organism's growth and development. Genomics captures the availability of whole DNA sequences for organisms and was made possible by Frederick Sanger's initial work (Sanger et al. 1977) as well as the more recent next-generation sequencing (NGS) technology. Genomics is broadly divided into two main areas: structural genomics, demonstrating the physical structure of whole genomes; and functional genomics, depicting the transcriptome (total transcripts) and the proteome (the complete display of encoded proteins). Schematic representation of different omics is given in Fig. 3.1 (Virkud et al. 2019). All of these are explained in detail in the following subdivisions.

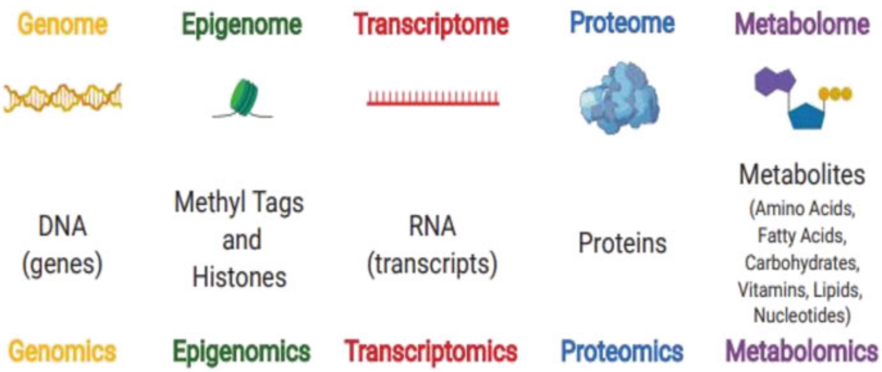


Fig. 3.1 Overview of the major "omics" fields (Virkud et al. 2019. *Adapted with permission*)

3.2.1.1 Structural Genomics

Structural genomics provides location to the entire set of genes in a genome, thus characterizing the structure of the genome. It proceeds through collective stages of analytic resolution of a genome, which begins with the allocation of genes and markers to single chromosomes, followed by their mapping within a chromosome and finally the creation of a physical map terminating in sequencing. In assigning genes or markers to individual chromosomes, many different methods are useful, for instance, linkage studies which give a rough idea of chromosomal position, pulsed field gel electrophoresis (PFGE) to separate chromosomal DNA of isolated fragments followed by locating new genes by hybridization, etc. (Lawrance et al. 1987; Sasaki et al. 2014).

The resolution is increased to the next level by determining the position of a gene or molecular marker on the chromosome, thus generating high-resolution chromosome maps using methods like meiotic linkage mapping, radiation hybrid mapping, etc. By directly manipulating cloned DNA fragments, a further improvement in mapping resolution is achieved [usually in high capacity vectors like bacterial artificial chromosome (BAC), yeast artificial chromosome (YAC), Cosmids, etc.]. The methods which involve the identification of a collection of cloned overlapping fragments that together constitute an entire chromosome or genome are commonly referred to as physical mapping because DNA is the physical material of the genome. The understanding of an individual genome structure is beneficial in mutating genes in that particular species. This study also makes it possible for processes such as transcription and translation to be explored. For several types of genetic analysis, including gene isolation and functional genomics, genetic and physical maps are an essential starting point.

3.2.1.2 Functional Genomics

The study which takes into account how genes and intergenic regions of the genome contribute to diverse biological functions is termed as functional genomics. It deals with a gene's complete structure, function, and regulation by incorporating molecular biology and cell biology studies. It is characterized by different areas of research, for instance, gene interaction mapping, analysis of single nucleotide polymorphisms

(SNPs), gene regulation (e.g. promoter analysis), microarray data analysis (gene expression studies), SAGE (serial analysis of gene expression; RNA sequencing for global gene expression studies in a cell), mutations, epigenetics, etc. (Kaushik et al. 2018). A widely used bioinformatics resource, the Database for Annotation, Visualization and Integrated Discovery (DAVID) allows characterization of functional genes, determines genes which are functionally related, enables gene/protein identifiers to be converted from one type to the other, and studies gene names in a set (Dennis et al. 2003; Huang et al. 2007).

An essential component of functional genomics is the human genome project (HGP) (Collins et al. 2003; Green et al. 2015) revealing that 3164.7 million nucleotide bases with a total of approximately 20,000 genes are found in the human genome. There are multiple fundamental strategies to functional genomics at different stages: genomics and epigenomics (DNA), transcriptomics (RNA), proteomics (proteins), and metabolomics (metabolites), the details of which are given in the following sections. Thus, it is expected that a wide-ranging model of the biological system under study will be provided by the compilation of all these data.

3.2.1.3 Nutritional Genomics

Human genome project (HGP) information provided an opportunity for researchers to understand the impact of genes and food bioactive compound interactions on human health. The term “nutrigenomics” or “nutritional genomics” has been coined to analyze this relationship between genes and nutrients. It encompasses the fields of genomics, transcriptomics, proteomics, physiology, nutrition, metabolomics, and epigenomics to search for and describe the common molecular-level interactions between genes and nutrients (Sales et al. 2014). Molecular tools are being utilized to identify, access, and understand the various results obtained between people or population groups through a certain diet (Cruz et al. 2003; Liu and Qian 2011; Dauncey 2012; Cozzolino and Cominetti 2013). Several instances of this interaction between gene and nutrients depend on their potential to bind to transcription factors. This binding thus affects the potential of transcription factors to associate with elements, leading to regulation of RNA polymerase binding.

Previous research with vitamins A, D, and fatty acids has shown that gene transcription can be triggered by their direct actions to activate nuclear receptors (Mahan and Stump 2005; Fialho et al. 2008; Cozzolino and Cominetti 2013; Kumar et al. 2014; Subramanian et al. 2016). Moreover, resveratrol, a compound found in wine and soy genistein, may have an indirect effect on the molecular signaling pathways, like the kappa B factor (Mahan and Stump 2005; Fialho et al. 2008; Dalmiel et al. 2012). The participation of these factors in the management of important molecules is attributed to illnesses varying from inflammation to cancer (Mahan and Stump 2005; Fialho et al. 2008).

3.2.2 Transcriptomics

The analysis of the transcriptome of an organism, the sum of all its RNA transcripts, is termed as transcriptomics. In the information network, mRNA acts as a transient intermediate molecule, while noncoding RNAs perform various different functions. The field of transcriptomics involves two predominant technologies: microarrays, which quantify the abundance of a given set of transcripts through their hybridization with a range of complementary probes, and RNA sequencing (RNA-Seq), which corresponds to the high-throughput sequencing of cDNA transcripts, where number of transcript counts is used to estimate the concentration. Large volumes of data are generated by transcriptomic analysis. In order to ensure their usefulness to the wider scientific community, raw or processed data may be stored in publicly accessible repositories such as Gene Expression Omnibus (GEO) (Edgar et al. 2002), ArrayExpress (Kolesnikov et al. 2015), etc.

Assessing an organism's gene expression patterns in various tissues, environments, or time intervals provides data on how genes are controlled, about functions of previously unannotated genes, and demonstrates features of the biology of an organism (Lowe et al. 2017). A significant application of this field lies in experimentation in diagnostics and disease profiling. The scope of using RNA-Seq to diagnose immune-related diseases is rapidly expanding due to its capability to distinguish populations of immune cells and sequence B-cell and T-cell receptor repertoires (Proserpio and Mahata 2016; Byron et al. 2016).

3.2.3 Proteomics

The large-scale analysis of the total protein complement of a cell line, tissue, or organism, i.e. its proteome, is referred to as proteomics (Wasinger et al. 1995; Wilkins et al. 1995; Anderson and Anderson 1996). Proteomics aims not only to characterize all proteins in a cell, but also to establish a precise three-dimensional cell map (3-D) that indicates where proteins are localized. Proteomics is thus considered to be the most important data set to describe a biological system as proteins are effectors of biological function, the levels of which depend on the corresponding levels of mRNA as well as on translational regulation of the host (Cox and Mann 2007). Expression proteomics refers to the quantitative analysis of protein expression between variable samples, while structural proteomics involves identifying and locating all proteins within a protein complex, determining their structure and analyzing all protein–protein interactions (Graves and Haystead 2002). Figure 3.2 illustrates the flowchart of various proteomics techniques.

An enormous amount of proteomics data is obtained with the help of high-throughput technologies. Various bioinformatics tools have been developed for predicting 3D structures, analyzing protein domains and motifs, interpretation of mass spectrometry results, etc. Evolutionary relationships can be inferred with the help of sequence and structure alignment tools (Vihinen 2001; Perez-Riverol et al. 2015). Proteomics-based techniques are used for various research environments,

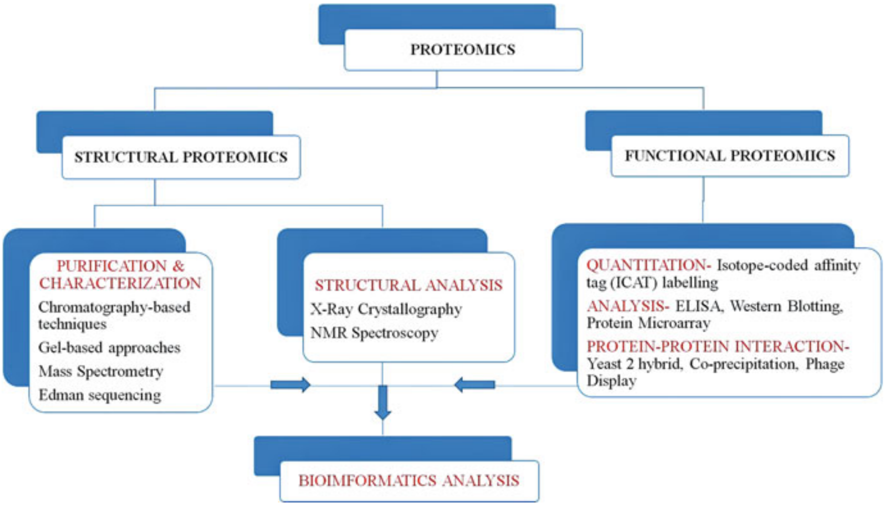

PROTEOMICS
STRUCTURAL PROTEOMICS
FUNCTIONAL PROTEOMICS
PURIFICATION & CHARACTERIZATION
Chromatography-based techniques
Gel-based approaches
Mass Spectrometry
Edman sequencing
STRUCTURAL ANALYSIS
X-Ray Crystallography
NMR Spectroscopy
QUANTITATION- Isotope-coded affinity tag (ICAT) labelling
ANALYSIS- ELISA, Western Blotting, Protein Microarray
PROTEIN-PROTEIN INTERACTION- Yeast 2 hybrid, Co-precipitation, Phage Display
BIOIMFORMATICS ANALYSIS

Fig. 3.2 Graphical representation of overview of the different proteomics techniques

such as diagnostic marker identification, vaccine production, the analysis of virulence mechanisms, the regulation of expression patterns in response to various signals, and the evaluation of protein pathways associated with several diseases (Aslam et al. 2017).

3.2.4 Metabolomics

Metabolomics is an emerging scientific area that explores the comprehensive estimation of all metabolites and low molecular weight molecules present in a biological sample. In response to chronic illnesses as well as monogenic diseases, detectable shifts in metabolite levels take place and these changes may show tissue specificity and temporal dynamics in comparison to the genome (Clish 2015). Metabolomics attempts to quantify molecules that have different physicochemical properties (e.g. varying in polarity from polar organic acids to nonpolar lipids) in comparison to genomic and proteomic strategies (Kuehnbaum and Britz-McKibbin 2013). Metabolomic techniques therefore help to classify the metabolome into metabolite sub-groups, depending on polarity of compounds, common functional groups, or similarity in structure, and hence, particular sample processing and analytical methods standardized for each sample are established (Clish 2015). Metabolomics has various health and disease applications, including personalized medicine, metabolic phenotyping, epidemiological studies, metabolome-wide association studies (MWAS), precision metabolomics, and as integrative omics, in conjunction with other omics sciences. List of some metabolic pathway databases and visualization tools include KaPPA-View, KEGG (Kyoto Encyclopedia of Genes and Genomes), HumanCyc, MetaMapp, etc. (Kusonmano et al. 2016).

3.2.5 Chemoinformatics

Chemoinformatics is the integration of computational and informational technique with the field of chemistry in the areas of topology, graphical theory, information extraction, and data collection in the chemical space. It deals with the conversion of data into information and information into knowledge for a broad range of applications, such as the advancement of biological systems research and development, software development techniques, drug design, etc. (Basuri and Meman 2011). Some of the major applications of chemoinformatics include:

1. **Prediction of Properties**: Analysis of physical, chemical, and biological properties such as adsorption, delivery, metabolism, excretion, and toxicity for drug design (ADME-Tox). For example, predicting the aqueous solubility, which is an important determinant of drug administration and absorption into the body (Strassberger et al. 2010).
2. **Analysis of Analytical Chemistry Data**: This includes analyzing samples to define and analyze the dynamic relationships between a sample's composition and its content, origin, or age (Maschio and Kowalski 2001).
3. **Computer-Assisted Structure Elucidation (CASE)**: The processing of large quantities of information includes elucidating a compound's composition from spectral analysis (Desany and Zhang 2004).
4. **Computer-Assisted Synthesis Design (CASD)**: The design for synthesis of an organic compound in consideration with the organic reactions, available starting materials, and economic effects (Molidor et al. 2003).
5. **Drug Design**: There is an increasing demand for development of a new drug in less time and at a minimal cost. Experimental methods like combinatorial chemistry and high-throughput docking, which in turn yield vast quantities of data for study, are also used in the drug design process (see Sect. 3.3.7 for details). Therefore, there is no question that the field of drug design is the most significant area of chemoinformatics at present (Ilyin et al. 2003).

3.2.6 Molecular Phylogeny

Phylogenetic analysis is a method to decipher the evolutionary history and relationship among a group of organisms. Phylogenetic trees are commonly constructed to study the evolutionary relationship among species and is the most important feature of phylogenetic analysis, which itself is evolving with the advancements in computer science. Molecular phylogenetics is the study of genes and other biological macromolecules "evolutionary relationships by analyzing mutations at several places in their sequences and forming hypotheses about the biomolecules" evolutionary relationships. They can serve as molecular fossils because genes are the medium for documenting the accumulated mutations. The evolutionary history of genes and even animals can be revealed by comparative assessment of molecular fossils from a variety of related animals. With the increase in availability of methods

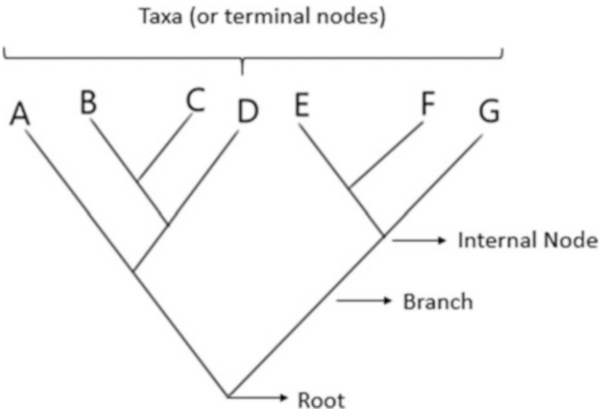


Fig. 3.3 Basic construct of a phylogenetic tree

and programs for phylogenetic tree construction, molecular phylogeny has become more popular. This is since molecular data are more abundant and easier to collect than fossil records, there is no sampling bias involved, which helps to mend the holes in real fossil records, with this data it is possible to create simpler and more accurate phylogenetic trees (Horiike 2016).

Based on the sequence similarity of molecules, such as DNA, RNA, proteins, it is also possible to infer evolutionary relationships between species. These similarities can be detected by multiple sequence alignment (MSA) through programs like Clustal W, Clustal X, and homology search (homologues are sequences that have common ancestry) through tools like BLAST (basic local alignment search tool) (Altschul et al. 1990). This is followed by use of methods to construct and interpret phylogenetic trees such as unweighted pair group method with arithmetic mean (UPGMA), neighbor-joining (NJ), maximum parsimony, maximum likelihood, and Bayesian method (Horiike 2016). Figure 3.3 shows a typical bifurcating phylogenetic tree that can be constructed for phylogenetic analysis.

Tree construction methods are classified into two groups. Primarily is the distance-based method that uses evolutionary distance matrix. UPGMA and NJ methods are the representative methods which make use of computational tools such as MEGA7 (Molecular Evolutionary Genetics Analysis version 7) and PHYLIP (PHYLogeny Inference Package), CLUSTALX, respectively. The advantage of the distance-based method is its short calculation time that allows handling of large amount of data. Another is the character-based approach, which uses the aligned sequences directly through tree inference. Maximum parsimony, Maximum likelihood, and Bayes method are the representative methods which use PHYLIP, MEGA7, PhyML, etc. as named in Table 3.1 (Horiike 2016).

Table 3.1 Table showing list of methods for infering phylogenetic trees (Horiike 2016)

Method	Group	Algorithm	Software
UPGMA	Distance matrix	Clustering for the shortest evolutionary distance	MEGA7
Neighbor-joining	Distance matrix	Clustering for minimum total branch distance	PHYLIP, Clustal X, MEGA 7
Maximum parsimony	Character-based	Searching tree with minimum total number of character-state changes	PHYLIP, MEGA 7
Maximum likelihood	Character-based	Searching tree with maximum likelihood	PHYLIP, PhyML, RAxML, FastTree, MEGA 7, TOPALi v2
Bayesian	Character-based	Searching tree with maximum posterior probability	MrBayes, TOPALi v2

3.2.7 Systems Biology

Systems biology is an approach in biomedical research to study the biological systems, involving interactions of the individual components of biological entities, such as molecules, cells, and organs. This is in unambiguous contrast with the reductionist strategy that has dominated biology for a long time (Wanjek 2011) and emphasizes that more than the sum of its components are the networks comprising entire living organisms. There are collaborative, integrative approaches between different scientific disciplines, such as computer science, biology, engineering, bioinformatics, physics, and others to predict how these processes develop over time and under varying conditions.

Through such hypothesis-driven research, the advent of sequencing and other high-throughput technologies has sparked the creation of new ways to develop solutions to the world's most pressing health and environmental challenges. Thus, the science of systems biology is focused on simulated computational and mathematical models of biological systems or processes (Raman and Chandra 2010). The most commonly studied models are the metabolic networks. For example, the model organism *Escherichia coli* has well known metabolic reactions, enzymes, cofactors, substrates, and products (Feist et al. 2007). However, this is the first step in understanding how these components work in spatial and temporal integration, and what the controls exercised on them are exactly. Although metabolic network topologies are well understood, the interactions that regulate this metabolism have yet to be explained (Gianchandani et al. 2006; Grüning et al. 2010), thereby, emphasizing the importance of metabolic networks in systems biology.

The creation of the Systems Biology Markup Language (SBML) has resulted in the need for successful sharing of formal, quantitative systems biology models (Hucka et al. 2004). Because biochemical network studies are a particularly successful area of systems biology, a variety of computational tools have been developed that address different needs in biochemical network analysis. The Systems Biology Workbench (SWB) is a set of systems biology tools, which includes biochemical network building, viewing, and editing programs, simulation tools, and model import and translation tools. CellDesigner, a Java-based software for the creation and editing of biochemical networks, is another extremely useful method (Funahashi et al. 2003).

Further, bioinformatics applications for research in systems biology use applications to visualize network architectures and overlay virtual and experimental data on the network. These tools include yEd graph editor for network editing and tools such as Cytoscape (Shannon et al. 2003) and Pathway Tools cellular overview diagram and Omics Viewer (Paley and Karp 2006) for visualization of "omics" data in the form of biochemical networks. The ability to develop predictive, multi-scale models helps our researchers to classify new disease biomarkers, stratify patients on the basis of specific genome profiles, target drugs, and other therapies. Moreover, the biology of structures provides the capacity for entirely new ways of assessment and innovation in biotechnology and computer science.

3.2.8 Synthetic Biology

Synthetic biology is another interdisciplinary territory that includes the utilization of engineering standards to biology. It focuses on the restructure and creation of biological components and frameworks that do not as of now exist in the normal world. Engineered science consolidates synthetic DNA synthesis with developing information on genomics to empower scientists to rapidly fabricate DNA sequences and form new genomes. Modified bacterial genomes have been synthesized and utilized in the creation of cutting edge biofuels, bio-items, renewable chemicals, bio-based chemicals (pharmaceutical intermediates, fine synthetic compounds, food ingredients), and in the medicinal services segment as well. However, systems biology studies complicated natural biological systems as it included modeling, simulation, and synthetic biology studies.

Progress in synthetic biology is empowered by ground-breaking bioinformatics devices that enable the design, construction, and test phases of the bioengineering cycle to be integrated. For the DESIGN and BUILD phases, bioinformatics tools provide tools for the discovery, synthesis, assembly, and optimization of components (enzymes and regulatory elements), devices (pathways), and structures (chassis). TEST methods include those for sampling, detection, and quantification of rapid prototyping metabolites (Carbonell et al. 2016). This includes tools such as the antiSMASH software (Weber et al. 2015) that recognizes and analyzes biosynthetic genomic regions in sequenced microbial genomes coding for natural products. The CanOE Strategy (Smith et al. 2012) and the Enzyme Function Initiative are methods for automated enzyme function annotation and prediction (Zhao et al. 2013; Gerlt et al. 2011). Possible biosynthetic routes can be predicted by an accumulation of pathway designing tools like BNICE and SimZyme (Moura et al. 2016).

3.3 Research Areas of Bioinformatics

3.3.1 Development of Biological Database

As biological data accumulates on a large scale and grows at an unprecedented rate, the construction of databases has become a crucial task in bioinformatics. In order to enable data retrieval and visualization, not only do biological databases store, organize, and distribute information in an organized and searchable way, but they also have computerized web application programming interfaces (APIs) to share and incorporate information from various database tools (Zou et al. 2015). As per a 2014 Molecular Biology Database Collection report published in the Nucleic Acids Research journal, a total of 1552 publicly available online databases exist (Fernandez-Suarez et al. 2014). Biological databases may roughly be classified into primary and secondary/derivative databases according to the degree of data curation. Primary databases, such as the NCBI Sequence Read Archive (SRA) (Kodama et al. 2012), constitute raw data as an archival repository, while secondary or derivative databases constitute curated or processed data, such as NCBI RefSeq (Pruitt et al. 2014). Table 3.2 gives some examples of important databases.

3.3.2 Sequence Analysis

Sequence analysis broadly represents computational evaluation of a DNA, RNA, or protein sequence, to mine information about its properties, such as biological function, structure, and evolution (Prjibelski et al. 2019). Sequence alignment is commonly used and is invaluable for biological sequence analysis and comparison. This method involves comparing two or more nucleotide sequences (DNA or RNA) or amino acid sequences by searching for a number of distinctive characters or patterns structured in them (Manohar and Shailendra 2012; Junqueira et al. 2014). The structure and function of a novel sequence can be easily predicted by doing sequence alignment.

There are two types of sequence alignment that can be performed using in silico tools such as comparing two (pairwise) or many sequences (multiple) for a string of characters. Alignment of three or more nucleotides or protein sequences refers to multiple sequence alignment. The genes which are similar are the ones that may be conserved among different species (Troy et al. 2001). Software tools and web services are often used for carrying out sequence analysis. BLAST (Altschul et al. 1990) is one of the most commonly used programs for sequence analysis based on pairwise sequence alignment. It carries out alignment as well as provides statistical information about the alignment. CLUSTAL W is a multiple sequence alignment (MSA) program for aligning homologous nucleotide or protein sequences (Thompson et al. 1997).

Table 3.2 Important human-related biological databases

Type of database	Examples
DNA	a) GenBank (USA), EMBL (Europe), DDBJ (Japan)—collection of all publicly available DNA sequences b) NCBI RefSeq—reference genome c) dbSNP—profiling of human genetic variation
RNA	a) RNAcentral—noncoding RNA sequence data b) miRBase—microRNA database
Protein	a) UniProt—collection of universal proteins b) PDB—primary database for 3D structures of biological macromolecules c) Pfam—identification of protein families and domains
Expression databases	a) GEO (Gene Expression Omnibus)—archive of gene expression data b) Human Protein Atlas—profiling expression information based on both RNA and protein data c) TiGER—tissue-specific gene expression and regulation
Enzyme database	a) ExPASy—Enzyme Nomenclature Database b) REBASE—The Restriction Enzyme Database
Pathway databases	a) KEGG PATHWAY—curated biological pathway resource on the molecular interaction and reaction networks b) MetaCyc—metabolic pathway database c) BioSilico—an integrated metabolic database d) BioCyc—pathway/genome database and software tools
Disease databases	a) HGMD—Human Gene Mutation Database b) CADgene—Coronary Artery Disease gene database c) ICGC—International Cancer Genome Consortium d) OMIM—Online Mendelian Inheritance in Man e) OMIA—Online Mendelian Inheritance in Animal
Literature	a) PubMed—database of biomedical literature from MEDLINE b) PubMed Central (PMC)—free, full-text literature archive
Chemical	a) PubChem—database of chemical compound from NCBI b) ChEBI—database of chemical compound from EMBL-EBI

3.3.3 Genome Analysis

Annotation of biological data is its main descriptive aspect. It refers to a textual representation of the biology associated with the data (Bell et al. 2012). Analysis of the genome requires DNA annotation, i.e. the method of specifying gene locations and coding regions in a genome to generate insights about the potential functions of the genes. Many resources and projects produce computational annotations to handle the exponentially increasing amount of biological sequence data because manually curated annotation is labor-intensive, time-consuming, and expensive (Boeckmann et al. 2003). In silico, gene prediction is relatively easy for the prokaryotes since all the genes are transcribed into the corresponding mRNA, followed by translation into proteins. For eukaryotic cells, however, it is more challenging because the coding DNA sequence is interrupted by introns (noncoding regions).

Bioinformatics has emerged as essential advantage for the various branch of biological sciences such as genomics, transcriptomics, proteomics, and metabolomics. It has so many types of biological databases and software tools. It is widely accessible through World Wide Web (WWW) (Teufel et al. 2006). Following are some important bioinformatics tools and software for genome analysis:

a) **Genomic Databases**: The sequence data generated is stored in large genomic databases/archives, the most frequently used of which are the European Molecular Biology Laboratory (EMBL)/European Bioinformatics Institute (EBI) database, the National Center for Biotechnology Information (NCBI-GenBank), and the DNA Data Bank of Japan (DDBJ).
b) **Genome Browsers**: In an attempt to provide easy access to sequence data, web-accessible tools, called genome browsers, have been created. Currently, the most widely used browsers are the Entrez Gene browser, the UCSC genome browser, and the EBI/Ensembl browser.
c) **Sequence Alignment**: For details, refer to Sect. 3.3.2.
d) **Ab-initio Gene Prediction**: It is a method in which genomic DNA is systematically searched for potential coding genes based on signal detection, which indicates the presence of coding regions in the vicinity and prediction is based on sequence information only. It can detect new genes with no similarity to known sequences or domains. Examples of ab-initio gene prediction programs include GENESCAN (Burge and Karlin 1997) and AUGUSTUS (Stanke et al. 2004).
e) **Expression Profiling**: For details, refer to Sect. 3.2.2.
f) **Promoter Prediction**: The promoter region is central to the regulation of the level of expression of a gene. PromoterScan (Prestridge 1995) has been used as one of the first tools with satisfactorily high precision for promoter prediction. Further progress has been made recently by PromoterInspector (Scherf et al. 2000) and Dragon Promoter Finder (Bajic et al. 2002) in the accuracy and sensitivity of algorithms for promoter prediction.

3.3.4 Three-dimensional (3D) Structure Prediction

Protein is one of the most complicated macromolecules of living organisms. Different amino acid protein sequences form different spatial shapes and structures that contribute to different cellular biological functionalities. Anfinsen in 1973 (Anfinsen, 1973) showed that in a protein's amino acid sequence, all the knowledge it requires to fold properly is encoded (called the dogma of Anfinsen). Although the protein folding mechanism is controlled by different physical rules, it is not simple to provide an accurate physical description of such complex macromolecule (including its interaction with surrounding solvent molecules) (Deng et al. 2018). This is where bioinformatics comes into play.

The bioinformatics tools for structure prediction mostly work on template-based structure prediction methods like homology modeling (based on sequence comparison) and threading methods (based on fold-recognition) (Huang et al. 2014). Tools like SWISS-MODEL, Modeller, JPred, 3D JIGSAW are some commonly used structure prediction tools for 3D modeling (Refer Chap. 2, Sect. 6). To date, many methods of prediction are very advanced and widely used by biologists, but there are still some challenges to address. Most of the prediction methods used today (including template-free methods) rely on the structural knowledge that is known, which is not too desirable. The latest advance in the prediction of protein structures can be predicted by reducing dependence on recognized structures and improving first-principle research (Deng et al. 2018).

3.3.5 Clinical Applications

The clinical application of bioinformatics is given a term called "Clinical Bioinformatics," which deals with associated sciences and technologies to study molecular mechanisms and probable therapies for human diseases (Wang and Liotta 2011). The term for the development of disease-specific biomarkers and individualized medicine is relatively recent and is the result of a convergence of the sciences of clinical informatics, bioinformatics, medical informatics, IT, mathematics, and omics. In addition, it also comprises others such as biomarker discovery and development, human tissue bank, mathematical medicine, and pharmacomics. The knowledge of clinical bioinformatics could also be instrumental in providing medical and biological information in personalized healthcare as it helps researchers in their daily medical practice for searching online biological databases with the help of bioinformatics.

It is possible to use bioinformatics in clinical diagnostics as well. The methods of bioinformatics can be used to identify the existence of genetic variations acting as markers for a disease or disorder. For example, genetic markers can be used to explain disease stratification in terms of symptoms and severity across populations in diseases like cystic fibrosis (caused by one of many distinct mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene located in chromosome 7), as well as to allow drugs to be targeted more effectively. Based on this, the CFTR gene has identified more than 1700 genetic variants for patients with cystic fibrosis.

3.3.6 Drug Discovery Research

Usually, drug discovery involves researchers discovering a target structure linked to a syndrome or illness in the human body, accompanied by screening for "primary" compounds that display affinity for the aim (Romano and Tatonetti 2019). The list of candidates is then narrowed down to identify the most promising leads that then go through the development process to test protection and efficacy in model organisms and, finally, in humans (Hughes et al. 2011). Data-driven drug finding practices data extracting on large data sources of candidate compounds and disease information to produce innovative therapeutic postulates rather than expecting for a single therapeutic assumption to deliver actionable results (Jorgensen 2004). The rising demand

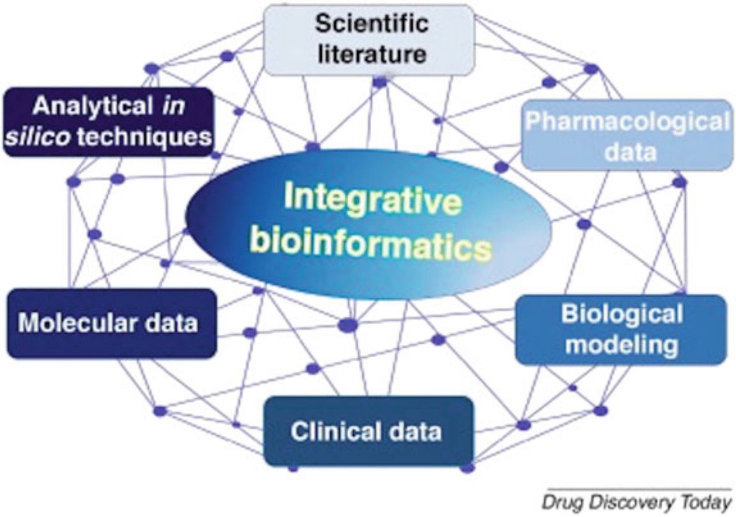


Fig. 3.4 Schematic representation of the integration of different data sets in bioinformatics (Ramharack and Soliman 2018. *Adapted with permission*)

to manufacture more and more drugs at low risk in a short period of time has led to a remarkable interest in bioinformatics (Ortega et al. 2012).

A computer-aided drug design (CADD) (Song et al. 2009; Cordeiro and Speck-Planche 2012) is a different field that is dedicated to bioinformatics-assisted drug designing. One of the key drivers of current bioinformatics approaches is the prediction and detection of biologically active candidates (Whittaker 2003) as well as mining and storing relevant information. The mining and storage of the human genome sequence by bioinformatics has facilitated to identify and categorize the nucleotide compositions of certain genes responsible for the coding of target proteins, in addition to identifying new targets that offer more potential for new drugs (Chen and Chen 2008; Katara et al. 2011). Another field where bioinformatics plays an important role is the approach of aim validation, as this supports to moderate the possible for failure in the phases of clinical testing and approval (Ratti and Trist 2001; Gilbert et al. 2003; Whittaker 2003). This also helps to ensure that during the approval procedure, more drug candidates are successful, making it additional cost-effective (Ortega et al. 2012). Moreover, bioinformatics will serve as an acceptable interface to provide pharmaceutical companies with new methods to opportunities to discover possible drug targets effectively and produce novel drugs (Whittaker 2003). Figure 3.4 shows the integration of biological data sets for better understanding of diseases to design more effective therapeutics (Ramharack and Soliman 2018).

3.3.7 Mathematical Modeling of Metabolic Processes

Mathematical modeling is used to explain inner and outer cell interactions and how they impact cell metabolism. In the investigation and simulation of phenotypes, this is calculated as metabolite concentrations and reaction fluxes over metabolic pathways, regulated by enzymes under different intrinsic and extrinsic conditions. It involves making an interpretation of cycles to mathematical problems with formal portrayals for an increased level of exactness and detail, since the objective is to arrive at the complexity and completeness of the behavior of a metabolic network. Depending on the situation in which the mechanism is analyzed, the two types of approaches to the numerical representation of biological processes vary. The stationary modeling takes into account the system operating at an equilibrium point where, after some time, the metabolite concentration is steady. Dynamic modeling, however, accepts the evolution of metabolite concentrations over time (Osvaldo et al. 2018).

Constructing models that make observable predictions of cell states over time is one of the most difficult tasks. This problem is currently addressed by new methods in silico, such as the reconstruction of dynamic models, the use of approaches to phenotype prediction, and the design of pathways through effective algorithms for strain optimization (Osvaldo et al. 2018). In biomedical science, systems biology and bioinformatics methods also help researching related data and properties (e.g. genome sequencing) to allow discoveries driven by modeling. This has also facilitated the development of genome-scale networks, the simulation of complex biological systems in silico, and the understanding of how metabolic flux distributions shift within a specific biological network to predict cellular phenotypes (McCloskey et al. 2013).

Furthermore, considered in many environmental and genetic situations, mathematical cellular metabolism modeling supports the tasks of metabolic engineering (ME) involving the design of appropriate strains, optimal gene deletion selection, or regulation of expression for the overproduction of compounds generated (Stephanopoulos et al. 1998; Burgard et al. 2003), thus, finding applications in the industrial sector as well.

3.4 Concluding Remarks

The high-throughput "omics" approaches to study biological samples such as genomics, transcriptomics, proteomics, and metabolomics have the ability to characterize all, or most, members of a family of molecules in a single analysis, aided by the use of bioinformatics tools and software. Originally developed for the analysis of biological sequences, bioinformatics now incorporates a wide range of subject areas including metagenomics, metatranscriptomics, metaproteomics, metabolomics, structural biology, systems and synthetic biology, metabolic and signaling pathways, high-throughput image analysis, gene expression studies, drug discovery, molecular genetics, and phylogenetic studies. Moreover, in silico tools and software are the prerequisites for quick sequence retrieval, efficient genome analysis, structure prediction, protein–protein interactions, rational drug designing, and phylogenetic analysis, thus having the potential to generate complete picture of an organism. This has opened wide avenues of research in biological sciences, as highlighted throughout the chapter. Therefore, in silico technology has revolutionized the way we look at basic scientific research, thereby opening doors for a variety of applications.

Competing Interests There is no competing interest.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Anderson NG, Anderson NL (1996) Twenty years of two-dimensional electrophoresis: past, present and future. Electrophoresis 17:443–453

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181 (4096):223–230

Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH (2017) Proteomics: technologies and their applications. J Chromatogr Sci 55(2):182–119

Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V (2002) Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics 18:198–199

Basuri TS, Meman AS (2011) Role of bioinformatics, cheminformatics and proteomic in biomarker identification and drug target validation in drug delivery processes. IJPSR 2(10):2521–2533

Bayat A (2002) Science, medicine, and the future: Bioinformatics. BMJ (Clin Res Ed) 324 (7344):1018–1022

Bell MJ, Gillespie CS, Swan D, Lord P (2012) An approach to describing and analysing bulk biological annotation quality: a case study using UniProtKB. Bioinformatics 28(18):i562–i568

Boeckmann B et al (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370

Burgard AP, Pharkya P, Maranas CD (2003) OptKnock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84:647–657

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW (2016) Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet 17:257–271

Carbonell P, Currin A, Jervis AJ, Rattray NJ, Swainston N, Yan C, Takano E, Breitling R (2016) Bioinformatics for the synthetic biology of natural products: integrating across the Design-Build-Test cycle. Nat Prod Rep 33(8):925–932

Chen YP, Chen F (2008) Identifying targets for drug discovery using bioinformatics. Expert Opin Ther Targ 12:383–389

Clish CB (2015) Metabolomics: an emerging but powerful tool for precision medicine. Cold Spring Harb Mol Case Stud 1(1):a000588

Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: lessons from large-scale biology. Science 300:286–290

Cordeiro MN, Speck-Planche A (2012) Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. Curr Top Med Chem 12(24):2703–2704

Cox J, Mann M (2007) Is proteomics the new genomics? Cell 130(3):395–398

Cozzolino SMF, Cominetti C (2013) Biochemical and physiological bases of nutrition in different stages of life in health and disease, 1st edn. Monole, São Paulo, Brazil

Cruz IBM, Taufer M, Schwanke CHA (2003) Genomics in the era of aging and its potential application in gerontology and geriatrics. In: Souza ACA (ed) Institute of Geriatrics and Gerontology PUCRS: the cradle of academic geriatrics in Brazil, 1st edn., pp 83–84

Dalmiel L, Vargas T, Molina AR (2012) Nutritional genomics for the characterization of the effect of bioactive molecules in lipid metabolism and related pathways. Electrophoresis 15:2266–2289

Dauncey MJ (2012) Recent advances in nutrition, genes and brain health. Proc Nutr Soc 71 (4):581–591

Deng H, Jia Y, Zhang Y (2018) Protein structure prediction. Int J Mod Phys B 32(18):1840009

Dennis G Jr, Sherman BT, Hosack DA et al (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4:P3

Desany B, Zhang Z (2004) Bioinformatics and cancer target discovery. DDT 9:18

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210

Feist AM, Henry CS, Reed JL et al (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol (3): article 121

Fernandez-Suarez XM, Rigden DJ, Galperin MY (2014) The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. Nucleic Acids Res 42:D1–D6

Fialho E, Moreno FS, Ong TPP (2008) Nutrition in the post-genomics: fundamentals and applications of omics tools. J Nutr 21(6):757–766

Funahashi A, Tanimura N, Morohashi M, Kitano H (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. BIOSILICO 1(1):159–162

Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, Cronan JE, Dunaway-Mariano D, Imker HJ, Jacobson MP, Minor W, Poulter CD, Raushel FM, Sali A, Shoichet BK, Sweedler JV (2011) The enzyme function initiative. Biochemistry 50:9950–9962

Gianchandani EP, Brautigan DL, Papin JA (2006) Systems analyses characterize integrated functions of biochemical networks. Trends Biochem Sci 31(5):284–291

Gilbert J, Henske P, Singh A (2003) Rebuilding big pharma's business model. In vivo Bus Med Rep 21:10

Graves PR, Haystead TA (2002) Molecular biologist's guide to proteomics. Microbiol Mol Biol Rev 66(1):39–63

Green ED, Watson JD, Collins FS (2015) Human Genome Project: twenty-five years of big biology. Nature 526:29–31

Grüning NM, Lehrach H, Ralser M (2010) Regulatory crosstalk of the metabolic network. Trends Biochem Sci 35(4):220–227

Horiike T (2016) An introduction to molecular phylogenetic analysis. J Agric Sci 4:36–45

Huang DW, Sherman BT, Tan Q et al (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8:R183

Huang YJP, Mao B, Aramini JM, Monteliono GT (2014) Assessment of template-based protein structure predictions in CASP10. Proteins 82(S2):43–56

Hucka M, Finney A, Bornstein BJ, Keating SM, Shapiro BE, Matthews J, Kovitz BL, Schilstra MJ, Funahashi A, Doyle JC, Kitano H (2004) Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. Syst Biol (Stevenage) 1(1):41–53

Hughes JP, Rees S, Kalindjian SB, Philpott KL (2011) Principles of early drug discovery. Br J Pharmacol 162:1239–1249

Ilyin SE, Pinhasov A, Vaidya AH, Amato FA, Kauffman J, Xin H, Gordon PA (2003) Emerging paradigms in applied bioinformatics. Biosilico 1:3

Jorgensen WL (2004) The many roles of computation in drug discovery. Science 303:1813–1818

Junqueira DM, Braun RL, Verli H (2014) Alinhamentos. In: Verli H (ed) Bioinformática da biologia à flexibilidade molecular. SBBq, São Paulo, pp 38–61

Katara P, Grover A, Kuntal H, Sharma V (2011) In silico prediction of drug targets in *Vibrio cholerae*. Protoplasma 248:799–804

Kaushik S, Kaushik S, Sharma D (2018) Functional genomics. Encyclopedia of Bioinformatics and Computational Biology 2:118–133

Kodama Y, Shumway M, Leinonen R (2012) International Nucleotide Sequence Database C. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40:D54–D56

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E et al (2015) ArrayExpress updated simplifying data submissions. Nucleic Acids Res 43:D1113–D1116

Kuehnbaum NL, Britz-McKibbin P (2013) New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era. Chem Rev 113:2437–2468

Kumar P, Periyasamy R, Das S, Neerukonda S, Mani I, Pandey KN (2014) All-trans retinoic acid and sodium butyrate enhance natriuretic peptide receptor a gene transcription: role of histone modification. Mol Pharmacol 85(6):946–957

Kusonmano K, Vongsangnak W, Chumnanpuen P (2016) Informatics for metabolomics. Adv Exp Med Biol 939:91–115

Lawrance SK, Smith CL, Srivastava R, Cantor CR, Weissman SM (1987) Megabase-scale mapping of the HLA gene complex by pulsed field gel electrophoresis. Science 235(4794):1387–1390

Liu B, Qian SB (2011) Translational regulation in nutrigenomics. Am Soc Nutr 2:511–519

Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T (2017) Transcriptomics technologies. PLoS Comput Biol 13(5):e1005457

Mahan LK, Stump SS (2005) Food, nutrition & diet therapy, 6th edn. Roca, Sao Paulo, Brazil

Manohar P, Shailendra S (2012) Protein sequence alignment: a review. World Appl Program 2:141–145

Maschio T, Kowalski T (2001) Bioinformatics – a patenting view. Trends Biotechnol 19:9

McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. Mol Syst Biol 9:661

Molidor R, Sturn A, Maurer M, Trajanoski Z (2003) New trends in bioinformatics: from genome sequence to personalized medicine. Exp Gerontol 38:1031–1036

Moura M, Finkle J, Stainbook S, Greene J, Broadbelt LJ, Tyo KE (2016) Evaluating enzymatic synthesis of small molecule drugs. J Metab Eng 33:138–147

Ortega SS, Cara LC, Salvador MK (2012) In silico pharmacology for a multidisciplinary drug discovery process. Drug Metabol Drug Interact 27:199–127

Osvaldo KD, Miguel R, Paulo M (2018) A review of dynamic modeling approaches and their application in computational strain optimization for metabolic engineering. Front Microbiol 9:1690

Paley SM, Karp PD (2006) The pathway tools cellular overview diagram and omics viewer. Nucleic Acids Res 34(13):3771–3778

Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA (2015) Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics 15 (5–6):930–949

Prestridge DS (1995) Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol 249:923–932

Prjibelski AD, Korobeynikov AI, Lapidus AL (2019) Encyclopedia of bioinformatics and computational biology 3:292–322

Proserpio V, Mahata B (2016) Single-cell technologies to study the immune system. Immunology 147:133–140

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O et al (2014) RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42:D756–D763

Raman K, Chandra N (2010) Systems biology. Resonance:131–153

Ramharack P, Soliman MES (2018) Bioinformatics-based tools in drug discovery: the cartography from single gene to integrative biological networks. Drug Discov Today 23(9):1658–1665

Ratti E, Trist D (2001) Continuing evolution of the drug discovery process. Pure Appl Chem 73:67–75

Romano JD, Tatonetti NP (2019) Informatics and computational methods in natural product drug discovery: a review and perspectives. Front Genet 10:368

Sales NM, Pelegrini PB, Goersch MC (2014) Nutrigenomics: definitions and advances of this new science. J Nutr Metab 202759

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463–5467

Sasaki AA, Fernandes GF, Rodrigues AM, Lima FM, Marini MM, Dos S Feitosa L, de Melo Teixeira M, Felipe MS, da Silveira JF, de Camargo ZP (2014) Chromosomal polymorphism in the Sporothrix schenckii complex. PLoS One 9(1):e86819

Scherf M, Klingenhoff A, Werner T (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol 297:599–606

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowskis B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504

Smith AAT, Belda E, Viari A, Medigue C, Vallenet D (2012) The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. PLoS Comput Biol 8:1–12

Song CM, Lim SJ, Tong JC (2009) Recent advances in computer-aided drug design. Brief Bioinform 10:579–591

Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32:W309–W312

Stephanopoulos GN, Aristidou AA, Nielsen J (1998) Metabolic engineering: principles and methodologies. Academic Press, San Diego, CA

Strassberger V, Fugmann T, Neri D, Roesli C (2010) Chemical proteomic and bioinformatic strategies for the identification and quantification of vascular antigens in cancer. J Proteomics 73:1954–1973

Subramanian U, Kumar P, Mani I, Chen D, Kessler I, Periyasamy R, Raghavaraju G, Pandey KN (2016) Retinoic acid and sodium butyrate suppress the cardiac expression of hypertrophic markers and proinflammatory mediators in Npr1 gene-disrupted haplotype mice. Physiol Genomics 48(7):477–490

Teufel A, Krupp M, Weinmann A, Galle PR (2006) Current bioinformatics tools in genomic biomedical research (Review). Int J Mol Med 17(6):967–973

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT, Cunningham P, Chamberlain AT, Sykes BC, Bradley DG (2001) Sequence-evolution - function: computational approaches in comparative genomics. Chapter 4: principles and methods of sequence analysis. Genetic evidence for near-eastern origins of European cattle. Nature 410:1091

Tsoka S, Ouzounis CA (2000) Recent developments and future directions in computational genomics. FEBS Lett 480:42–48

Vihinen M (2001) Bioinformatics in proteomics. Biomol Eng 18(5):241–248

Virkud YV, Kelly RS, Wood C, Lasky-Su JA (2019) The nuts and bolts of omics for the clinical allergist. Ann Allergy Asthma Immunol 123(6):558–563

Wang X, Liotta L (2011) Clinical bioinformatics: a new emerging science. J Clin Bioinform 1 (1):1–10

Wanjek C (2011) Systems biology as defined by NIH an intellectual resource for integrative biology. NIH Catalyst 19:6

Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL, Humphery- Smith I (1995) Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. Electrophoresis 16:1090–1094

Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH (2015) AntiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43:W237–W243

Whittaker P (2003) What is the relevance of Bioinformatics to pharmacology? Trend Pharmacol Sci 24:434–439

Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL (1995) Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev 13:19–50

Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, Bonanno JB, Hillerich BS, Seidel RD, Babbitt PC, Almo SC, Sweedler JV, Gerlt JA, Cronan JE, Jacobson MP (2013) Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature 502:698–702

Zou D, Ma L, Yu J, Zhang Z (2015) Biological databases for human research. Genom Proteom Bioinf 13(1):55–63

Online Resources

https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/what-genomics

https://www.who.int/genomics/geneticsVSgenomics/en/

http://www.scfbio-iitd.res.in/tutorial/genomeanalysis.html

https://medical-dictionary.thefreedictionary.com/Ab+Initio+Gene+Prediction

https://isbscience.org/about/what-is-systems-biology/

https://isbscience.org/about/what-is-systems-biology/

https://archive.bio.org/articles/synthetic-biology-explained

https://www.jli.edu.in/blog/applications-of-clinical-bioinformatics/

https://www.ddw-online.com/informatics/p323016-application-of-bioinformatics-in-support-of-precision-medicine.html

https://www.ddw-online.com/informatics/p323016-application-of-bioinformatics-in-support-of-precision-medicine.html

https://www.jli.edu.in/blog/applications-of-clinical-bioinformatics/

https://isbscience.org/about/what-is-systems-biology/

https://app.biorender.com/community/gallery/s-5eceae0e8b256400b59142f6-what-are-omics-sciences

https://biocyc.org/

Protein Analysis: From Sequence to Structure

4

Jaykumar Jani and Anju Pappachan

Abstract

Proteins are primary molecules that control most of the cellular processes. The sequence of a protein is linked to its structure which in turn is linked to its function. Understanding and integrating protein sequence, structure, and function information is necessary to address many challenging areas of Biology including protein engineering, structural biology, and drug discovery. Bioinformatics deals with protein sequences, structures, predictions, and analysis. Accessibility of these data and availability of high-throughput analysis tools will supplement experimental work in order to understand proteins better. Prediction of three-dimensional structures of proteins and studying the structural features are very necessary to understand various diseases and aid in disease diagnosis and drug discovery. In this chapter we discuss about various databases and *in silico* tools and methods related to protein sequence and structure analysis.

Keywords

Sequence · Protein structure prediction · Protein analysis · *In silico* analysis · Protein database · Homology modelling

4.1 Introduction

Proteins are the key players that control almost all activities which sustain living organisms. Even though the genome of an organism consists of information for survival, proteins are the versatile macromolecules that regulate virtually all life

J. Jani · A. Pappachan (✉)
School of Life Sciences, Central University of Gujarat, Gandhinagar, Gujarat, India
e-mail: jaykumar.jani@cug.ac.in; anju.p@cug.ac.in

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_4

processes within a cell. By providing structural and catalytic support proteins regulate various dynamic process of cells. Cytoskeletal proteins are examples of structural proteins that maintain cellular integrity and overall shape. Other proteins maintain cellular homeostasis by catalyzing various processes like DNA replication, transcription, translation, metabolism, cell communication and provide defence and immunity (Cohn 1939; Nelson et al. 2008). Defective proteins results in many disease conditions like Alzheimer's disease and sickle cell anaemia to name a few (Chou 2004). Study of proteins is of interest not only to biologists but also chemists because proteins are intriguing chemical entities and analysing their structures and how they carry out various functions are of prime importance. Detailed study of protein structure and function also helps to understand their molecular mechanism and role in various diseases. Throughout the kingdom of life from bacteria to higher eukaryotes, proteins are polypeptides made up of the same ubiquitous 20 amino acids. So, understanding the chemistry of amino acid is central to understand the molecular biochemistry of proteins. How amino acids are linked to one another through various kinds of covalent and non-covalent interactions give rise to proteins of varying structures, which can be grouped under distinct protein families that perform diverse functions (Nelson et al. 2008).

Study of proteins have traditionally been carried out using *in vitro* and *in vivo* techniques. But in modern protein chemistry, *in silico* studies are equally important. The wealth of sequence and structural data that has come as an outcome of the genome projects made it necessary for protein chemists to turn to the computers as laboratories to perform virtually various experiments in order to understand proteins better. Today, there are many bioinformatics tools and databases which help to correlate protein sequences with their structure and function. Identification of protein structure through conventional biophysical techniques like X-ray crystallography, NMR and Cryo-electron microscopy can be lengthy and complex which can be made easy with the development of structural bioinformatics which deals with prediction and analysis of the three-dimensional structure of bio-macromolecules (Marco 2009). An *in silico analysis* of protein sequence and structure can both complement and supplement experimental work.

The key challenge in bioinformatics is how to retrieve and analyse meaningful data and use it to enhance our understanding of biological molecules. Different protein databases store different pieces of information and address different aspects of protein analysis. In this chapter we will discuss some of the commonly used protein databases and tools available for protein sequence and structure analysis. We provide a flow chart on how to characterize proteins computationally starting from their sequence and proceeding to their structural analysis (Fig. 4.1). We also discuss some of the recent examples of how such *in silico* analysis is helping in the structure-based drug discovery and medical biology.

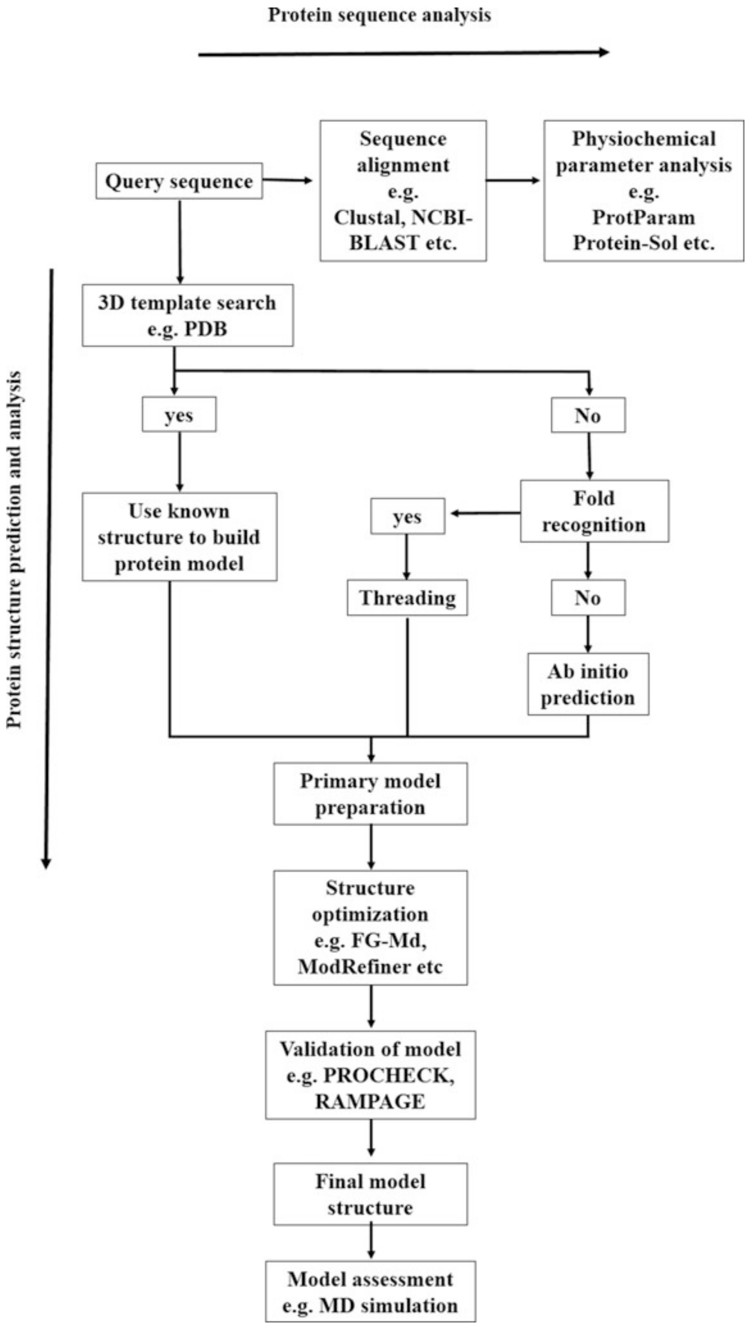

Protein sequence analysis
Query sequence
Sequence alignment e.g. Clustal, NCBI-BLAST etc.
Physiochemical parameter analysis e.g. ProtParam Protein-Sol etc.
Protein structure prediction and analysis
3D template search e.g. PDB
yes
No
Use known structure to build protein model
yes
Fold recognition
Threading
No
Ab initio prediction
Primary model preparation
Structure optimization e.g. FG-Md, ModRefiner etc
Validation of model e.g. PROCHECK, RAMPAGE
Final model structure
Model assessment e.g. MD simulation

Fig. 4.1 Flow chart on protein sequence analysis to structure prediction and analysis

4.2 Protein Structure Overview

Protein can be classified into different groups based on its structure, chemical nature, and biological role. Complex structural detail of protein can be studied at primary, secondary, tertiary, and quaternary levels of structural organization (Nelson et al. 2008).

4.2.1 Primary Structure

A linear linking of amino acids with each other as a chain via a peptide bond is represented as the primary structure of a protein. The polypeptide chain has an N-terminus and a C-terminus based on the presence of free amino or carboxyl group, respectively. The peptide bond is planar and rigid (non-rotatable) in nature as it partially shares two pairs of an electron. Whereas N-Cα and Cα-C have some freedom to rotate (−180 to +180), which helps proteins to acquire a three-dimensional structure (Nelson et al. 2008).

4.2.2 Secondary Structure

Local arrangement of some part of a polypeptide in particular conformation is referred to as protein secondary structure. The most common secondary structures are α-helices and β-strands, others are loops, turns, and coils. The secondary structure is mainly stabilized by Hydrogen bond (H-bond). The geometry has specific phi (φ) and psi (ψ) dihedral angle which can be studied by Ramachandran plot (Nelson et al. 2008).

4.2.2.1 Alpha(α) Helix

Alpha helix is most abundant in proteins compared to other helices (Kendrew et al. 1958; Pauling et al. 1951). Each helical turn is composed of 3.6 residues and has negative Φ and ψ angles ($\Phi = -64$ +/−7 and ψ −41+/−7). The α-helix repeats itself at every 0.54 nm, the radius of helices is 0.23 nm, and residue transition distance is 0.15 nm. The hydrogen bond between the nitrogen of amide of the 1st amino acid and carboxyl oxygen atom of 5th (i + 4) amino acid is a characteristic feature of α-helix (Pauling et al. 1951).

4.2.2.2 The β-strand

It shows extended conformation compared to α-helices. The distance between adjacent amino acids is around 3.5 Å. The pattern of H-bond in the β-strand can be parallel, antiparallel, or mixed type based on the direction of the strand from the amino to carboxyl terminal (Richardson 1977).

4.2.2.3 3_{10} Helices

Compared to regular α-helices this structure is found less frequently. The H-bond pattern is i + 3 instead of i + 4 in α-helices (Taylor 1941). The backbone dihedral angles (phi and psi) for 3_{10} helices are −49 and −26, respectively (Ramakrishnan and Ramachandran 1965).

4.2.2.4 β-turns

β-turns are irregular in shape and length, they connect two β-strands and help polypeptide to change the direction. β-turns are also known as reverse turns. β-turns are usually found on the surface of protein which enables them to interact with other proteins and molecules (Venkatachalam 1968).

4.2.3 Tertiary Structure

Tertiary structure is a three-dimensional arrangement of local secondary structure in a specific conformation. This structure is supported by various interactions including hydrogen bond, hydrophobic interaction, disulphide bridges, salt bridges, and Van der Waals interaction. Other than the various interactions post-translational modifications significantly contribute to protein folding. Based on current knowledge tertiary structure of the protein can be classified into three classes which are α-protein, β-protein, α+β-protein (Nelson et al. 2008).

4.2.4 Quaternary Structure

The quaternary structure represents complex interaction among polypeptide chain, this complex is made up of multiple polypeptide subunits but operates as a single functional unit. The subunits can be same or different. The overall structure is stabilized by hydrogen bonds, salt bridges, and various other intramolecular interactions (Nelson et al. 2008).

4.2.5 Domains, Motifs, and Folds

Consideration of various domains, motifs, and folds becomes very important while predicting the structure and function of the protein as they have evolutionarily conserved sequence and are mainly found in the active site of protein which is responsible for catalysis (Nelson et al. 2008).

4.2.5.1 Domain

It is a conserved part of the polypeptide chain and can individually form its three-dimensional structure irrespective of other domains in the protein. Even it can execute its function irrespective to rest of the protein. A single protein may have more than one domain. Chimeric proteins with desired activity can be generated through protein engineering utilizing domain swapping (Nelson et al. 2008).

4.2.5.2 Motifs

Motifs are conserved sequence of amino acids found among proteins having similar catalytic activity. One motif may have more than one secondary structure element, e.g. Helix turn helix (Nelson et al. 2008).

4.2.5.3 Fold

Folds are similar to motifs and represent general protein architecture. Proteins with the same folds show the same combinations of secondary structure (Nelson et al. 2008).

4.3 Classification of Proteins Based on Protein Folding Patterns

In order to obtain structure-based information retrieval, databases have been developed deriving information from the Protein Data Bank. Similarities in protein folding pattern has been used to organize and group proteins. The prominent structural classification databases used heavily by biologists to understand protein structure are SCOP and CATH (Ghoorah et al. 2015). Such classifications are useful because they reflect both structural and evolutionary relatedness.

4.3.1 CATH (Class, Architecture, Topology, Homology)

This database groups protein based on topology, homology, class, and architecture. The topology level classification clusters proteins based on the overall shape and secondary structure. Homology based classification groups protein by their sequence identity along with protein domain similarity shared with the ancestor. Class of protein is mainly determined by their secondary structure and fold pattern and includes; all α, all β, α-β, etc. The architecture of proteins represents an overall structure and shape of a protein generated by different secondary structure organization. Architecture level classification system groups protein based on its secondary structure arrangement in three-dimensional space (Ghoorah et al. 2015; Orengo et al. 1997).

4.3.2 SCOP (Structural Classification of Proteins)

SCOP is an open-access database created in 1994. This database maintained by MCR Laboratory of Molecular Biology UK was created with a purpose to provide evolutionary information and structural similarity between all proteins with known structure. The database organizes protein structures in a hierarchy starting from domains at the lowest level. Set of domains are classified into families of homologues. Families that share common structure and function are grouped into superfamilies. Superfamilies that share a common folding topology are grouped as folds. Each fold group may belong to one of the general classes—α, β, α + β, and small proteins which often have minimal secondary structures. This database classifies protein based on Family, Superfamily, fold, IUPR (Intrinsically Unstructured Protein Region), Classes, and protein type (Ghoorah et al. 2015; Murzin et al. 1995; Andreeva et al. 2020).

4.4 Commonly Used Databases to Retrieve Protein Sequence and Structure Information

There are various sequence, structure, and composite databases which provide different information regarding proteins. Sequence databases provide protein sequence information and structure databases like PDB provide three-dimensional structural information about protein, and the composite database integrates information from various primary databases. The different composite database uses different algorithms and criteria to yield diverse information on proteins (Chen et al. 2017). Table 4.1 gives a list of commonly used protein databases.

4.4.1 Commonly Used Protein Sequence Databases

Primary databases mainly consist of experimentally derived information, for example, protein sequence, structure, etc. Commonly used primary database for proteins is PIR (Chen et al. 2017).

4.4.1.1 Protein Information Resource (PIR)

PIR was established in 1984 with the purpose to support genomic, proteomic, and system biology research. The database was developed at the National Biomedical Research Foundation (NBRF). Initially, information was obtained and compiled from Atlas of protein sequence and structure published by Margaret Dayhoff.

Table 4.1 Web-links for protein databases

Tool name	Weblink
SCOP	http://scop.mrc-lmb.cam.ac.uk/
CATH	http://www.cathdb.info/
PIR	https://proteininformationresource.org/
Swiss-Prot/Uniprot	https://www.uniprot.org/
PROSITE	https://prosite.expasy.org/
PRINT	http://130.88.97.239/PRINTS/index.php
BRENDA	https://www.brenda-enzymes.org/
Pfam	https://pfam.xfam.org/
PDB	rcsb.org

Later in 2002 PIR, along with its international partner created a single worldwide database UniProt by combining PIR-PSD, Swiss-Prot, and TrEMBL (Wu et al. 2003).

4.4.2 Structure Database

4.4.2.1 PDB

Protein Data Bank is a repository of macromolecular structures experimentally deciphered by X-ray crystallography, NMR spectroscopy, and Cryo-EM all around the world. Initially, the database was created as a joint project by Cambridge Crystallographic Data Center, UK and Brookhaven National Laboratory, the USA in 1971. In 2003 the database becomes an international organization. Now there are four members which are PDBj, PDBe, Research Collaboration for Structural Bioinformatics (RCSB), and Biological Magnetic Resonance Data Bank (BMRB) who deal with data deposition, data processing, and distribution. The information submitted to the database is reviewed manually and computationally for its authenticity. Each submitted structure is given unique four letter accession ID called PDB ID. The database can be dug by protein name, PDB ID, author name, deposition date, etc. PDB also contains information regarding protein secondary structure, experimental procedure, experimental data, and ligand information. The protein structure coordinate file can be downloaded as a .pdb file and can be visualized using structure visualization software such as Pymol, VMD, Rasmol, etc. The main purpose of the database is to provide structural information of biologically important macromolecules. Further some secondary and curated databases utilize information from PDB to predict protein structure (Berman 2008).

4.4.3 Composite Databases

Composite databases utilize information from different primary and secondary databases and use a complex combination of computational algorithms in order to provide vital information like biological role, a conserved region of the protein, active site residue, signature sequence, etc. (Chen et al. 2017). Some of them are listed below:

4.4.3.1 Swiss-Prot

Swiss-Prot is designed by EMBL (European Molecular Biology Laboratory) and Department of Medical Biochemistry at University of Geneva collectively. In 2002, Swiss-Prot became UniProt Knowledgebase (UniProtKB) with supplement information from TrEMBL and PIR protein database. Today, UniProtKB provides detailed information about protein function, structure, post-translational modification, etc., with minimum redundancy (Bairoch and Apweiler 2000).

4.4.3.2 PROSITE

PROSITE is a secondary database that contains information about conserved motifs of proteins which relates to its biological function. Multiple sequence alignment (MSA) is performed by a database to provide information related to the query sequence. When a search is made for a new protein sequence in the database it gives two types of information. First, it gives information about sequence patterns and enlists other proteins with the same pattern. Second, it gives detail about the protein family and its denoted biological role (Hulo et al. 2006).

4.4.3.3 PRINT

This database classifies protein into different families based on protein fingerprints. Fingerprints are multiple small conserved motifs identified by sequence alignment. Motifs are not necessarily present in the contiguous sequence, but they might come together in 3D space upon protein folding, which defines active site or interacting site of the protein. Thus the study of fingerprint represents protein fold and function better than single motif (Attwood et al. 2000).

4.4.3.4 BRENDA (BRaunschweig ENzyme DAtabase)

BRENDA database is specifically for enzymes and its biological pathway. It gives information about the functional and molecular properties of enzymes that have been classified by IUBMB (International Union of Biochemistry and Molecular Biology). The information available in the database is obtained by manual extraction from literature, text mining, data mining, and computational prediction. Every enzyme classified in BRENDA contains information about its biochemical reaction and kinetic property such as substrate and product of the corresponding enzyme (Schomburg et al. 2002).

4.4.3.5 Pfam

Pfam is a protein family database. Entry in Pfam is classified as family, domain, repeats, and motifs. Search can be made using protein sequence, domain, keyword, or taxonomy. As a result, it provides Pfam annotations for domain architecture, sequence alignment, interaction with other proteins, and protein structure in PDB (Finn et al. 2014).

4.5 Protein Sequence Analysis

The sequence of the protein determines the structure and the function of proteins. A thorough analysis of the protein sequence will throw light on its biological role, active site, stability, post-translational modification sites, regulatory elements, etc. Today there are several databases and tools available which predict protein features based on its sequence composition.

4.5.1 Protein Sequence Alignment

Knowledge of residue to residue correspondence between sequences will help to understand patterns of conservation and variability among sequences and infer evolutionary relationships. Two or more protein sequences share similarity if they have evolved from a common ancestor. Sequence similarity beyond a certain threshold indicates that the proteins share a common structure and biological function. Alignment of multiple protein sequences helps to understand protein features which might appear non-significant in pairwise alignment. Patterns of amino acid conservation can give information on domains, active site, and distant relationships may be detected. In short sequence alignment tools permit the researcher to predict the function of gene and protein fastly and accurately *in silico* by comparing query sequence with previously characterized protein, which could not be easily possible manually in the laboratory (Chenna et al. 2003). For a meaningful analysis, multiple sequence alignment should have both closely and distantly-related sequences. Various sequence alignment tools based on different algorithms are available. Clustal maintained by EMBL-EBI is one of the widely used multiple sequence alignment tools (Do and Katoh 2008).

4.5.1.1 Clustal

Clustal includes a series of programs commonly used in bioinformatics for sequence alignment purposes. Originally the program was developed in 1988 and managed by EMBL-EBI. There are many versions of Clustal based on the development/upgradation of an algorithm, Clustal Omega is the current standard version. All versions of Clustal perform multiple sequence alignment from a series of pairwise alignments, and assess it on the basis of scores based on a scoring matrix. These values will be used by the algorithm for distance measurement which reflects the evolutionary distance between sequences and the tool can build a phylogenetic tree using the neighbour-joining approach (Chenna et al. 2003).

4.5.1.2 Sequence Alignment in Database Searching

When complete genomes were determined, in order to identify the unknown function of many proteins coded by the genome, databases can be searched to identify their homologues by sequence alignment. The most commonly used such tool by biologists all over the world is the NCBI BLAST (Basic Local Alignment Search Tool).

BLAST

BLAST is a fast, accurate, and most commonly used method worldwide to find sequence similarity between a query sequence and sequences available in the databases. The sequence is queried against a specified database, and produces a report of those proteins in the database that are related to the query sequence. BLAST provides different options for standard and specialized data mining. Standard BLAST includes BLASTP (protein query against a protein database), BLASTN (DNA nucleotide query against DNA database), TBLASTN (protein query against translated nucleotide sequence database), BLASTX (translated nucleotide sequence against protein database), PHI-BLAST (Pattern Hit Initiated-BLAST that finds homologous protein sequences which also contains a regular pattern), and PSI-BLAST (Position-Specific Iterated-BLAST). While specialized search includes SmartBLAST (to find protein having high similarity to query sequence), PrimerBLAST (to design primer specific to the template), GlobalAlign (to compare two sequences entirely), CD-Search (to find conserved domain architecture), IgBLAST (to search immunoglobulins and T-cell receptors sequence), MOLE-BLAST (to establish the taxonomy for uncultured or environmental sequences), etc. (Altschul et al. 1990; Madden et al. 2019).

MSAs contain patterns that characterize families of proteins. There are several methods for applying MSAs of known proteins to identify related sequences in database searches, important ones being Profiles, PSI-BLAST, and Hidden Markov Model (HMM). All the three methods are useful to identify distantly-related sequences in a database search. Profiles contain conserved patterns found in a MSA of a set of homologous sequences. These patterns can be used to identify other homologous proteins by matching the query sequences from the database against the sequences in the alignment table, with higher weight to the conserved positions than variable regions. PSI-BLAST a modification of BLAST starts with a normal BLAST, then derives pattern information from MSA of initial hits and reprobes the database using the pattern. This process is iterated, by refining the pattern in successive cycles. HMM is more powerful than the other two to find distant relatives and predicting protein folding patterns. These are computational structures for describing fine patterns that define homologous protein families (Mount 2009).

4.5.2 Physicochemical Parameters from Sequence Analysis

Understanding various physiochemical parameters of protein such as molecular weight, extinction coefficient, half-life, hydropathicity index, solubility, and isoelectric point (PI) is very essential to understand protein behaviour and function. Parameters such as solubility of protein affect protein folding, interaction with macromolecules and ligands. To design novel therapeutics and to optimize recombinant protein production this prediction becomes useful. Bioinformatics tools like ProtParam and Protein-Sol are generally used for computational prediction of physicochemical properties of proteins based on sequence information (Gasteiger et al. 2005; Hebditch et al. 2017).

4.5.2.1 ProtParam

ExPASy is a bioinformatics tool which provides access to the various database in the field of life sciences like proteomics, genomics, transcriptomics, population genetics, etc. This portal is operated by Swiss Institute of Bioinformatics (SIB). ProtParam is one of many tools available on the ExPASy server which calculates various parameters of protein which are given below (Gasteiger et al. 2005).

Molecular Weight

The molecular weight of a protein is calculated by adding the average isotopic mass of each amino acid in the sequence.

Theoretical PI

The isoelectric point (PI) of protein depends on the pKa value of amino acid. The pKa value depends on the side-chain composition of amino acid. However, the pH of the solution where protein is present significantly affects the PI and solubility of the protein.

Grand Average of Hydropathicity (GRAVY)

GRAVY index is used to represent the hydrophobicity of a given protein. It gives sum of hydropathy value of each amino acid in the sequence divided by total length of the protein. The positive and negative GRAVY value represents hydrophobic and hydrophilic nature of protein, respectively. This calculation is done by hydropathy values given by Kyte and Doolittle.

Half-life

This is the predicted time required for half of the protein to degrade after its synthesis in the cellular system.

Instability Index

This parameter represents the stability of a given amino acid sequence in the test tube. If the value is lower than 40 it is considered stable and if the value is greater than 40 it is considered as unstable.

Extinction Coefficient

The extinction coefficient represents the absorbance of light by a given medium at a particular wavelength. Experimentally this value can be calculated by using the reference of known amino acid sequence. Computationally it is predicted by analysing number of aromatic amino acids in a given amino acid sequence.

4.5.2.2 Protein–Sol

Protein–sol is an online open-access tool (http://protein-sol.manchester.ac.uk). This tool predicts solubility of a given amino acid sequence, the algorithm of the tool calculates 35 features of sequence which include twenty amino acid composition scores, seven other composites, protein length, folding propensity, disorder propensity, beta-strand propensities, Kyte-Doolittle hydropathy, PI, sequence entropy, and absolute charge. If the predicted solubility score is >0.45 then the protein is predicted to be soluble, if the value is <0.45 then solubility is less (Hebditch et al. 2017).

4.6 Protein Structure Prediction

4.6.1 Secondary Structure Prediction

Local secondary structure can be predicted by utilizing information of its amino acid sequence. It is the first crucial step to tertiary structure prediction. Available methods focus to identify conserved local secondary structures such as helices, strands, and turns. These structures form at the early stage of protein folding. Thus, understanding of protein secondary structure is essential to study the protein folding process also. There are many prediction methods available which use different algorithm for secondary structure prediction. The Chou–Fasman method was considered as a breakthrough method having almost 50–60% accuracy in prediction. However, recent methods have an improvised algorithm with an increased accuracy of up to 60–65% (Kabsch and Sander 1983). Apart from the use of amino acid sequence for secondary structure prediction, consideration of microenvironment of protein and solvent accessibility of protein improvises prediction.

4.6.1.1 Chou–Fasman Method

This is one of the earliest methods developed by Peter Y Chou and Gerad D Fasman in order to predict the secondary structure of a protein. This method is based on analysis derived from data generated by X-ray crystallography. It analyses the relative frequency of each amino acid to occur at a particular position in protein secondary structures. By studying verified data it was found that each amino acid has a certain propensity to prefer one secondary structure over other or a specific position in the secondary structure, e.g. proline and glycine are found at the end of the helix. Consideration of frequency of specific amino acid rather than available chemical and physical theories for structure prediction makes this method less accurate. Nevertheless, Chen *et al.,* in 2006 improvised this method which made it predict secondary structure more accurately (Chou and Fasman 1974).

4.6.1.2 (Garnier–Osguthorpe–Robson) GOR Method

It is an information theory-based method. In addition to Chou method, it considers the conditional probability of each amino acid to form a secondary structure to predict the location of secondary structure in a given sequence. The original method is more accurate in predicting α-helices than β-strands (Garnier et al. 1978). This method has approximately 65% accuracy (Mount and Mount 2001).

4.6.1.3 Neural Network-Based Method

JPRED, SPINE, PHD, and PSIPRED are neural network-based prediction methods. This method predicts helices and sheets with higher accuracy. The commonly used neural network-based methods use a two-layer neural network prediction approach. The first layer network utilizes sequence to structure approach where it predicts the secondary structure of a protein by considering central residue utilizing a position-specific scoring matrix (PSSM) or MSA. In the second layer, it uses structure to structure approach, and filters outs output from the first layer to generate a final structure with higher accuracy. The accuracy of the predicted structure by this method is up to 70% (Lin et al. 2005).

4.6.2 Protein Tertiary Structure Prediction

The most successful approach for predicting protein tertiary structure is the template-based homology modelling. It is based on the knowledge that homologous protein sequences fold into similar three-dimensional structures. The general criteria are that two sequences must be at least 25% identical to assume structural similarity between them. To predict the three-dimensional structure of a protein, homology modelling starts with doing a database search to identify its homologues whose structures are solved. Now this structure is used as a template to predict the unknown protein structure. Then their amino acid sequences are aligned and structurally conserved regions are assigned based on closely related amino acid sequences. The atomic coordinates of these regions are then used to construct a partial model of the unknown protein. Side chains that are different between the two proteins within these regions are replaced with the correct ones taken from suitable structure libraries. In this partial model, now the gaps are filled by loop searching and modelling of the loops. At the end of this process, a complete model with certain errors in bond length, bond angle, etc., may be obtained which has to be corrected by molecular mechanics and energy minimization (Marks et al. 2012).

The main problem in three-dimensional structure prediction is the calculation of free energy and obtaining structure with the globally lowest energy. Nevertheless, due to recent advancements in technology, several automated bioinformatics tools are now available to do this. Mainly two types of approaches are used for protein model structure preparation (1) template-based and (2) template independent. Both methods have their advantages and disadvantages (Marks et al. 2012; Kc 2017). However, template-based methods are more accurate than other methods (Kc 2017; Zhang and Skolnick 2004). Few commonly used tools are explained below. Individuals can access the CAMEO website (https://www.cameo3d.org/), which is an automated server to provide continuous assessment of protein structure prediction services in order to decide on a tool for protein structure prediction (Haas et al. 2018).

4.6.2.1 Template-based Method for Predicting Tertiary Structure of Proteins

SWISS-MODEL

SWISS-MODEL is a widely used modelling tool as it is fast, accurate, and user friendly. This server consists of three integrated compounds (1) SWISS-MODEL pipeline—contains software for database related to protein modelling (2) SWISS-MODEL Workspace—provides virtual workspace and handles complex tasks during model preparation (3) SWISS-MODEL Repository—provides updated information regarding 3-D protein model of model organisms. The structure prediction process by SWISS-MODEL consists of the following steps: template searching, target-template alignment, structure building, and last, evaluation of the model. For the template searching and alignment, it uses BLAST and HHblits. If the query sequence is identical to previously known structure, then it copies coordinate information from that and builds homologous structure. However, if the structure is non-identical or has a patch of the unaligned region, it builds structure from information available in the fragment library. The final model is evaluated by QMEAN, which is knowledge-based scoring, and given as output. An optimized model can be downloaded as a PDB file (Schwede et al. 2003; Waterhouse et al. 2018).

Modeller

Modeller was developed by Andrej Sali Laboratory at the University of California, San Francisco. This tool is used for tertiary and quaternary structure prediction. It derives important information about protein structure from experimental data generated by NMR spectroscopy, site-directed mutagenesis, fluorescence spectroscopy, image reconstructions from electron microscopic studies, etc. This information is utilized to understand various parameters such as bond length, bond angle, and dihedral angle in the protein model structure building. To build modelled structure MODELLER uses following sequential steps; (i) searching for the available evaluated structure related to the query sequence, (ii) alignment of query and template sequence, (iii) model preparation, and (iv) evaluation of the final model. The DOPE method is used for model evaluation. Other than model building it also performs fold assignment, phylogenetic tree preparation, and *de novo* modelling of protein loop (Webb and Sali 2016).

I-TASSER

Developed by Yang Zhang Lab, upgraded version of I-TASSER models structure using threading method. In order to generate a protein model from the query sequence, it performs multiple steps. First, it searches for a super secondary structure related to query in PDB, using multiple threading approaches also called LOMETS [50]. Then, the different fragments of the modelled structure are combined using the Monte Carlo method. Multiple models of protein having lower energy levels are generated using Replica Exchange Monte Carlo Simulation (REMC). Coordinates of all the models are clustered by SPICKER method and average values of coordinates from all models are taken further for model preparation. Lastly, FG-MD algorithm is used to reconstruct all the atoms of the model having low free energy states. As a final output, five full-length models with atomic resolution and estimated accuracy are shown up. If in case given template does not have any previously available homologous structure for modelling, then the structure is prepared from scratch using *ab initio*-based approach by QUARK tool. QUARK is an integral part of I-TASSER structure prediction pipeline but these steps are only used when domains in the template are <300 residues (Roy et al. 2010; Xu and Zhang 2012).

4.6.2.2 Template Free Method for Predicting Tertiary Structure of Proteins

Even though the template-based method is more accurate, the template free method is very crucial for proteins that do not have a satisfactory template or have novel fold (s). The limitation of this method is low accuracy of the force field and it requires a greater computational facility for a query having >150 residues. Differences between template based and template free method is that the template free approach utilizes the basic principles of protein folding and does not need a homologous structure. Therefore this method is capable to model novel proteins even with new folds (Kc 2017). Rosetta is one of the methods which performs template free structure prediction. It generates a full-length model based on 3–9 residues fragment available from the known structure. The fragments are selected based on sequence similarity. Monte Carlo method is used for the assembly of a different fragment to give rise to the final full-length structure (Rohl et al. 2004). QUARK is another fragment-based structure prediction tool that is developed by Yang Zhang Lab, it uses 1–20 residue fragments (Xu and Zhang 2012). These fragments are assembled by REMC and atomic-level knowledge-based force fields are used to generate the final model. Other template free structure prediction methods are FRAGFOLD (Jones 2001), SCRATCH (Cheng et al. 2005), etc.

4.6.3 CASP

The Critical Assessment of protein Structure Prediction (CASP) primarily helps in advancing the methods for protein 3-D structure prediction from the amino acid sequence. It provides an opportunity to research groups to test their structure prediction method and compare it with other available methods. CASPs performs worldwide experiments at an interval of every two years which critically evaluates the current state and progress in protein structure prediction and what is the future scope for development. Till now thirteen CASPs experiments have been performed, the assessment and result of each experiment was published in Proteins: Structure, Function, and Bioinformatics journal. These analyses help the individual researcher to choose appropriate structure prediction method for their research work (Kinch et al. 2019).

4.7 Evaluation, Refinement, and Analysis of Predicted Protein Structure

4.7.1 Evaluation of Predicted Structure

Evaluation of modelled protein structure is a common step performed to ensure that the predicted structure is closest to the original structure. This is done by studying stereo-chemical properties such as bond angle, torsion angle, bond length, and planarity of bonds. The G-factor is a measurement used to study how usual or unusual are the stereo-chemical parameters of given protein model. Lower the G-factor: lower the probability of a particular conformation (Wlodawer 2017).

4.7.2 Structure Refinement

Due to the limitation of a force field and all-atom reconstruction the quality of predicted structure may not be very good. So, the refinement of a predicted structure is a necessary step in protein structure prediction. The aim of refinement is to improve the model structure quality with minor improvement of coordinates in the backbone and side-chain atoms. Refinement will help to get a structure with high stereo-chemical quality which is nearer to the native structure. Potential energy minimization (PEM) techniques and molecular dynamics help to get a structure with lower energy. FG-MD is one of the methods which performs atomic-level molecular dynamics simulation to obtain a lower energy structure without much change in overall structure (Zhang et al. 2011). Mod-Refiner which is also used for structure refinement uses Monte Carlo simulation for energy minimization. This method usually refines backbone structure first, from the primary C_α traces. After refining the backbone at minimum energy, it performs another round of simulation to reconstruct side-chain atoms and gives a final refined model with lower minimum free energy. The refined model can be validated using Ramachandran analysis (Xu and Zhang 2011; Feig 2017) to see if there are stearic clashes between the atoms in the structure. PROCHECK (Laskowski et al. 1993), RAMPAGE, and Moleman2 (Kleywegt and Jones 1996) are extensively used online tools for structure validation.

4.7.3 Structure Analysis

4.7.3.1 Molecular Dynamics Simulation

MD simulations are efficient tools to effectively understand protein structure to function relationships. How proteins function require knowledge of structure as well as dynamics. Molecular dynamics simulations provide powerful tools for exploring the conformational energy landscape accessible to these molecules, Though the method was developed in the 1950s, with the advancement in computational facilities and MD algorithms, this technique has achieved time scales close to that of biological processes and has helped us to move from the analysis of single structures, to the analysis of conformational ensembles. Biologists mainly use this method to study the conformation dynamics of protein, refinement of protein structure, and to understand the interaction of the protein with other molecules. Structure prediction studies performed through MD simulation can be tested using a community-wide experiment in CASP. GROning Machine for Chemical Simulation (GROMACS) is one of the common open-access software packages used to perform simulation of proteins, lipids, and nucleic acids. Once a complex structure of a protein with the ligand is prepared by docking or obtained from PDB repository it can be used as an input file in GROMACS. By applying script code for the different force fields, the movement of the molecule over time can be created in MD run(s). The output of the simulation can be analysed and visualized in the supplemented tool provided in the MD package (Abraham et al. 2015; Hollingsworth and Dror 2018).

4.8 Protein Interaction Studies Using *In Silico* Methods

Proteins rarely act alone. For various metabolic and regulatory processes, they may be associated with ligands or nucleic acids or other proteins. Understanding the molecular and structural basis of these interactions is very necessary for the functional elucidation of the proteins. There are several *in silico* methods to predict and characterize protein–ligand/nucleic acid/other protein interacting sites. In order to predict interaction of protein with other molecules large number of available structural data are being utilized to develop and improvise available prediction algorithms. The empirical, force field, knowledge based, and machine learning are four scoring functions currently in use (Böhm 1994). These scoring functions use different approaches to calculate binding energy of protein with another molecule. SWISS Dock (Grosdidier et al. 2011) is a commonly used tool to study protein–ligand interaction. Manual docking and simulation studies are also helpful to understand these interactions.

4.8.1 Protein–Protein Interaction (PPI)

Interacting proteins are necessary for proper functioning of various cellular processes. There are several examples like proteinase-inhibitor complexes, antigen–antibody interactions, various signalling complexes, RNA polymerase assembly, etc. Experimental study of protein–protein interaction is costly and time-consuming, and such studies can be made easy computationally. Various computational tools are available for PPI prediction, primarily all tools utilize protein sequence information for analysis (Jones and Thornton 1997). Previously studied protein and structural information are useful to identify a surface patch of protein that may be found at the interface site. PPI interaction can be studied online using fully automated tools and offline by using manual docking software. These tools give information about binding geometry and binding energy (Kangueane and Nilofer 2018). Some of the available tools are PrISE, InterPreTS, iLoops, Struct2Net which are structure-based prediction tools. PPI spider, Path2PPI, POINeT, RedNemo are PPI network prediction tools. TRI_tool, HIVsemi, ChiPPI, InterPORC are model organism-based PPI prediction tools. STRING, SPRINT, HSPPIP, BindML+, and iFrag are other PPI prediction tools (Kangueane and Nilofer 2018; Rao et al. 2014).

4.8.2 Protein DNA Interaction

Protein DNA interactions are very important for the fundamental processes like DNA replication, transcription, and translation. Its importance in epigenetic regulation is also now well recognized.

Transcription factor and histone proteins are examples of protein with multiple substrate specificity which makes them difficult to learn. However, there are numerous bioinformatics tools which predicts DNA–protein integration. Mainly two approaches are used for this prediction: sequence-based and structure-based. The structure-based approach requires protein structure to predict interaction and a sequence-based approach utilizes previously available sequence information to predict interaction (Sarai and Kono 2005). Examples of such tools are DBS-PSSM, DBS-Pred, DISIS, DISPLAR, DP-Bind, BindN, FoldX, and DNAbinder (Sarai and Kono 2005; Si et al. 2015).

4.8.3 Protein–Carbohydrate Interaction

Protein–carbohydrate interactions play a crucial role in a biological system in processes of cell signalling, inflammation, host–pathogen interaction, cell adhesion, etc. Among all carbohydrate interacting proteins, antibodies and lectins are well characterized (Chandra et al. 2006; Sacchettini et al. 2001). We have very limited information about protein–carbohydrate interaction because carbohydrates are the very diverse molecules which can adopt a wide range of conformations. The information generated through protein crystallographic methods is a limitation as it gives a snapshot of only one particular conformation in which it was crystallized (Taherzadeh et al. 2016). BALLDock, SLICK, Vina-Carb, and PROCARB are commonly used tools for protein–carbohydrate interaction prediction (Taherzadeh et al. 2016; Malik et al. 2010; Kerzmann et al. 2006).

4.9 Applications of Protein Sequence and Structure Analysis in Drug Discovery

Earlier, novel drug discovery was either by chance or a trial and error process which is usually performed by a high throughput screening method. However, advancement in protein structure prediction and docking algorithms reduced the cost and time needed for this process. Bioinformatics helps in different aspects of drug discovery and development starting from target selection to prediction of a lead compound to its improvement. Protein sequence and structure analysis is important to select a potential drug target against a disease. Knowledge about multi-protein complexes makes it possible to target specific protein–protein interaction.

Even if the tertiary structure of a potential drug target may not be available, with a predicted protein structure, we can create a hypothesis about its function, interaction with other macromolecules, and its regulatory aspect in the biological system. For

Table 4.2 Drugs currently under clinical trial for COVID-19 treatment

Target	Antiviral treatment
RNA polymerase	Remdesivir, favipiravir, ribavirin, umifenovir, galidesivir, oseltamivir, sofosbuvir, methylcobalamin
3CL protease	Lopinavir/Ritonavir, Ivermectin
PL protease	Disulfiram
Protein S	Griffithsin
Miscellaneous	Resveratrol, Loperamide, Losartan, Chloroquine, Hydroxychloroquine

example, understanding of protein structure allows us to design site-directed mutations that alter its function or multimeric status (Takeda-Shitaka et al. 2004). A detailed study of the structures will help to design and test potential ligands and for selecting structural features for combinatorial synthesis of libraries.

The importance of protein structure prediction in drug discovery is evident during the current COVID-19 pandemic situation caused by novel coronavirus later denoted as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). To identify potential drug targets and drug candidates, researchers are using computational approaches to predict protein structure and carrying out docking and simulation studies to screen a range of drug candidates against this virus *in silico* (Elmezayen et al. 2020; Joshi et al. 2020; Narayanan and Nair 2020). The shortlisted molecules are being studied both *in vitro* and *in vivo* for future use. Considering this pandemic situation Zhang Lab has provided predicted 3D model structure and its functional annotation of the COVID-19 proteins coded by the genome of SARS-CoV-2 which can be directly used for docking and simulations studies (https://zhanglab.ccmb.med.umich.edu/COVID-19/). Many structural and non-structural proteins of the virus as well as host-based drug targets are studied *in silico* for druggability. Already studied antiviral drugs for various protein targets are listed in Table 4.2 (Gil et al. 2020). Most of the mentioned drugs are under clinical trial for COVID-19 disease (Huang et al. 2020)

Antibodies are very crucial protein molecules for both basic research and pharmaceutical applications. Atomic-level structural information is required to understand the molecular specificity of antibody which further illustrates its biological importance. Several computational tools are available which deals with different antibody feature predictions. For example, Fv modelling of antibody used to study paratope, epitope, and protein docking. These tools precisely give information about residues that are involved in antigen–antibody interaction. This information further utilizes to increase or decrease antigen–antibody interaction by mutation studies *in vitro*. SAbPred is an online server that contains multiple tools used to predict antibody structure and other features (Dunbar et al. 2016).

Another important group of proteins are membrane proteins which are challenging to crystallize. Approximately 25% of the total proteins in a cell are membrane proteins and yet there are only few structures available. Since crystallization of this protein is very difficult, protein structure modelling remains the next option for structural study. There are plenty of reports where the researchers have used modelled membrane protein structures to screen for various drugs (Becker et al. 2004; Hauser et al. 2018). Even there are dedicated tools for modelling of GPCR family proteins such as GPCR-SSFE 2.0, GPCRM, and GOMoDo (Worth et al. 2017; Miszta et al. 2018; Sandal et al. 2013).

4.10 Conclusion

There are several databases of protein sequence and structures which are not only repositories of validated and annotated data, but also provide several tools to analyse these data. Once a new protein is discovered, the biological function can be understood by sequence comparisons with homologous proteins because proteins with related functions have related amino acid sequences. Such comparisons also throw light on the evolution of these proteins. Families of proteins with related functions have evolved from a common ancestor. Such proteins will show similar three-dimensional structure too which means that the three-dimensional structure of an unknown protein can be predicted by homology modelling if a homologous structure is already known. Due to the tremendous advances in our knowledge of protein folding as well as machine learning tools and algorithms, protein structure prediction methods have improved significantly in the past decade. This has facilitated the prediction of model protein structure with greater accuracy and closer to the native structure. These protein structures can be further analysed to understand their structure–function relationships. One of the major applications of such studies is in drug discovery and development. However few challenges need to be addressed for future development such as modelling of multi-domain proteins, prediction of structure involving loop-mediated interactions, simulation of macromolecular complexes, better algorithms to understand protein folding, etc. With the advancement in computational facilities and development of powerful algorithms, such *in silico* analysis of protein sequences and structures can make tremendous impact on major challenges in biology.

References

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1–2:19–25

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andreeva A, Kulesha E, Gough J, Murzin AG (2020) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res 48:D376–D382

Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W (2000) PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Res 28:225–227

Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48

Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A, Kalid O, Bar-Haim S, Warshaviak D, Fichman M, Noiman S (2004) G protein-coupled receptors: In silico drug discovery in 3D. Proc Natl Acad Sci U S A 101:11304

Berman HM (2008) The protein data bank: a historical perspective. Acta Crystallogr A 64:88–95

Böhm HJ (1994) On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure. J Comput Aided Mol Des 8:623–632

Chandra NR, Kumar N, Jeyakani J, Singh DD, Gowda SB, Prathima MN (2006) Lectindb: a plant lectin database. Glycobiology 16:938–946

Chen C, Huang H, Wu CH (2017) Protein bioinformatics databases and resources. Methods Mol Biol (Clifton, N.J.) 1558:3–39

Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33:W72–W76

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31:3497–3500

Chou KC (2004) Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13:222–245

Cohn EJ (1939) Proteins as chemical substances and as biological components. Bull N Y Acad Med 15:639

Do CB, Katoh K (2008) Protein multiple sequence alignment. Methods Mol Biol 484:379–413

Dunbar J, Krawczyk K, Leem J, Marks C, Nowak J, Regep C, Georges G, Kelm S, Popovic B, Deane CM (2016) SAbPred: a structure-based antibody prediction server. Nucleic Acids Res 44: W474–W478

Elmezayen AD, Al-Obaidi A, Şahin AT, Yelekçi K (2020) Drug repurposing for coronavirus (COVID-19): in silico screening of known drugs against coronavirus 3CL hydrolase and protease enzymes. J Biomol Struct Dyn:1–13

Feig M (2017) Computational protein structure refinement: almost there, yet still so far to go. WIREs Comput Mol Sci 7:e1307

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120:97–120

Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. The proteomics protocols handbook. Springer

Ghoorah AW, Devignes M-D, Alborzi SZ, Smaïl-Tabbone M, Ritchie DW (2015) A structure-based classification and analysis of protein domain family binding sites and their interactions. Biology 4:327–343

Gil C, Ginex T, Maestro I, Nozal V, Barrado-Gil L, Cuesta-Geijo M, Urquiza J, Ramírez D, Alonso C, Campillo NE, Martinez A (2020) COVID-19: drug targets and potential treatments. J Med Chem

Grosdidier A, Zoete V, Michielin O (2011) SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res 39:W270–W277

Haas J, Barbato A, Behringer D, Studer G, Roth S, Bertoni M, Mostaguir K, Gumienny R, Schwede T (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins 86(Suppl 1):387–398

Hauser AS, Chavali S, Masuho I, Jahn LJ, Martemyanov KA, Gloriam DE, Babu MM (2018) Pharmacogenomics of GPCR drug targets. Cell 172:41–54.e19

Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J (2017) Protein-Sol: a web tool for predicting protein solubility from sequence. Bioinformatics (Oxford, England) 33:3098–3100

Hollingsworth SA, Dror RO (2018) Molecular dynamics simulation for all. Neuron 99:1129–1143

Huang X, Pearce R, Zhang Y (2020) De novo design of protein peptides to block association of the SARS-CoV-2 spike protein with human ACE2. Aging 12:11263

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJA (2006) The PROSITE database. Nucleic Acids Res 34:D227–D230

Jones DT (2001) Predicting novel protein folds by using FRAGFOLD. Proteins 45:127–132

Jones S, Thornton JM (1997) Analysis of protein-protein interaction sites using surface patches. J Mol Biol 272:121–132

Joshi T, Joshi T, Sharma P, Mathpal S, Pundir H, Bhatt V, Chandra S (2020) In silico screening of natural compounds against COVID-19 by targeting Mpro and ACE2 using molecular docking. Eur Rev Med Pharmacol Sci 24:4529–4536

Kabsch W, Sander C (1983) How good are predictions of protein secondary structure? FEBS Lett 155:179–182

Kangueane P, Nilofer C (2018) Protein-protein and domain-domain interactions. Springer

Kc DB (2017) Recent advances in sequence-based protein structure prediction. Brief Bioinform 18:1021–1032

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181:662–666

Kerzmann A, Neumann D, Kohlbacher O (2006) SLICK– scoring and energy functions for protein–carbohydrate interactions. J Chem Inf Model 46:1635–1642

Kinch LN, Kryshtafovych A, Monastyrskyy B, Grishin NV (2019) CASP13 target classification into tertiary structure prediction categories. Proteins Struct Funct Bioinform 87:1021–1036

Kleywegt GJ, Jones TA (1996) Phi/psi-chology: Ramachandran revisited. Structure 4:1395–1400

Laskowski RA, Macarthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291

Lin K, Simossis VA, Taylor WR, Heringa J (2005) A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21:152–159

Madden TL, Busby B, Ye J (2019) Reply to the paper: misunderstood parameters of NCBI BLAST impacts the correctness of bioinformatics workflows. Bioinformatics 35:2699–2700

Malik A, Firoz A, Jha V, Ahmad S (2010) PROCARB: a database of known and modelled carbohydrate-binding protein structures with sequence-based prediction tools. Adv Bioinform 2010

Marco W (2009) Structural bioinformatics: from the sequence to structure and function. Curr Bioinform 4:54–87

Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30:1072–1080

Miszta P, Pasznik P, Jakowiecki J, Sztyler A, Latek D, Filipek S (2018) GPCRM: a homology modeling web service with triple membrane-fitted quality assessment of GPCR models. Nucleic Acids Res 46:W387–W395

Mount DW (2009) Using hidden Markov models to align multiple sequences. Cold Spring Harb Protoc, 2009, pdb.top41

Mount DW, Mount DW (2001) Bioinformatics: sequence and genome analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540

Narayanan N, Nair DT (2020) Vitamin B12 may inhibit RNA-dependent-RNA polymerase activity of nsp12 from the SARS-CoV-2 virus. IUBMB Life

Nelson DL, Lehninger AL, Cox MM (2008) Lehninger principles of biochemistry. Macmillan

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH – a hierarchic classification of protein domain structures. Structure 5:1093–1109

Pauling L, Corey RB, Branson HR (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci 37:205

Ramakrishnan C, Ramachandran G (1965) Stereochemical criteria for polypeptide and protein chain conformations: II. Allowed conformations for a pair of peptide units. Biophys J 5:909–933

Rao VS, Srinivas K, Sujini GN, Kumar GNS (2014) Protein-protein interaction detection: methods and analysis. Int J Proteomics 2014:147648

Richardson JS (1977) β-Sheet topology and the relatedness of proteins. Nature 268:495–500

Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738

Sacchettini JC, Baum LG, Brewer CF (2001) Multivalent protein− carbohydrate interactions. a new paradigm for supermolecular assembly and signal transduction. Biochemistry 40:3009–3015

Sandal M, Duy TP, Cona M, Zung H, Carloni P, Musiani F, Giorgetti A (2013) GOMoDo: a GPCRs online modeling and docking webserver. PLoS ONE 8:e74092

Sarai A, Kono H (2005) Protein-DNA recognition patterns and predictions. Annu Rev Biophys Biomol Struct 34:379–398

Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res 30:47–49

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31:3381–3385

Si J, Zhao R, Wu R (2015) An overview of the prediction of protein DNA-binding sites. Int J Mol Sci 16:5194–5215

Taherzadeh G, Zhou Y, Liew AW-C, Yang Y (2016) Sequence-based prediction of protein–carbohydrate binding sites using support vector machines. J Chem Inf Model 56:2115–2122

Takeda-Shitaka M, Takaya D, Chiba C, Tanaka H, Umeyama H (2004) Protein structure prediction in structure based drug design. Curr Med Chem 11:551–558

Taylor HS (1941) Large molecules through atomic spectacles. Proc Am Philos Soc:1–12

Venkatachalam CM (1968) Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 6:1425–1436

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, De Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46:W296–W303

Webb B, Sali A (2016) Comparative protein structure modeling using MODELLER. Curr Protoc Bioinform 54:5.6.1–5.6.37

Wlodawer A (2017) Stereochemistry and validation of macromolecular structures. Methods Mol Biol 1607:595–610

Worth CL, Kreuchwig F, Tiemann JKS, Kreuchwig A, Ritschel M, Kleinau G, Hildebrand PW, Krause G (2017) GPCR-SSFE 2.0-a fragment-based molecular modeling web tool for Class A G-protein coupled receptors. Nucleic Acids Res 45:W408–w415

Wu CH, Yeh L-SL, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC (2003) The protein information resource. Nucleic Acids Res 31:345–347

Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 101:2525–2534

Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80:1715–1735

Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101:7594–7599

Zhang J, Liang Y, Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 19:1784–1795

5

Computational Evolutionary Biology

Subhamoy Banerjee

Abstract

Evolution is the dynamic process where a species or population undergo change in heritable characteristics. The study of evolution is called evolutionary biology. The role of computational tools and algorithms has become important for the study of evolutionary process. The key molecular aspect of evolution is sequence variation which is detected by comparing DNA or protein sequences. Different computational tools have been developed to align the obtained sequences and identification of sequence variation. Phylogenetics is a representation of similarity or dissimilarity among the species or genes or proteins. The variation of DNA sequence occurs by substitution of the bases and thereby it affects amino acid sequence. Evolutionary dating has become a crucial tool for estimation of species divergence. The application of evolutionary genomics is spanning from studying human evolution to the evolution of varieties of viruses. Many viruses pose serious threats to human health. Thus, studying viral evolution has become extremely important from biomedical aspect.

Keywords

Phylogenetics · Phylodynamics · Molecular clock · Bayesian analysis · Human evolution · Virus evolution

S. Banerjee (✉)
School of Life Sciences B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
e-mail: subhamoy.sls@crescent.education

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_5

5.1 Introduction

In the first half of twentieth century, evolutionary biology was largely considered as a field of analyzing fossils. But, after the discovery of DNA, few branches of biology like molecular biology, genetics has rapidly evolved with new perspectives. The microbial adaptation to environmental stress like antibiotic has been explained from the light of mutation. Evolutionary biology witnessed a paradigm shift from handling fossils to analyzing mutations in genes. During that time, another field of biology, namely bioinformatics has silently sprouted with a considerable application of computational algorithms. Algorithms and statistics have become an integral part of evolutionary biology for calculating the number of mutations in a given site to construct phylogenetic tree of a novel pathogenic virus. In the 1980s, world has witnessed Human immune virus (HIV) claiming millions of lives worldwide. Computational analysis of the gene sequences of HIV subtypes helped the scientist to understand the nature of the virus and its mutation rate. From the beginning of twentieth century, the world has witnessed several cases of epidemic and pandemic. In 2014, Sierra Leone of West Africa witnessed a serious viral outbreak with high mortality rate named Ebola. The zoonotic host of Ebola is bat. The genome analysis of the viral strains isolated from different patient samples and bat samples revealed that the outbreak started from single person who came in contact to Ebola infected bat (Futuyma and Kirkpatrick 2017). Since 2020, we are witnessing one of the worst pandemic in human history named Covid-19. Viral genome sequencing from the patient sample followed by computational analysis revealed that the possible source is bat and the nearest relative of SARS-CoV-2 (causative agent of Covid-19) is SARS. This information is produced by using different computational tools like multiple sequence alignment, phylogenetic tree construction, estimation of the time of the most recent common ancestor (tMRCA) using Bayesian analysis, etc.

Computational evolutionary biology is hugely applied in evolutionary genetics for analysis of the ancient genomes of human and other species, molecular anthropology, tracking spread of an infectious agent, genetic polymorphism detection, etc. The advancement of genome sequencing technology has yielded huge volume of sequence data. These sequences have opened a unique opportunity to the scientists to discover new insights in evolution. Also, the huge volume of data pose challenge to manage and analyze the data to extract meaningful information. With advancements of novel and high throughput sequencing techniques have opened many subfields in evolutionary biology like molecular anthropology, population genetics, archeogenetics, macro-evolution, etc. The analysis of sequencing data reveals information on evolutionary process, epidemiological influence, etc. Analyzing raw genome data yields various information like mammalian promoter architecture and evolution (Carninci et al. 2006), identifying evolutionary relationship between closely related and potentially hybridizing species like flycatcher (Nater et al. 2015), identification of mutations associated with clonal evolution of breast cancer (Wang et al. 2014), etc. Genome annotation helps us to identify protein coding genes and their respective structural and functional annotations. The functional annotation uses the method of identification of homologous proteins through finding best match in the database. In evolutionary point of view, homologous means descendant of common ancestor. Different genetic events like mutation (insertion or deletion), duplication, inversion, etc., which lead to change in sequence level which in turn influences the structural and functional alteration. The structural and functional alteration can be like change in active site or ligand recognition for enzymes and change in DNA binding domain in transcription factors which lead to change in expression level or target gene (Carninci et al. 2006). Homologues are subdivided into orthologs and paralogs. Orthologs are result of gene duplication event and share functional equivalence. Paralogs are the result of duplication event and used to have different functions. Phylogenetic analysis is used to study the paralogy event and functional annotation can be inferred from it. Neutral theory of evolution suggests functionally most significant sites come under most selection pressure which result in functional shift. Functional shifts are calculated by calculating non-synonymous substitution alone or the ratio of non-synonymous to synonymous substitution. The high value of the ratio indicates higher chance of origin of functionally different proteins. Different software packages have been developed to analyze the substitution rate and prediction method. Some of the software packages are PAML version 4 (phylogenetic analysis by maximum likelihood) (Yang 2007), PhyML (PHYlogenetic inferences using Maximum Likelihood) (Guindon et al. 2005), MrBayes (Huelsenbeck and Ronquist 2001), RAxML (randomized axelerated maximum likelihood) (Stamatakis 2014), etc.

PAML: This software package is a collection of programs for phylogenetic analysis of DNA and proteins. It has varieties of phylogenetic models to test different hypothesis like estimation of synonymous and non-synonymous rate of substitution of DNA when two sequences are compared. It is also used for reconstruction of ancestral genes and proteins, species divergence time calculation, etc. (Yang 2007).

PhyML: PhyML is another software which uses maximum likelihood method to construct phylogenetic tree. This algorithm follows "hill climbing" approach that adjusts both branch length and tree topology simultaneously, which needs few iterations to reach the optimum (Guindon and Gascuel 2003). The significantly less computing time is an advantage for this method.

MrBayes: This software package calculates the posterior probability of phylogenetic tree by using Bayesian phylogenetic inference. It uses Markov Chain Monte Carlo (MCMC) method to construct the phylogenetic tree. MrBayes 3.2 includes different methods like relaxed clock, dating, model averaging, etc. The output includes substitution ratio, ancestral states, branch length, etc. (Ronquist et al. 2012).

RAxML: It uses phylogenetic analysis of large datasets using maximum likelihood method. It uses general time reversible model and supports different types of dataset including RNA secondary structure data. It supports bootstrapping to optimize the phylogenetic tree construction. It also has the feature of analyzing next generation sequencing data (Stamatakis 2014).

Multiple sequence alignment (MSA) is a preliminary step for sequence data analysis. There are several tools available for MSA like MUSCLE, MAFFT, etc. MSA generally follows progressive alignment algorithm where a reference tree is generated followed by pairwise alignment by dynamic programming and MSA is determined.

MUSCLE (MUltiple Sequence Comparison by Log-Expectation): It is a widely used program. It is fast and able to handle partially matched dataset. It uses Markov model for distance measurement. Initially it calculates pairwise distance by counting k-mer frequency and performs progressive alignment. Further, reference tree is constructed using UPGMA or Neighbor joining method. Tree refinement is done using Markov model. The tree refinement terminates when the node length no longer decreases. Further, the tree is divided into two parts and MSA is performed and the result is realigned. If the sum of pair score of the realigned nodes becomes greater than previous value, then the realigned tree is further subdivided into two parts and the hole process is repeated until the score remains unchanged. This way, MUSCLE aligns multiple sequences (Edgar 2004).

MAFFT (Multiple Alignment using Fast Fourier Transform): It is a fast algorithm with ability to handle larger amino acid or DNA sequence. Given the speed of the process, the accuracy of the alignment is also found satisfactory and the iterative approach used in MAFFT was found consistently accurate (Nuin et al, 2006). It uses progressive alignment with Fast Fourier Transform for clustering purpose. It is able to rapidly identify homologous regions. Two different heuristics are implemented in MAFFT—progressive alignment and iterative alignment (Katoh et al. 2002). MAFFT has three steps, in which the first step is progressive alignment conducted using a shared 6-tuples followed guide tree construction using UPGMA method. The second step is recalculation of distance matrix and redoing the progressive alignment process and in the last step, iterative refinement is used to optimize weighted sum of pair score (Nuin et al. 2006).

5.2 Substitution Model

The substitution model is applied for either DNA or genome and amino acids in protein. Here, nucleotide and amino acid substitution model shall be separately discussed.

5.2.1 Nucleotide Substitution Model

The distance between two DNA sequences is defined as the number of substitutions per site and considering evolution rate is constant, the sequences should further diverge. Substitution models essentially calculate the number of substitutions per site and calculate the distance between two sequences. In simplistic way, the extent of sequence divergence is the ratio or proportion (p) of the nucleotide sites at which two nucleotides are differing from each other. So, $p = n_d/n$, where n_d is the number of nucleotides that are different between two nucleotides and n is the number of nucleotides screened, respectively (Nei and Kumar 2000). This is called p distance. As there are 4 different nucleotides, there could be 16 different possible pair can be generated. Out of these, four are identical which are AA, TT, CC and GG. Four nucleotide pairs undergo transition—AG, GA, TC and CT. There are eight possibilities of transversion—AT, TA, AC, CA, GT, TG, GC and CG. Theoretically, transversion should be twice frequent than transition but in reality, transition is found to be more frequent than transversion (Nei and Kumar 2000). The quantitation of nucleotide substitution is very important in order to estimate the number of nucleotide substitutions. There are different models that have been proposed for the purpose. The standard nucleotide substitution models are—Jukes–Cantor model, Kimura model, equal input model, Tamura model, HKY model, Tamura–Nei model, general reversible model and unrestricted model.

1. Jukes–Cantor Model: It is one of the simplest nucleotide substitution model. In this model, the base substitution frequency is considered to be same with a probability α (per unit time). Hence, the chance of a nucleotide to be substituted at any given site, $r = 3\alpha$. Considering two sequences X and Y diverged from a common ancestral sequence time t years before. Let us assume the proportion of identical nucleotide sequence is p_t and the ratio of different nucleotides is q_t, where $q_t = 1\text{-}p_t$. Now, the probability for identical and non-identical sequences at time t+1 is measured accordingly (Nei and Kumar 2000). So, for the unchanged nucleotides, the difference equation is

$$p_{t+1} = (1 - 2r)p_t + (2/3)r(1 - p_t)$$

which can be rewritten as $p_{t+1} - p_t = (2r/3) - (8r/3)p_t$

By converting the difference equation into continuous function, dp/dt, the equation becomes

$$dp/dt = (2r/3) - (8r/3)p_t$$

For solving non-synonymous nucleotides

$$V(d') = 9q(1\text{-}q)/(3\text{-}4q)^2 * n$$

where d′ of d can be obtained by observed value of q which is q′.

2. Kimura two parameter model: Transition type nucleotide substitution is considered to be more frequent than transversion type nucleotide substitution. Kimura proposed a model to estimate the total substitution. Total substitution rate, r per year (or per unit time) is $r = \alpha + 2\beta$. The transition frequency $P = ¼(1 - 2e^{-4(\alpha + \beta)t} + e^{-8\beta t})$ and transversion frequency $Q = 1/2(1\text{-}e^{-8\beta t})$, where t is the time after divergence. According to Kimura model, the frequency of each nucleotide is 0.25 at equilibrium. In this sense, this model is similar to Jukes–Cantor model and Kimura model is widely used (Kimura 1980; Nei and Kumar 2000).
3. Tajima and Nei model: This model proposed was proposed by Tajima and Nei in 1984. This method is more robust than many other models. The nucleotide frequency is considered constant in order to calculate the number of nucleotide substitution (d). d is estimated as

$$d = -b * \ln\left(1 - \frac{p}{b}\right)$$

where $b = 1/2\left[1 - \sum_{i=1}^{4} gi^2 + \frac{p^2}{c}\right]$ (Nei and Kumar 2000; Tajima and Nei 1984).

4. Tamura's model: In reality, nucleotide frequency does not become equal, on the contrary to some substitution models. Tamura proposed a substitution model where substitution frequency is different for transition and transversion and denoted as α and β, respectively. Also, AT and GC content is also considered different (θ_1 and θ_2, respectively) and the estimation of substitution is dependent on both the parameters. Here, the nucleotide substitution d is determined as

$$d = -hln\left(1 - \frac{P}{h} - Q\right) - 1/2(1 - h)\ln(1 - 2Q)$$

   where h = 2θ(1-θ)
   Tamura's method is able to compute different evolutionary parameters.

5. HKY model: It was proposed by Hasegawa, Kishino and Yano at 1985. It is based on Markov chain principle. It was proposed to estimate the divergence of mitochondrial DNA between primates and ungulates (Hasegawa et al. 1985). HKY model estimates genetic distances indirectly and it takes into consideration multiple changes in a site. Phylogenetic relationship was established using Maximum Likelihood method.
6. Tamura and Nei method: It was proposed by Tamura and Nei in 1993 (Tamura and Nei 1993). It considers different substitution values for transition and transversion. Transversion is assumed to have same frequency. It also differentiates two kinds of transition (purine → purine and pyrimidine → pyrimidine). For purine, $\alpha_{AG} = \alpha_{GA}$ and $\alpha_{transversion} = \alpha_1$; $\alpha_{CT} = \alpha_{TC} = \alpha_2$. In this model, P and Q are the transitional and transversional mutation frequency, respectively.
7. General time reversible model: GTR is probably the most popular substitution model in last one decade (Sumner et al. 2012). A Markov chain is called time reversible if

$$\boldsymbol{\pi_i q_{ij} = \pi_j q_{ji}} \qquad \textbf{for all, I ! = j}$$

$\boldsymbol{\pi_i}$ is the proportion of time the Markov chain spends at the *i*th state and $\boldsymbol{\pi_i q_{ij}}$ is the information flow $r =$ from *i*th to *j*th state. The rate matrix is written as a symmetrical matrix is multiplied by diagonal matrix.

$$\mathbf{Q}=\mathbf{q_{ij}}=\begin{bmatrix} . & a\pi C & b\pi A & c\pi G \\ a\pi T & . & d\pi A & e\pi G \\ b\pi T & d\pi C & . & f\pi G \\ c\pi T & e\pi T & f\pi A & . \end{bmatrix}$$

$$=\begin{bmatrix} . & a & b & c \\ a & . & d & e \\ b & d & . & f \\ c & e & f & . \end{bmatrix}\begin{bmatrix} \pi T & 0 & 0 & 0 \\ 0 & \pi A & 0 & 0 \\ 0 & 0 & \pi G & 0 \\ 0 & 0 & 0 & \pi A \end{bmatrix}$$

5.2.2 Amino Acid Substitution Model

From evolutionary point of view, studying amino acid substitution is more informative than nucleotide substitution because amino acid is more conserved than nucleotides and gives important insights into long term evolution. There are different statistical methods to measure distance between two amino acid sequences. The distance, also termed as evolutionary distance is extremely important for construction of phylogenetic tree and estimation of divergence time. One method is to measure the number of different amino acids (n_d) between two sequences. This method can be applicable if the sequence length of all the peptides is same, but it does not happen in reality (Nei and Kumar 2000). Introducing the gaps in multiple sequence alignment in order to show insertion-deletion (indel) mutation is a very common practice. There, direct measure of n_d is not possible. However, the ratio of number of differences over total number of amino acids (n_d/n) is a more meaningful approach. The ratio is known as p distance. In reality, p is not strictly proportional to time (t). One of the reason might be the multiple amino acid substitution at the same site. Poisson distribution gives a better estimate between the relation of p and t with the equation $p(k; t) = e^{-rt} * (rt)^k/k!$ where r is the rate of amino acid substitution per year at a given site. The rate is considered uniform in all the bases. It was found that this assumption does hold true in real life scenario. The observation suggests that the amino acid substitution rate is higher in functionally trivial sites compared to the important one. It was described that if the number of amino acid substitutions per site follows gamma distribution (Γ), it will become negative binomial distribution also (Nei and Kumar 2000).

So, $f(r) = \left(\frac{b^a}{\Gamma(a)}\right)e^{-br}r^{a-1}$

Where $a = \bar{r}^2/V(\bar{r})$ and $b = \bar{r}/V(\bar{r})$ and $\bar{r}$ is mean and $V(\bar{r})$ is the variance of r.

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} dt$$

Where a is the gamma parameter which determines the shape of the distribution and b is the scaling parameter.

5.3 Molecular Clock Estimation

Biological sequences as we see it is a product of evolutionary history and phylogenetic tree is a visual representation of this history (Haubold and Weihei 2006). The evolutionary diagrams are best represented by bifurcating tree. A tree has nodes and edges. Two important feature of a phylogenetic tree are (1) topology and (2) branch length. In topology, the nodes are arranged as per closely related or most similar sequences. This way, the branching is done by placing closest to furthest. Branch length is considered to be proportional to time. Knowing the mutation rate, the internal node can be drawn. Molecular clock hypothesis suggest that mutation rate of DNA and protein is constant over time. Also, it is hypothesized that the mutation rate is constant among different organisms, which implies, as they share common origin, the species divergence is directly proportional to time. Kimura had suggested genetic drift hypothesis in which he proposed that majority of the new mutation has no effect on the evolutionary fitness, as a result it either fixed within a population or spread through the population without any effect or randomly lost. The mutation fixed within a population termed as substitution rate and appearance of new mutation is known as mutation rate. Kimura showed substitution rate is equivalent to mutation rate, i.e. if mutation rate is constant in the species, then substitution rate shall also be constant throughout the tree of life. This is considered as strict molecular clock. But, in reality, mutation and substitution rates are variable throughout the course of evolution. That is why the concept of relaxed molecular clock is put forward which assumes the rate of events like mutation and substitution varies in a limited scale. However, calibrating the molecular clock is an important step irrespective of which model is adopted. The proper knowledge divergence is required to calibrate the clock so that it can be used for other events.

There are different nucleotide substitution models to estimate the amount of genetic changes. Some of the most commonly used models are generalized time reversible (GTR) model, Jukes–Cantor (JC) model, Hasegawa–Kishino–Yano (HKY) model. Substitution model has three components—substitution matrix that describes the relative rate of changes between nucleotides (Purine and Pyrimidine), the frequency of occurrence of each nucleotide, and the variation of substitution rates across the sites. GTR model considers different substitution rates for each nucleotides and different rates for transition and transversion. The estimation of time in the phylogenetic tree depends on the substitution model.

To handle the non-uniformity in rates among gene and lineage have led to various molecular clock models. These models are suitable to handle large data and different calibrations are available (Ho and Duchene 2014). The process of estimating

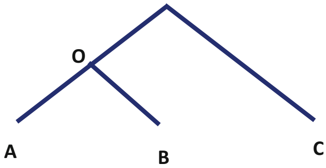


Fig. 5.1 Rooted tree for three species A, B, and C. O is the ancestral node

evolutionary time scale can be subdivided into five steps: (1) dataset assembly, (2) choice of calibration model, (3) selection of molecular clock method and model for rate variation, (4) analysis, and (5) interpretation (Sauquet 2013). Calibrating the molecular clock is very important and tricky part. The variation in rates can be divided into three components: gene effects, lineage effects, and gene-by-lineage effects (Ho and Duchene 2014). The rates can be expressed as substitutions per site per year or per million years. Gene effect or locus effect affects evolutionary rate across the genes. Lineage effect is the rate that varies across taxa but all genes are affected equally. Gene-by-lineage effects work in combination and gene specific rates vary among lineage. The classification of molecular clock can be done by number of distinct rates (k) across the number of branches (n) in the tree. For, strict molecular clock, $k = 1$ and for relaxed molecular clock, $n \geq k > 1$. Another type of molecular clock is local clock, where the assumption is evolutionary close relatives share same rate. So, the phylogenetic tree used to have distinct clusters of closely related lineages and each cluster used to have different rates. Random local clock model is implemented by phylogenetic tree construction software BEAST, where phylogeny, node time, and rate are estimated simultaneously. In this framework, Bayesian stochastic search is implemented to infer whether a branch inherits substitution from parental node by calculating posterior probability.

There are different tests of the molecular clock such as relative rate tests and likelihood ratio test. Relative rate test is the simplest of clock hypothesis test. Two species A and B evolve at the same rate relative to a third species C which is an out-group. Thus, all the rates measured here are relative to C, hence the name is relative rate test. In a hypothetical tree (Fig. 5.1), O is the ancestral node and distance of A and B, i.e., d_{OA} and d_{OB} should be equal. Likelihood ratio test of the clock is applied to any tree size. If there are s number of species, under the clock model of H_0, there are s-1 internal nodes for rooted tree. The number of substitutions per site is uniform.

As any model of molecular clock estimation works on a number of assumptions, Bayesian framework, by far, considered to be the best. It is based on Bayes theorem which states

$$P(\theta|D) = P(\theta)P(D|\theta)/P(D)$$

where P(θ|D) is the posterior probability, P(θ) is the prior probability, P(D|θ) is likelihood, and P(D) is the probability of the data. The likelihood value is estimated by likelihood model that includes substitution model and choice of specific tree. Prior distribution is estimated by evolutionary models. In this model, hyperparameters are used to estimate the priors, which is called hyperpriors. The probabilities of different parameters are then obtained from likelihood score and the posterior probability distribution is obtained. The probability of the data, P(D) is difficult to obtain analytically and it was circumvented by implementation of Markov Chain Monte Carlo (MCMC) method. It generates a huge number of sample data by stochastic method to determine posterior distribution in Bayesian phylogenetics. There are different algorithms for MCMC simulation and one of the most widely used algorithm is Metropolis–Hastings (Bromham et al. 2018; Hastings 1970; Metropolis et al. 1953). In this algorithm, one model tree is generated with a set of values assigned to the model parameters and likelihood is estimated. Then, it modifies one or more model parameters and recalculates the likelihood score and prior probability. Thus, it generates two related set of sequences separated by at least one parameter. The ratio of the posterior probability of these two alternate events is calculated. Likewise, depending on the ratio, the procedure moves on step by step. There are different software that use Bayesian model for molecular clock analysis, e.g. Bayesian Evolutionary Analysis Sampling Trees (BEAST), Molecular Evolutionary Genetics Analysis (MEGA), MrBayes, BayesPhylogenies, BayesTraits, etc.

1. BEAST: It is one of the most popular evolutionary phylogenetic and phylodynamic analysis software. BEAST analyzes the data using posterior distribution of the input data by using MCMC method. It can use sequence data of DNA (nucleotide), protein (amino acid), codon models, microsatellite, and SNPs (Bouckaert et al. 2019). The program is comprised of various standalone programs such as visualization tool BEAUti, Loganalyser, and Logcombiner to run MCMC analysis and logging and analysis, DensiTree, TRACER (for visualization and analysis of MCMC tree file), etc. The time calculation of phylogenetic tree is one of the most important feature of BEAST. The ancestral sequence obtained from phylogenetic tree used to have time dimension. But those sequences not necessarily exist at the node. The classic binary rooted time tree structure is important in some cases like population and transmission trees where branches represent entire species or population and branching event represent speciation, sampled ancestor or fossil sampling, structured population which is the tree branches are colored to represent specific subpopulation, identification of recombination event, species network, polytomies, i.e. individual gives rise to multiple lineages at the same time (Bouckaert et al. 2019). BEAST has been developed based on many seminal work on the topic and updated multiple times since 2007. The current version is BEAST v2.6.3.

Bayesian phylodynamic network construction needs a specific model for substitution, molecular clock estimation, population dynamic model generating tool. The site model that includes substitution model together with molecular clock model calculates likelihood data, P(D|T, θ). Model averaging and model comparison is done by bModelTest package. Molecular clock model in BEAST follows relaxed and random local clock model. FLC package is used to apply relaxed and random local clock model. Different tree models are used to construct phylodynamic model. There are two types of tree models: (a) model for unstructured population and (b) model for structured population.

a. Model for unstructured population: For unstructured tree model, two assumptions are used—first birth-death model where birth and death rate is assumed to build the tree. The second approach is coalescent model. Here, depending on the background population and changes in the effective size of the population, the parameterization is done.
b. Model for structured population: For structured population, the model is analogous to unstructured population models, i.e. two approaches are (1) multistate birth-death model and (2) structured coalescent approach. In multistate birth-death model, using bdmm package, BEAST v 2.5 can quantify parameter like migration rate from ancestral lineage without MCMC sampling. In structured coalescent model, MultiType Tree package can sample ancestral state of all lineages using MCMC sampling.

Overall, BEAST is a very important and robust tool ancestral sequence reconstruction and other phylodynamic analysis which include usage of molecular clock.

2. MEGA: It is another popular software with diverse application to homologous gene sequence analysis from multi-gene family and infer evolutionary relationships (Kumar et al. 2008). It is a desktop application with context dependent user friendly features. The input section has browse functionality to easily retrieve the input data. MEGA reformatted the data in MEGA format with . meg extension. Varieties of file types like clustal (.an), PAUP (.nexus), Phylip interleaved (.phylip), FASTA (.fasta), GCG (.gcg), PIR (.pir), NCBI xml format (. xml), etc., are supported and converted to MAGA format. Tree input format (phylogenetic tree) is Newick format. MEGA uses TreeExplorer to visualize and reconstruct the data. Editing and computing basic statistical parameters like codon frequencies, transition/transversion ratio, etc., can be achieved and users can specify substitution models for evolutionary analysis. Among other parameters, nucleotide pair frequencies, relative synonymous codon usage (RSCU) values, Disparity index, etc., can be computed by MEGA. User can choose among varieties of nucleotide substitution models like Jukes–Cantor, Tajima–Nei, Kimura 2 parameter, Tamura 3 parameter, Tamura–Nei model, etc. For amino acid substitution model, different models like no. of differences, p distance, equal input model, Poisson correction distance, PAM model, etc., are used. The other substitution model is synonymous and non-synonymous substitution. Different models like Nei–Gojobori method, Li–Wu–Luo method, Pamilo–Bianchi–Li method, Kumar method, etc., are available in MEGA. Different phylogenetic tree construction algorithms like Neighbor Joining (NJ)/ Unweighted Paired Group Method with Arithmetic Mean (UPGMA) method, Maximum Parsimony (MP), Maximum Likelihood (ML), etc., are available in MEGA. Users can choose the model as per their requirement. MEGA offers

statistical test for constructed tree which includes interior branch test and bootstrap test. Molecular clock test includes Tajima's test and ML based molecular clock test. MEGA also offers reconstruction of ancient state by Maximum Likelihood method. ML estimates ancestral state of each node of phylogenetic tree. Maximum parsimony method is applied for inferring ancestral state when sequence diversity is low.

5.4 Tools for Genome Biology and Evolution

The wealth of genomic data has accelerated the research on evolutionary biology. The huge information of genome of different species has driven the scientists to develop novel computational tools to infer useful information from the vast repertoire of data which is rapidly increasing also. Some of those tools are IMPUTOR, DeCoSTAR, MultiTWIN, POPBAM, VISTA genome browser and tools, PipMaker, Expasy tools including Bgee, OMA, ALF, BayeScan, etc.

IMPUTOR: The next generation sequences have generated huge sequence data but relatively short read lengths influences the quality of the data. IMPUTOR compares sequences by constructing a high confidence phylogenetic tree and imputes for a set of recombining sequences. It functions via the principle of parsimony which considers ancestral sites will not get reversion. It imputes missing variants and also corrects nonmissing sites that may arise false sequencing errors (Jobin et al. 2018).

DeCoSTAR: This software reconstructs ancestral genes and genomes. It organizes the ancestral sequences in form of adjacencies of ancestral sequences. It can also improve the assembly of fragmented genome by recognizing scaffolding fragments formed by evolutionary pressure. Ancestral genes or domains are calculated by considering phylogenetic tree which is constructed by taking into account different events that influence gene evolution like gain, loss, duplication, or transfer. This software not only able to reconstruct gene domains generate like breakage, fusion, etc., but also able to handle large datasets (Duchemin et al. 2017).

MultiTwin: This software package assumes a multipartite graphical approach to construct and analyze evolution at different levels of organization. This software takes into account different levels of biological organization like genes, genomes, communities, or environment. It comprehensively analyze sequence based classification. This type of graph is useful in comparative analysis of genome in microbes, gene sharing between their cellular genome, transposable elements, etc. This tool can be used to decipher pathogenicity traits among microbial community (Corel et al. 2018).

POPBAM: Next generation sequencing results yield multiple short fragments. POPBAM software is a collection of tools for evolutionary analysis of whole genome alignment among multiple species. It uses BAM formatted file. BAM is a compressed binary version of sequence alignment map/file (SAM). POPBAM uses sequence assembly file and calls variant sites and calculates different statistical parameter related to evolutionary biology such as nucleotide diversity, linkage disequilibrium, population divergence, etc. It follows sliding window method across the chromosome to perform statistical analysis. It is a fast and memory efficient program (Garrigan 2013).

VISTA genome browser: VISTA genome browser is a set of tools for comparative genomics. It starts with raw genomic data and processes it to results ready for visualization. It was originally developed for long genome sequence alignment and later developed to compute multiple sequence alignment and visualize results and analyze the result for important footprints. It can perform whole genome alignment for multiple sequences. It can identify deleterious single nucleotide polymorphisms (SNPs) (Brudno et al. 2007). It uses a tool named LAGAN for alignment. There are different types of tools are there, namely pairwise LAGAN, multi LAGAN, shuffle LAGAN. LAGAN is glocal alignment algorithm. The local alignment is done by BLAT or CHAOS. VISTA has different comparative genomics server like mVISTA which aligns and compares sequences between multiple species, wgVISTA which aligns large sequences like microbial whole genome, gVISTA which aligns whole genome and finds ortholog, and phyloVISTA which aligns different sequences from different species considering their phylogenetic relationship (http://genome.lbl.gov/vista/customAlignment.shtml).

PipMaker: It is an early genomic analysis software for comparing two long DNA sequences. It identifies conserved segments and displays high resolution plots. The plots are displayed as percent identity plot (pip), whose axis shows degree of similarity and the sequence position. It can compare two closely related genomic sequences (e.g., human and mouse) and infer useful knowledge like cis-regulatory region and protein coding regions. It can distinguish between protein coding and noncoding regions. Also, gene regulatory elements can be identified as noncoding sequences that diverged 100–300 million years ago (Schwartz et al. 2000).

5.5 Insights Into Human Evolution

Computational role in evolutionary biology has played a crucial role in demystifying human evolution. Advancement in sequencing technology has enabled scientists to generate sequencing data but without proper analysis, it is just a "data graveyard." Also, most of the ancient DNA is highly contaminated. The source of contamination is bacteria and the person who is handling the sample. So, obtaining high quality DNA sample from fossils (e.g., Neanderthals) is very difficult. The development of metagenomic approach has opened up a new era where identification of mixed sample is possible (Noonan 2010). The partial annotation of Neanderthal genome revealed that they belong to hominid group and most closely related to modern human. The genetic changes have been occurred in last few hundred thousand years. The time based modifications in human genome sequences have revealed a major morphological, behavioral, and cognitive changes (Green et al. 2006). Structured vocal communication is an important parameter for human and FoxP2 gene is thought to play an important role in language processing. In comparison to human, Neanderthal FOXP2 gene has shown two nucleotide changes in 911 and 977 in exon

11. (Krause et al. 2007). The genomic study reveals that Neanderthals interbred with ancient human and non-African modern human carry Neanderthal genes. The analysis revealed that modern East Asian population carry more amount of Neanderthal genes compared to European population. Male hybrid sterility was found as one of the significant deleterious mutation acquired from Neanderthal (Sankararaman et al. 2014). One comparative analysis of genome copy number variation between human and other 9 primates has revealed that some genes like AQP7 has human specific duplication which drives species specific evolution (Dumas et al. 2007). Thus, different scientific work indicate that human evolutionary biology is steadily enriching with help of computational tool.

5.6 Role in Viral Evolution

Genome wide association study (GWAS) is a very useful tool to identify gene association to diseases. GRASP2 database is a publicly available repository 8.87 million SNP-disease association (Karim et al. 2016). In another study, a novel human adenoviral pathogen causing pneumonia was identified and with help of recombination analysis, phylogenomics and phylodynamics study reveal it has three hosts (human, chimpanzee, and bonobo) and it is able to cross species infection (Dehghan et al. 2019). In another study, the group has analyzed genome of 95 strains of human adenovirus and performed different in silico analysis like recombination and structural analysis, phylogenetic analysis, etc. They identified horizontal genome transfer by recombination is an important feature for adenovirus evolution and may pose serious consequence on human health (Ismail et al. 2018). Transposable elements are very important evolutionary tool. They are found in most species and they regulate varieties of cellular mechanism. There are different tools and databases like RepBase, Dfam, GyDB, SINEbase, TREP, RiTE, RepeatMasker, etc., have been created to identify and catalog transposable elements (Goerner-Potvin and Bourque 2018). Nipah virus is a zoonotic virus and assumed to be a threat to animal and human. In one study, authors have collected all the available Nipah virus genome sequence and analyzed by phylogenetics and molecular evolution study. The receptor analysis and other studies indicate the variability among two strains and towards its adaptive evolution (Li et al. 2020a, 2020b). Flaviviridae family of viruses causes major health hazards. It is a single strand RNA virus. The synonymous codon usage pattern of flaviviridae family has been analyzed and correspondence analysis study revealed it is constituted of two groups (Yao et al. 2019). Ebola virus is one of the most deadly virus currently in circulation. Authors have analyzed Ebola virus data and identified that positive genetic selection has happened on GP and L genes and there could be more strains that is able to make human to human transmission (Liu et al. 2015). Upon studying Ebola virus data and performing phylodynamic assessment of intervention strategies, hypothetical impact assessment, role of barriers on virus transmission, etc., helps to make strategy to stop viral spread (Dellicour et al. 2018). Simulation is an important step to predict evolutionary dynamics of virus genome. One research group has created a software named SANTA-SIM for simulating viral sequence evolutionary dynamics considering recombination event and mutational selection, and it moves forward through time (Jariani et al. 2019).

5.7 Conclusion

In conclusion, computational evolutionary biology has a multifaceted application in the field of evolutionary biology, studying virus evolution, etc. The reconstruction of ancient sequences is applied in improving strains and sequences for modifying metabolic pathways. Thus, creating a new field named systems metabolic engineering which is an amalgamation of metabolic engineering, synthetic biology, and evolutionary engineering (Choi et al. 2019). Molecular clock hypothesis has been extensively used to analyze evolution and spread of SARS-CoV-2 which is ravaging human life since 2020 (Benvenuto et al. 2020; Lai et al. 2020; Zehender et al. 2020). The recent Covid-19 outbreak has reinforced the significance of studying computational evolutionary biology as the viruses are constantly evolving and many of them are zoonotic. The suitable prediction model is necessary to fight against future events of pandemic. Computational tools and algorithms may play a crucial role in that. The advancement in the field of machine learning (ML) can play a significant role in that approach. Researchers have used ML based prediction for viral genome classification (Remita et al. 2017) and also predicted human adaptive influenza A virus based on its nucleotide composition (Li et al. 2020a, b). Different applications of artificial intelligence in identifying disease causing viruses have already discussed in review articles (Park et al. 2020). The role of computational tools in tracing human evolution is also extensively used. The sequence data analysis of Neanderthals proved to be extremely crucial in understanding diversification of Homo sapiens from Neanderthals and the interbreeding of two species. Overall, computational evolutionary biology is a very important and diverse topic for study and work.

References

Benvenuto D, Giovanetti M, Salemi M, Prosperi M, De Flora C, Junior Alcantara LC, Ciccozzi M (2020) The global spread of 2019-nCoV: a molecular evolutionary analysis. Pathog Glob Health 114(2):64–67. https://doi.org/10.1080/20477724.2020.1725339

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchene S, Fourment M, Gavryushkina A, Drummond AJ (2019) BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 15(4):e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Bromham L, Duchene S, Hua X, Ritchie AM, Duchene DA, Ho SYW (2018) Bayesian molecular dating: opening up the black box. Biol Rev Camb Philos Soc 93(2):1165–1191. https://doi.org/10.1111/brv.12390

Brudno M, Poliakov A, Minovitsky S, Ratnere I, Dubchak I (2007) Multiple whole genome alignments and novel biomedical applications at the VISTA portal. Nucleic Acids Res 35: W669–W674. https://doi.org/10.1093/nar/gkm279

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38 (6):626–635. https://doi.org/10.1038/ng1789

Choi KR, Jang WD, Yang D, Cho JS, Park D, Lee SY (2019) Systems metabolic engineering strategies: integrating systems and synthetic biology with metabolic engineering. Trends Biotechnol 37(8):817–837. https://doi.org/10.1016/j.tibtech.2019.01.003

Corel E, Pathmanathan JS, Watson AK, Karkar S, Lopez P, Bapteste E (2018) MultiTwin: a software suite to analyze evolution at multiple levels of organization using multipartite graphs. Genome Biol Evol 10(10):2777–2784. https://doi.org/10.1093/gbe/evy209

Dehghan S, Seto J, Liu EB, Ismail AM, Madupu R, Heim A, Seto D (2019) A zoonotic adenoviral human pathogen emerged through genomic recombination among human and nonhuman simian hosts. J Virol 93:18. https://doi.org/10.1128/JVI.00564-19

Dellicour S, Baele G, Dudas G, Faria NR, Pybus OG, Suchard MA, Lemey P (2018) Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. Nat Commun 9 (1):2222. https://doi.org/10.1038/s41467-018-03763-2

Duchemin W, Anselmetti Y, Patterson M, Ponty Y, Berard S, Chauve C, Tannier E (2017) DeCoSTAR: reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. Genome Biol Evol 9(5):1312–1319. https://doi.org/10.1093/gbe/evx069

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM (2007) Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res 17(9):1266–1277. https://doi.org/10.1101/gr.6557307

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf 5:113. https://doi.org/10.1186/1471-2105-5-113

Futuyma DJ, Kirkpatrick M (2017) Evolution, 4th edn. Sinauer Associates, Inc., Sunderland

Garrigan D (2013) POPBAM: tools for evolutionary analysis of short read sequence alignments. Evol Bioinformatics Online 9:343–353. https://doi.org/10.4137/EBO.S12751

Goerner-Potvin P, Bourque G (2018) Computational tools to unmask transposable elements. Nat Rev Genet 19(11):688–704. https://doi.org/10.1038/s41576-018-0050-x

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Paabo S (2006) Analysis of one million base pairs of Neanderthal DNA. Nature 444(7117):330–336. https://doi.org/10.1038/nature05336

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52(5):696–704. https://doi.org/10.1080/10635150390235520

Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33:W557–W559. https://doi.org/10.1093/nar/gki352

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22(2):160–174. https://doi.org/10.1007/BF02101694

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97

Haubold B, Weihei T (2006) Phylogeny introduction to evolutionary biology: an evolutionary approach. Berhauser Verlag, Basel, pp 143–168

Ho SY, Duchene S (2014) Molecular-clock methods for estimating evolutionary rates and timescales. Mol Ecol 23(24):5947–5965. https://doi.org/10.1111/mec.12953

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Ismail AM, Cui T, Dommaraju K, Singh G, Dehghan S, Seto J, Seto D (2018) Genomic analysis of a large set of currently-and historically-important human adenovirus pathogens. Emerg Microbes Infect 7(1):10. https://doi.org/10.1038/s41426-017-0004-y

Jariani A, Warth C, Deforche K, Libin P, Drummond AJ, Rambaut A, Theys K (2019) SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination. Virus Evol 5(1):vez003. https://doi.org/10.1093/ve/vez003

Jobin M, Schurz H, Henn BM (2018) IMPUTOR: phylogenetically aware software for imputation of errors in next-generation sequencing. Genome Biol Evol 10(5):1248–1254. https://doi.org/10.1093/gbe/evy088

Karim S, NourEldin HF, Abusamra H, Salem N, Alhathli E, Dudley J, Kumar S (2016) e-GRASP: an integrated evolutionary and GRASP resource for exploring disease associations. BMC Genomics 17(Suppl 9):770. https://doi.org/10.1186/s12864-016-3088-1

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30(14):3059–3066. https://doi.org/10.1093/nar/gkf436

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16(2):111–120. https://doi.org/10.1007/BF01731581

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Paabo S (2007) The derived FOXP2 variant of modern humans was shared with Neandertals. Curr Biol 17 (21):1908–1912. https://doi.org/10.1016/j.cub.2007.10.008

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9(4):299–306. https://doi.org/10.1093/bib/bbn017

Lai A, Bergna A, Acciarri C, Galli M, Zehender G (2020) Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. J Med Virol 92(6):675–679. https://doi.org/10.1002/jmv.25723

Li K, Yan S, Wang N, He W, Guan H, He C, Su S (2020a) Emergence and adaptive evolution of Nipah virus. Transbound Emerg Dis 67(1):121–132. https://doi.org/10.1111/tbed.13330

Li J, Zhang S, Li B, Hu Y, Kang XP, Wu XY, Jiang T (2020b) Machine learning methods for predicting human-adaptive influenza A viruses based on viral nucleotide compositions. Mol Biol Evol 37(4):1224–1236. https://doi.org/10.1093/molbev/msz276

Liu SQ, Deng CL, Yuan ZM, Rayner S, Zhang B (2015) Identifying the pattern of molecular evolution for Zaire ebolavirus in the 2014 outbreak in West Africa. Infect Genet Evol 32:51–59. https://doi.org/10.1016/j.meegid.2015.02.024

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092

Nater A, Burri R, Kawakami T, Smeds L, Ellegren H (2015) Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst Biol 64(6):1000–1017. https://doi.org/10.1093/sysbio/syv045

Nei M, Kumar S (2000) Molecular evoution and phylogenetics. Oxford University Press, Oxford

Noonan JP (2010) Neanderthal genomics and the evolution of modern humans. Genome Res 20 (5):547–553. https://doi.org/10.1101/gr.076000.108

Nuin P A, Wang Z, Tillier ER (2006) The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7:471. https://doi.org/10.1186/1471-2105-7-471

Park Y, Casey D, Joshi I, Zhu J, Cheng F (2020) Emergence of new disease: how can artificial intelligence help? Trends Mol Med 26(7):627–629. https://doi.org/10.1016/j.molmed.2020.04.007

Remita MA, Halioui A, Malick Diouara AA, Daigle B, Kiani G, Diallo AB (2017) A machine learning approach for viral genome classification. BMC Bioinf 18(1):208. https://doi.org/10.1186/s12859-017-1602-3

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61(3):539–542. https://doi.org/10.1093/sysbio/sys029

Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Reich D (2014) The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507(7492):354–357. https://doi.org/10.1038/nature12961

Sauquet H (2013) A practical guide to molecular dating. Comptes Rendus Palevol 12(6):355–367

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Miller W (2000) PipMaker--a web server for aligning two genomic DNA sequences. Genome Res 10(4):577–586. https://doi.org/10.1101/gr.10.4.577

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Sumner JG, Jarvis PD, Fernandez-Sanchez J, Kaine BT, Woodhams MD, Holland BR (2012) Is the general time-reversible model bad for molecular phylogenetics? Syst Biol 61(6):1069–1074. https://doi.org/10.1093/sysbio/sys042

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1(3):269–285. https://doi.org/10.1093/oxfordjournals.molbev.a040317

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10(3):512–526. https://doi.org/10.1093/oxfordjournals.molbev.a040023

Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Navin NE (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature 512(7513):155–160. https://doi.org/10.1038/nature13600

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24 (8):1586–1591. https://doi.org/10.1093/molbev/msm088

Yao H, Chen M, Tang Z (2019) Analysis of synonymous codon usage bias in flaviviridae virus. Biomed Res Int 2019:5857285. https://doi.org/10.1155/2019/5857285

Zehender G, Lai A, Bergna A, Meroni L, Riva A, Balotta C, Galli M (2020) Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy. J Med Virol. https://doi.org/10.1002/jmv.25794

6 Web-Based Bioinformatics Approach Towards Analysis of Regulatory Sequences

B. Sharan Sharma, Sonal R. Bakshi, Preeti Sharma, and Ramtej J. Verma

Abstract

Coding sequences, making up only a very small percentage of human genome, have been relatively well studied than regulatory sequences. Hence, importance of majority of the elements and sequences that control the regulation and expression of protein-coding genes remains unknown. To realize the role of regulatory sequences in health and diseases, structural and regulatory information needs to be transcribed from them to be incorporated with the protein-coding counterpart. This will certainly provide an opportunity to realize the information generated from non-coding sequences in basic and clinical research. With advancements in computational biology and bioinformatics tools and techniques, web-based bioinformatics tools provide ample opportunities to study regulatory elements which help understand their biological relevance especially when their dysregulation can result in disease. Analysis of regulatory sequences via powerful online approaches has given a boost to comparative genomics studies enabling accurate annotations of sequences under study. This chapter provides a bird's eye view of

B. S. Sharma (✉)
GeneXplore Diagnostics and Research Centre, Ahmedabad, Gujarat, India

Rivaara Labs Pvt Ltd, KD Hospital, Ahmedabad, Gujarat, India

S. R. Bakshi
Institute of Science, Nirma University, Ahmedabad, Gujarat, India

P. Sharma
Department of Zoology, Biomedical Technology and Human Genetics, Gujarat University, Ahmedabad, Gujarat, India

PanGenomics International Pvt Ltd, Sterling Accuris Diagnostics, Ahmedabad, Gujarat, India

R. J. Verma
Department of Zoology, Biomedical Technology and Human Genetics, Gujarat University, Ahmedabad, Gujarat, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_6

web-based online approaches for analysing regulatory sequences in mammalian genomes. As our knowledge on involvement of regulatory regions in diseases in increasing, it is only rationale to develop additional sophisticated user-friendly bioinformatics tools to analyse the regulatory sequences for the identification of unique regions or novel variants potentially involved in the pathogenesis of disease.

Keywords

Cis-regulatory elements · RSAT · Transcription factor · TFBS · Online bioinformatics tools

6.1 Introduction

A growing number of online bioinformatics tools provides a reliable platform towards in silico analyses of DNA sequences especially cis-regulatory elements (CREs). CREs are non-coding regions of the genome that generally control expression of linked genes. These regulatory elements include proximal regulatory elements—promoters; and distal regulatory elements—silencers, enhancers, insulators, and locus control regions (LCRs) (Budd 2012). Regulatory elements form an important component of mammalian genome owing to their distal locations and precise spatial and temporal regulation of associated genes. Regulatory DNA elements possess several transcription factor binding sites (TFBSs), also known as motifs, binding of these sites with specific transcription factors is a prerequisite to control the gene expression. Laboratory based molecular and genetic methods have been useful in identifying short stretches of repetitive DNA sequences providing binding sites to different transcription factors. However, in large scale studies laboratory methods become time-consuming and too arduous. With advancements in bioinformatics tools and techniques, computational methods are now available to efficiently and systematically identify and analyse regulatory DNA elements. Simple and user-friendly online web-based approaches are promising computational methods for systematic discovery of cis-acting regulatory elements and conserved motifs. Through this chapter, we provide a bird's eye view of current web-based online tools for analysing regulatory sequences in mammalian genomes.

6.2 Why Care About Regulatory Sequences?

Mammalian genome consists of coding and non-coding DNA sequences. In the past, much attention has been paid on coding regions of the DNA as they are the manufacturing units of functioning proteins and slightest of variation in protein-coding regions could have devastating effects on the proper functioning of protein that may result in disease conditions. On the other hand, mutations in regulatory elements, generally, alter expression of associated gene (increased or decreased expression) and not the protein structure, therefore, these mutations are less likely to exert phenotypic impact. However, a rising number of studies have identified variations outside the coding regions that are robustly associated with various quantitative traits and complex diseases (Worsley-Hunt et al. 2011; Mokry et al. 2016; Chatterjee and Ahituv 2017). The primary reason being mutations outside coding regions and in regulatory DNA sequences can disrupt binding sites of transcription factor or can create new ones eventually altering mechanisms of transcriptional regulation. Hitherto, in the past two decades, much stress has been given to protein-coding changes in a DNA sequence. Now, it is apparent that strategic focus will be on variations within regulatory sequences as well since precise functioning of various classes of regulatory elements is prerequisite towards error-free expression of linked genes, and failure of this molecular machinery can lead to serious magnitudes resulting in disease conditions.

6.3 The Marriage of Omics and Bioinformatics

Major disciplines of '-omic' technologies have evolved rapidly over the last decade or so (Sharma et al. 2019a). Consequently, with rapid advances in sequencing and microarray technologies, availability of genomic data has expanded tremendously. Handling of this genomic data holds many potential applications not only to understand normal physiology but also disease conditions. Hence, role of bioinformatics tools and techniques are fundamental. And, it would be appropriate to say that the marriage of omics with bioinformatics forges a frontier to provide scientists with unlimited opportunities towards breakthrough discoveries in medicine research.

Omics and bioinformatics are just not only used for studying genes and biological signalling pathways responsible for human diseases but are also used to identify potential new targets for their applications in therapeutic drugs and therapy (Allen and Cagle 2009). Omics revolution and rise of bioinformatics tools and techniques has upgraded our knowledge and capability to handle huge data-sets. Omics combined with high-throughput studies have facilitated mapping of a large number of genetic variants within coding as well as non-coding regulatory sequences (LaFramboise 2009; Begum et al. 2012; Gloss and Dinger 2018; Perenthaler et al. 2019).

Comparative genomic approaches using computational tools are bestowing a notable impression on the study of regulatory regions especially in mammals and at present denote the most systematic and sound approaches of envisaging non-coding sequences expected to control the gene expression patterns. By subjecting genomic sequences to in silico comparisons and subsequent investigations, we are slowly but surely moving towards a better catalogue of recurrent regulatory motifs responsible for the fundamental biological processes (Loots 2008). Analysis of data generated through genomic projects have given a valuable foundation for studying particular genetic variants causative to simple and complex ailments (Hasin et al. 2017), and as our knowledge on involvement of regulatory elements in diseases in increasing, it is only rationale to develop additional sophisticated user-friendly bioinformatics tools to analyse the regulatory sequences for the identification of unique regions or novel variants potentially involved in the pathogenesis of disease.

6.4 Web-Based Tools as Powerful Assets to Analyse Regulatory Sequences

Powerful computational approaches in the form of web-based tools have given a boost to comparative genomics studies enabling accurate annotations of regulatory sequences. Detection and analysis of non-coding regulatory elements with conventional approaches has been slow and laborious in general (Ogino et al. 2012). Progress in comparative genomics and computational biology has helped in improving this situation. With the availability of large number of genomic data, genome-wide identification and analysis of non-coding regulatory elements is now possible with the power of comparative genomics and bioinformatics tools and techniques. Significance and power of the web-based comparative genomics approaches has already been demonstrated by a number of studies (Turatsinze et al. 2008; Sand et al. 2009; Nguyen et al. 2018; Sharma et al. 2019b). Some of the commonly used sequence analysis web-based bioinformatics tools are listed in Table 6.1.

6.5 Discovery of Over-Represented Oligonucleotides (motifs) in Regulatory Sequences

With advancements in computational algorithms, over-represented oligonucleotides or motifs can be discovered which truly represent the transcription factor binding sites (TFBSs) in a sequence. In general, motifs are short and widespread DNA sequences which are significantly over-represented compared to random patterns. Binding of transcription factors (TFs) to these motifs represents important episode in regulating gene expression by regulatory elements of the genome. Motif analysis is an advanced field of sequence analysis in the current era which is possible because of the high-throughput DNA sequencing technologies such as ChIP-Seq which involves chromatin immunoprecipitation followed by next-generation sequencing to study a plethora of DNA–protein interactions in vivo (Liu et al. 2010). Several user-friendly online tools now exist for de-novo motif discovery and downstream analysis of discovered motifs. Regulatory sequence analysis tools (RSAT) (Medina-Rivera et al. 2015) and MEME Suite (Bailey et al. 2009) are examples of very popular tools for analysing regulatory elements including motif-based analyses.

Position weight matrices (PWMs) or Position-specific scoring matrices (PSSMs) are the standard model to describe binding motifs (Kiesel et al. 2018). In PSSMs, each motif location supplies additively and autonomously from other locations to the total binding energy (Kiesel et al. 2018). PSSMs based database of transcription factors for eukaryotic transcription factor binding profiles are available. JASPAR (Khan et al. 2018) and HOCOMOCO (Kulakovskiy et al. 2018) are two of the

Table 6.1 Web-based bioinformatics tools for regulatory sequence analysis

Tool	Application/description	URL	Reference(s)
RSAT (regulatory sequence analysis tools)	Collection of software tools for the detection and analysis of cis-regulatory elements	http://www.rsat.eu	Van Helden (2003), Nguyen et al. (2018)
The MEME suite (multiple expression motifs for motif elicitation)	A software toolkit for performing motif-based sequence analysis	http://meme-suite.org	Bailey et al. (2015)
Motif combinator	To search for combinations of cis-regulatory motifs	http://emu.src.riken.jp/combinator	Kato and Tsunoda (2007)
cREMaG (cis-regulatory elements in the mammalian genome)	In silico studies of the promoter properties of co-regulated mammalian genes	http://www.cremag.org	Piechota et al. (2010)
PRECISE (prediction of regulatory CIS-acting elements)	for prediction of cis-acting regulatory elements	http://www.dpw.wau.nl/pv/pub/precise/	Trindade et al. (2005)
WebMOTIFS	Automated discovery, filtering and scoring of DNA sequence motifs	http://fraenkel-nsf.csbi.mit.edu/webmotifs.html	Romer et al. (2007)
PWMScan (position weight matrix scan)	For scanning entire genomes with a position-specific weight matrix	http://ccg.vital-it.ch/pwmscan	Ambrosini et al. (2018)
BaMM Tools (Bayesian Markov models)	For de novo motif discovery and regulatory sequence analysis	https://bammmotif.mpibpc.mpg.de	Kiesel et al. (2018)
Cister (cis-element cluster finder)	For detecting regulatory regions in DNA sequences, by searching for clusters of cis-elements	http://sullivan.bu.edu/~mfrith/cister.shtml	Frith et al. (2001)
i-cisTarget (integrative-cisTarget)	For the prediction of regulatory features and cis-regulatory modules	https://gbiomed.kuleuven.be/apps/lcb/i-cisTarget/	Herrmann et al. (2012)
MatInspector	Identifies TFBS in nucleotide sequences using a large library of weight matrices	https://www.genomatix.de/online_help/help_matinspector/matinspector_help.html	Quandt et al. (1995), Cartharius et al. (2005)
Cluster-buster	For finding clusters of pre-specified motifs in DNA sequences	http://zlab.bu.edu/cluster-buster/	Frith et al. (2003)
TFmotifView	To study the distribution of known TF motifs in genomic regions	http://bardet.u-strasbg.fr/tfmotifview/	Leporcq et al. (2020)

popular examples of online TF databases which are widely used to predict, compare, and analyse unknown and/or novel motifs and transcription factors with transcription factor binding models stored in these database.

In one of our original and recent works, a systematic bioinformatics approach was adapted to detect and analyse motifs in human locus control regions (LCRs) (Sharma et al. 2019b). LCRs are important, however, not much studied, cis-acting regulatory sequences that control expression of linked genes in a position-independent and copy-number dependent manner. Using web-based RSAT suite, motifs of biological relevance could be discovered in the important human LCRs to help understand their unique regulatory features. LCRs form an important component of integrating vectors owning to their unique expression control abilities, therefore, identification of unique regulatory signatures present within LCR sequences will be contributory in the design of new generation of regulatory elements. One such example was described in the design of a non-viral mammalian expression vector in which the primary transgene was under the transcriptional control of elements of LCR (Sharma and Verma 2020). Such vector design provides a framework for strong regulation with non-viral features which confer certain advantages over viral vectors.

In order to understand the regulation of a gene, understanding of the mechanism of DNA–protein interaction at the molecular level is important (Sharma et al. 2020). This interaction, in general, involves various classes of regulatory elements for eventual and faithful expression of genes. This expression machinery, thus, is an essential aspect of cellular functioning failure of which could lead to serious consequences resulting in diseases. With advancements in computational biology and bioinformatics tools and techniques, it is now easier to discover unique regulatory signatures (URS) of regulatory elements in order to understand the molecular machinery. Web-based bioinformatics tools such as RSAT and MEME are powerful tools to better understand regulatory elements in order to predict their potential role in health and diseases. We have used RSAT suite in the past to study LCRs, some of the features of RSAT suite are outlined in the next section.

6.6 Regulatory Sequence Analysis Tools (RSAT)

Regulatory Sequence Analysis Tools (RSAT) is a well-documented and a popular suite of modular tools that detects and analyse cis-regulatory elements in genome sequences. RSAT web server has been running without interruption since late 1990s (Van Helden et al. 1998). Initially named yeast-tools, RSAT at beginning was restricted to yeast genome only (Van Helden et al. 2000a). During early 2000s, RSAT web server was mainly centred on oligo-analysis and dyad-analysis based on string-based pattern-discovery algorithms (Van Helden et al. 2000b; Van Helden 2003). Later, RSAT was upgraded to analyse sequences PSSMs, and for the identification of conserved elements in promoters of orthologous genes (phylogenetic footprints) by the inclusion of new tools (Thomas-Chollier et al. 2008). RSAT update in 2010 supported 1794 genomes which included 1120 bacteria, 88 archaea, 98 fungi, 16 metazoa, and 461 phages (Thomas-Chollier et al. 2011). This update described 13 new programs added to the 30 tools of the previous version. A series of protocols were also described through different publications to give step-by-step instructions about option choices and result interpretation for the popular tools of RSAT (Janky and van Helden 2007; Sand and Helden 2007; Defrance et al. 2008; Sand et al. 2008; Turatsinze et al. 2008). RSAT 2015 version offered access to a large number of genomes from all kingdoms, assisted by a new taxon-specific organization of the public servers, and was also expanded to diversify its applications, including comparison and clustering of motifs, regulatory variants analyses and comparative genomics (Medina-Rivera et al. 2015). The 20th anniversary article of RSAT provided updates on the novelties included in RSAT 2018 suite, and also presented various access and training modalities (Nguyen et al. 2018).

Proper sequence dataset is the beginning point of any genomic analysis, RSAT suite provides an application called as ‘retrieve-ensembl-seq’ that significantly eases the retrieval of sequences from the Ensembl database in a user-friendly fashion (Sand et al. 2009). RSAT suite has been useful in detecting and analysing putative cis-regulatory elements and regions enriched in such elements (Turatsinze et al. 2008). Peak-motifs of RSAT, as a comprehensive pipeline, efficiently discovers motifs and identify putative transcription factors in ChIP-seq and similar data (ChIP-PET, ChIP-on-chip, CLIP-seq). Biological validity of peak-motifs was demonstrated by recovering the correct motifs from ChIP-seq sets corresponding to known transcription factors, moreover predicted specific motifs and transcription factors in an original analysis (Thomas-Chollier et al. 2012). Recently ‘Variation-tools’ program has been included in the well maintained suite RSAT which provide an accessible resource for expert and non-expert users to analyse regulatory variants in a web interface for as many as fifteen organisms with flexibility to upload personal variant and PSSM collections (Santana-Garcia et al. 2019).

Development of RSAT from yeast genome to high-throughput sequencing era has impacted the analysis of regulatory elements in a user-friendly and positive way. Comprehensive identification of functional elements, most notably regulatory motifs, has a fundamental importance towards biomedical research (Hashim et al. 2019), especially when their dysregulation might lead to pathological conditions. Computational identification of other functional elements such as DNase I-hypersensitive sites, zinc finger domains, and other regulatory signatures within the regulatory sequences is an important aspect of biomedical research. Development of additional user-friendly online algorithms/tools will be instrumental to gain regulatory insights of complex sequences.

6.7 Future Perspectives

In the current era, analysing complex sequences, involved in gene regulation, using user-friendly computational and bioinformatics tools is an advanced sequence analysis field with tremendous scope in omics technologies. Such kind of sequence analysis is not only a fine finishing tool for wet lab results but also offers a cradle of novel biological knowledge ranging from improved sequence binding models to exploration of specific binding site arrangements.

Regulatory sequence analysis also serves in finding out actual in vivo binding pattern of a particular protein.

We are in a phase where laboratories without having the need to undertake deep sequencing projects can mine data available in public platforms, and this is where web-based tools are handy, enabling us to work from any computer, toward systematic analysis of mined data. This helps users to extract and summarize relevant information intuitively from huge amount of genomic sequence data. Furthermore, bioinformatics suites available online have made it easier to comprehensively study regulatory elements which are of fundamental importance towards biomedical research. Web-based sequence analysis tools promote genomic research greatly for their flexible accessibility, ease-of-use, and quality performance.

Newer and improved technologies such as single molecule real-time (SMRT) sequencing have potential applications and utilities for medical diagnostics (Ardui et al. 2018). SMRT technology confers advantages like long-read lengths and high consensus accuracy (Nakano et al. 2017). Genetic variations like structural variations and complex rearrangements might affect gene expression by disrupting non-coding regulatory elements (Mitsuhashi and Matsumoto 2019). Identification of such variations has been a challenge and this is where clinically oriented studies using long-read sequencer have proved to be useful. SMRT sequencing has enabled the detection of precise breakpoints which are otherwise not easy to detect using short-read sequencing as breakpoints frequently occur in repetitive regions (Stancu et al. 2017). Diverse detection strategies can identify different types of variations across the genome and hence the advances like SMRT sequencing promise telomere to telomere, gap free coverage which increase the chances of unravelling the non-coding regions and therefore the need to be better equipped with user-friendly and easily accessible bioinformatics tools.

As summarized in this chapter, web-based sequence analysis tools are useful tools in computational biology, especially in regulatory sequence research. However, there are few challenges as well such as how to define false positive or false negative results. Possible solution is still the gold standard wet lab approach to validate the dry lab results. In conclusion, with the progressive development of new in silico technologies, web-based bioinformatics tools are becoming central platform for researchers to extract useful information and generate knowledge, promoting scientific discoveries.

References

Allen TC, Cagle PT (2009) Bioinformatics and omics. In: Basic concepts of molecular pathology. Springer, Boston, pp 55–60

Ambrosini G, Groux R, Bucher P (2018) PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. Bioinformatics 34(14):2483–2484

Ardui S, Ameur A, Vermeesch JR, Hestand MS (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res 46(5):2159–2168

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37(suppl_2):W202–W208

Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME suite. Nucleic Acids Res 43(W1): W39–W49

Begum F, Ghosh D, Tseng GC, Feingold E (2012) Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Res 40(9):3777–3784

Budd A (2012) Introduction to genome biology: features, processes, and structures. In: Evolutionary genomics. Humana Press, Totowa, pp 3–49

Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21(13):2933–2942

Chatterjee S, Ahituv N (2017) Gene regulatory elements, major drivers of human disease. Annu Rev Genomics Hum Genet 18:45–63

Defrance M, Sand O, Van Helden J (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. Nat Protoc 3(10):1589

Frith MC, Hansen U, Weng Z (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics 17(10):878–889

Frith MC, Li MC, Weng Z (2003) Cluster-buster: finding dense clusters of motifs in DNA sequences. Nucleic Acids Res 31(13):3666–3668

Gloss BS, Dinger ME (2018) Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med 50(8):1–8

Hashim FA, Mabrouk MS, Atabany WA (2019) Comparative analysis of DNA motif discovery algorithms: a systemic review. Curr Cancer Ther Rev 15(1):4–26

Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genome Biol 18(1):83

Herrmann C, Van de Sande B, Potier D, Aerts S (2012) i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. Nucleic Acids Res 40(15): e114

Janky RS, van Helden J (2007) Discovery of conserved motifs in promoters of orthologous genes in prokaryotes. In: Comparative genomics. Humana Press, Totowa, pp 293–308

Kato M, Tsunoda T (2007) MotifCombinator: a web-based tool to search for combinations of cis-regulatory motifs. BMC Bioinf 8(1):100

Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, Lee R, Bessy A, Chèneby J, Kulkarni SR, Tan G, Baranasic D (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res 46(1):D260–D266

Kiesel A, Roth C, Ge W, Wess M, Meier M, Söding J (2018) The BaMM web server for de-novo motif discovery and regulatory sequence analysis. Nucleic Acids Res 46(W1):W215–W220

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, Kolpakov FA (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res 46(D1):D252–D259

LaFramboise T (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res 37(13):4181–4193

Leporcq C, Spill Y, Balaramane D, Toussaint C, Weber M, Bardet AF (2020) TFmotifView: a webserver for the visualization of transcription factor motifs in genomic regions. Nucleic Acids Res 48:W208–W217

Liu ET, Pott S, Huss MQ (2010) A: ChIP-seq technologies and the study of gene regulation. BMC Biol 8(1):56

Loots GG (2008) Genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis. Adv Genet 61:269–293

Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM (2015) RSAT 2015: regulatory sequence analysis tools. Nucleic Acids Res 43(W1):W50–W56

Mitsuhashi S, Matsumoto N (2019) Long-read sequencing for rare human genetic diseases. J Hum Genet 2019:1–9

Mokry M, Harakalova M, Asselbergs FW, de Bakker PI, Nieuwenhuis EE (2016) Extensive association of common disease variants with regulatory sequence. PLoS One 11(11):e0165893

Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, Shinzato M, Minami M, Nakanishi T, Teruya K, Satou K (2017) Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell 30(3):149–161

Nguyen NT, Contreras-Moreira B, Castro-Mondragon JA, Santana-Garcia W, Ossio R, Robles-Espinoza CD, Bahin M, Collombet S, Vincens P, Thieffry D, Helden J (2018) RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res 46(W1):W209–W214

Ogino H, Ochi H, Uchiyama C, Louie S, Grainger RM (2012) Comparative genomics-based identification and analysis of cis-regulatory elements. In: Xenopus protocols. Humana Press, Totowa, pp 245–263

Perenthaler E, Yousefi S, Niggl E, Barakat S (2019) Beyond the exome: the non-coding genome and enhancers in malformations of cortical development. Front Cell Neurosci 13:352

Piechota M, Korostynski M, Przewlocki R (2010) Identification of cis-regulatory elements in the mammalian genome: the cREMaG database. PLoS One 5:8

Quandt K, Frech K, Karas H, Wingender E, Werner T (1995) Matlnd and Matlnspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res 23(23):4878–4884

Romer KA, Kayombya GR, Fraenkel E (2007) WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. Nucleic Acids Res 35(suppl_2):W217–W220

Sand O, Helden J (2007) Discovery of motifs in promoters of coregulated genes. In: Comparative genomics. Humana Press, Totowa, pp 329–347

Sand O, Thomas-Chollier M, Vervisch E, Van Helden J (2008) Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services—an example with ChIP-chip data. Nat Protoc 3(10):1604

Sand O, Thomas-Chollier M, Van Helden J (2009) Retrieve-ensembl-seq: user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl. Bioinformatics 25 (20):2739–2740

Santana-Garcia W, Rocha-Acevedo M, Ramirez-Navarro L, Mbouamboua Y, Thieffry D, Thomas-Chollier M, Contreras-Moreira B, van Helden J, Medina-Rivera A (2019) RSAT variation-tools: an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding. Comput Struct Biotechnol J 17:1415–1428

Sharma BS, Verma RJ (2020) Prediction and design of zinc finger target sites for an important human regulatory region (locus control region). Res J Biotechnol 15(7):78–82

Sharma BS, Prabhakaran V, Desai AP, Bajpai J, Verma RJ, Swain PK (2019a) Post-translational modifications (PTMs), from a cancer perspective: an overview. Oncogen J 2(3):12

Sharma BS, Swain PK, Verma RJ (2019b) A systematic bioinformatics approach to motif-based analysis of human locus control regions. J Comput Biol 26(12):1427–1437

Sharma BS, Prabhakaran V, Verma RJ (2020) Design of non-viral vector with improved regulatory features towards therapeutic application. Bioinformation 16(4):307–313

Stancu MC, Van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, De Ligt J, Pregno G, Giachino D, Mandrile G, Valle-Inclan JE, Korzelius J (2017) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun 8(1):1–3

Thomas-Chollier M, Sand O, Turatsinze JV, Janky RS, Defrance M, Vervisch E, Brohee S, van Helden J (2008) RSAT: regulatory sequence analysis tools. Nucleic Acids Res 36(suppl_2): W119–W127

Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J (2011) RSAT 2011: regulatory sequence analysis tools. Nucleic Acids Res 39(suppl_2):W86–W91

Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 40(4):e31

Trindade LM, Van Berloo R, Fiers MW, Visser RG (2005) PRECISE: software for prediction of cis-acting regulatory elements. J Hered 96(5):618–622

Turatsinze JV, Thomas-Chollier M, Defrance M, Van Helden J (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc 3 (10):1578

Van Helden J (2003) Regulatory sequence analysis tools. Nucleic Acids Res 31(13):3593–3596

Van Helden J, André B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281 (5):827–842

Van Helden J, André B, Collado-Vides J (2000a) A web site for the computational analysis of yeast regulatory sequences. Yeast 16(2):177–187

Van Helden J, Rios AF, Collado-Vides J (2000b) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res 28(8):1808–1818

Worsley-Hunt R, Bernard V, Wasserman WW (2011) Identification of cis-regulatory sequence variations in individual genome sequences. Genome Med 3(10):65

An Overview of Bioinformatics Resources for SNP Analysis

7

Sudarkodi Sukumar, Arunika Krishnan, and Subhamoy Banerjee

Abstract

Genetic variations are pivotal in causing intra-species diversity of organisms. The advent of high throughput genomic technologies has led to the large-scale cataloguing of genetic variations not just in humans, but in several other organisms. Single nucleotide polymorphisms (SNPs) are the major form of genetic variations. SNPs are essential in understanding the evolution of phenotypic differences of organisms and furthermore, are being used as markers in diagnostics and therapeutics for various diseases. As the genomic sequence data are increasing extensively, efficient tools are required to analyse and functionally interpret the SNPs. Various bioinformatics and statistic tools have been employed over the years for SNP analysis. In this chapter, a detailed account is presented on the various bioinformatic approaches existing for SNP analysis for both human and other non-human genomes. Furthermore, the challenges and gaps to be addressed in the bioinformatics field are discussed in order to study SNPs efficiently in the future.

Keywords

Single nucleotide polymorphisms · Variants · SNPs · Bioinformatics · SNP Analysis

S. Sukumar (✉) · A. Krishnan · S. Banerjee
School of Life Sciences, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
e-mail: sudarkodi.sls@crescent.education

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_7

Abbreviations

bp	Base pair
CNV	Copy Number Variations
cSNPs	coding region SNPs
GWAS	Genome Wide Association Studies
ncSNPs	non-coding region SNPs
NGS	Next Generation Sequencing Technology
nsSNPs	non-synonymous SNPs
QTL	Quantitative Trait Loci
SNPs	Single Nucleotide Polymorphisms
SNVs	Single Nucleotide Variants
sSNPs	Synonymous SNPs
STR	Short Tandem Repeats
SVM	Support Vector Machine
UTR	Untranslated region
VCF	Variant Call Format
VNTRs	Variable Number of Tandem Repeats
WGS	Whole Genome Sequences

7.1 Introduction

Variations, be it similarities or differences between organisms, have been a significant trigger for the study of genetics since Mendel's experiments on pea plants. Mendel and his successors such as Morgan aimed to associate a phenotypic trait with unknown genetic factors. But post the completion of the human genome project (Lander et al. 2001), genetic scientists are aiming to associate known genetic variations with unknown phenotypic traits. Thus, the advancements brought about by the human genome project has completely reframed the prespectives of geneticists over a century. Apart from the main goal of documenting the sequence of approximately 25,000 human genes, the other equally significant objective of the human genome project was to establish tools to analyse the enormous data (Collins et al. 2003). Bioinformatics has been undoubtedly at the core of achieving this objective. Bioinformatics has advanced considerably and has an indispensable role at every step of genomic analysis, beginning from processing of whole genome sequences (WGS); variant identification; annotation; and functional interpretation of variations (Roos 2001).

7.1.1 Types of SNPs

Genetic variations are of different types such as genomic structural variations, single nucleotide polymorphisms (SNPs), insertions/deletions or indels, block substitutions, inversions, variable number of tandem repeats (VNTRs), and copy number variations (CNVs). The most common and abundant form of genetic variation in the human genome are SNPs (Collins et al. 1998). SNPs are basically single nucleotide substitutions that occur in at least 1% of a population. Substitutions can be either transitions (purine ↔ purine or pyrimidine ↔ pyrimidine) or transversions (purine ↔ pyrimidine), although transitions are most common (Wang et al. 1998). In the human genome on an average, a SNP can occur about every 1000 base pair (bp) (Taillon-Miller et al. 1998) and are responsible for 90% of genomic variations between individual humans (Brookes 1999). Although SNPs are mostly biallelic, significant research suggests the presence of triallelic (Hodgkinson and Eyre-Walker 2010) and as well as tetra-allelic SNPs (Phillips et al. 2015). SNPs can occur in the coding region or in the non-coding region of a genome. Based on their likely effect on proteins in the form of amino acid variants, the coding region SNPs (cSNPs) can be further classified into synonymous SNPs (sSNPs) and non-synonymous SNPs (nsSNPs) (Fig. 7.1).

Synonymous SNPs are substitutions that do not result in amino acid changes in the translated protein sequence owing to the degeneracy of the genetic code. Hence, sSNPs were also referred to as silent mutations for long, assuming the absence of any downstream effects on the protein. However, studies have proven that sSNPs can indeed affect protein conformation and function through different mechanisms (Hunt et al. 2009) and are even associated with human diseases (Brest et al. 2011; Nackley et al. 2006). Interestingly sSNPs have also found use in population genetic analysis on par with nsSNPs (Gutacker et al. 2006).

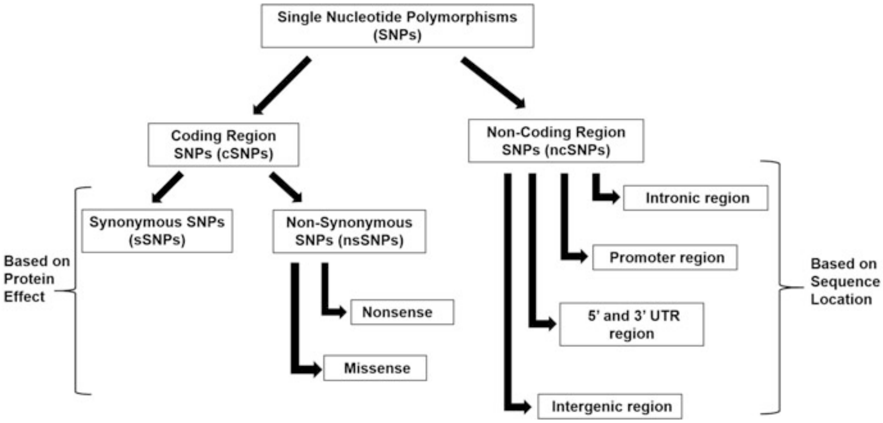

Single Nucleotide Polymorphisms (SNPs)
Coding Region SNPs (cSNPs)
Non-Coding Region SNPs (ncSNPs)
Synonymous SNPs (sSNPs)
Non-Synonymous SNPs (nsSNPs)
Nonsense
Missense
Based on Protein Effect
Intronic region
Promoter region
5' and 3' UTR region
Intergenic region
Based on Sequence Location

Fig. 7.1 Classification of Single Nucleotide Polymorphisms (SNPs)

Non-synonymous SNPs are substitutions that result in an altered amino acid sequence and are further classified into missense and nonsense SNPs. Missense nsSNPs lead to a different codon and therefore, an altered amino acid in the polypeptide sequence. Missense nsSNPs can affect the structural and/or functional properties of a protein. Thus, a missense mutation could be beneficial, neutral, mildly deleterious or even be lethal towards a protein function and thereby in its disease-causing potential (Z. Wang and Moult 2001). Even very rare missense SNPs could be deleterious and cause complex human diseases (Kryukov et al. 2007). Significantly, about 50% of all disease-causing mutations are missense SNPs (Krawczak et al. 2000). On the other hand, nonsense SNPs (nsSNPs) lead to a premature stop codon and hence, a truncated protein sequence.

Non-coding region SNPs (ncSNPs) are those that occur in stretches of DNA such as the introns, 3′ & 5′ untranslated regions (UTR), promoters and intergenic regions. With increasing studies on the importance of non-coding regions, several investigations have been accelerated in analysing the significance of ncSNPs. NcSNPs in UTR and intronic regions are capable of affecting alternative splicing, splice-site recognition by interfering with protein binding in those regions (Mansur et al. 2018). NcSNPs especially in promoter regions are expected to be regulatory in nature, affecting the expression of neighbouring genes and thereby associate with disease phenotypes (Zhang and Lupski 2015).

7.1.2 Applications of SNPs

Identification of SNPs in different species is made possible by the application of next generation sequencing technology (NGS) in the development of large genotypic arrays. High density occurrence, cheaper cost of establishing assays, adaptability of such assays between different laboratories are the factors in the favour of using SNPs for studying genetic variations between individual animals, humans, plants or even microbes within a population. Hence, SNP has varied applications such as haplotype mapping, linkage disequilibrium analysis, population genetic analysis, phylogenetic reconstruction, forensic analysis and implementation of personalized medicine (Fig. 7.2).

7.1.2.1 Strain Genotyping

Identification and characterization of microbes especially infectious pathogens at strain level are useful for pathogen typing and drug resistance screening, which aids accurate diagnosis and tailored treatment regime (Li et al. 2009b). Also, SNP genotyping has been extensively utilized to study the global spread and for phylogenetic reconstruction of pathogens such as *Mycobacterium tuberculosis* (Coll et al. 2014), Group A *Streptococcus* (Beres et al. 2006), *Brucella melitensis* (Tan et al. 2015), and *Plasmodium vivax* (Baniecki et al. 2015).

Apart from the research of pathogens, SNP genotyping is also valuable to study strain relationships of model organisms such as the mouse (Petkov et al. 2004; Moran et al. 2006), yeast (Wilkening et al. 2013) and drosophila (Berger et al. 2001).

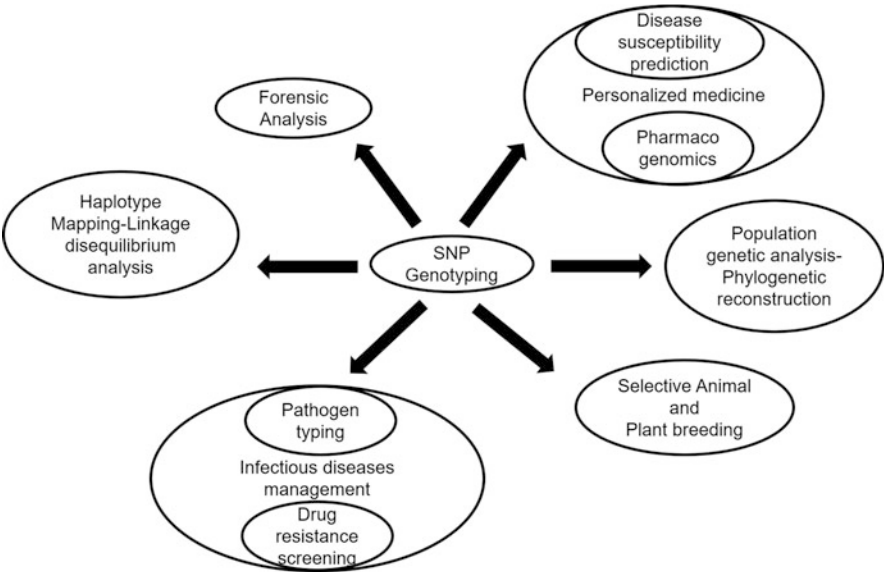

Disease susceptibility prediction
Personalized medicine
Pharmaco genomics
Forensic Analysis
Haplotype Mapping-Linkage disequilibrium analysis
SNP Genotyping
Population genetic analysis-Phylogenetic reconstruction
Selective Animal and Plant breeding
Pathogen typing
Infectious diseases management
Drug resistance screening

Fig. 7.2 Scheme of varied bioinformatics enabled applications of SNPs

7.1.2.2 Selective Breeding of Plants and Animals

The heritable nature of SNP is useful in high resolution genetic mapping and thereby facilitating selective marker assisted-breeding in plants and animals. Next generation sequencing of popular plant species such as *Oryza sativa* (Shen et al. 2004) and *Arabidopsis thaliana* (Pisupati et al. 2017) has propelled the sequencing of other important crops as well, thereby leading to large-scale generation of SNP data in plants (Lai et al. 2012). Similarly SNPs are useful in increasing breeding value of livestock by targeting quantitative trait loci (QTL) that are responsible for milk production, nutrient content of meat, eggs, etc. (Dekkers 2012).

7.1.2.3 Forensic Analysis

Conventionally SNPs were considered less informative due to its biallelic nature and secondary only to short tandem repeats (STR) for forensic analysis. But advancements in molecular biology techniques have encouraged the use of SNPs in place of STR especially for challenging samples that have less amount of template DNA or that are badly degraded. SNPs are informative and provide investigative leads in areas such as identity, ancestry, lineage and phenotypic traits (Budowle and Van Daal 2008). The 1000 Genomes project has enabled the discovery of tetra-allelic SNPs with more discriminatory power than biallelic SNPs in forensic applications (Sobrino et al. 2005).

7.1.2.4 Personalized Medicine

SNPs are not just responsible for variations in common physical traits between individuals but also influences the differences in disease susceptibility and drug response between individuals (Peterson et al. 2013; Ahmed et al. 2016). Such SNPs are as crucial in the case of infectious diseases as for metabolic diseases (Wang et al. 2018; Nogales and Dediego 2019) including even the Covid-19 pandemic (Hou et al. 2020). Therefore, it is indispensable to account these SNPs to offer personalized diagnostics as well as treatment options to combat diseases.

It is essential to note that none of these diverse applications of SNPs would be possible without advancements in technology that would aid in the identification, prediction and validation of SNPs. Undoubtedly, developments in bioinformatics are indispensable for the study of SNPs. This chapter provides a comprehensive view of the latest and updated bioinformatics resources and software tools for SNP analysis for not just human genomes but across the genomes of other organisms as well.

7.2 SNP Discovery and Identification

SNP discovery methods are diverse such as array-based hybridization, amplification-based (polymerase chain reaction-PCR) methods and sequencing. SNP arrays and PCR methods are more suitable for targeted detection of SNPs in specific genomic regions or genes. On the other hand, WGS is a promising and preferred method for SNP discovery for a genome wide and untargeted approach. Some excellent reviews are available for the readers for a deeper knowledge on SNP discovery and identification methods (Kwok and Chen 2003; Nielsen et al. 2011). Briefly after sequencing, the reads are aligned to a reference genome and differences in bases are detected between the two sequences, provided they are associated with a confidence or statistical score. This process is called SNP calling or also referred to as variant calling.

Some of the popularly used read alignment tools are MAQ alignment (Li et al. 2008), stampy (Lunter and Goodson 2011), BWA aligner (Li and Durbin 2009), BWA-MEM (Md et al. 2019), Bowtie aligner (Langmead et al. 2009), and its successor Bowtie2 (Langmead and Salzberg 2012). BWA and the Bowtie aligners are based on the data compression algorithm called the Burrows–Wheeler Transform (BWT), which is fast and memory efficient but less sensitive. Researchers who do not intend to compromise sensitivity for speed may opt for hash-based alignment such as MAQ and stampy. Some of the popularly used SNP calling algorithms are Broad Institute's GATK (Genome Analysis Tool Kit) (McKenna et al. 2010) (https://gatk.broadinstitute.org/hc/en-us), Sequence alignment/map (SAMtools) (Li et al. 2009a), VarScan (Koboldt et al. 2009) and Free Bayes (Garrison and Marth 2016). All these tools are capable of multiple sample SNP calling. Of significance is the SAMtool, which also introduced the SAM/BAM file format to store read alignments to further enable variant calling and thus facilitated the portability of NGS data. For further information, the readers can refer to the review by Chang Xu, which provides a detailed account of SNP calling algorithms for detecting somatic mutations (Xu 2018). All these tools are constantly evolving to match the advancements in sequencing technology in terms of read length, coverage, etc. These algorithms are freely available as stand-alone packages majorly for Linux platforms and some are available on Windows/Mac as well. Also, there are reviews that guide researchers in the selection of SNP calling methods and software on a general note (Altmann et al. 2012) and as well as specific to an organism (Olson et al. 2015). Still for experimental biologists with limited computational expertise and facilities, there are several simplified options to perform variant calling. GALAXY is one such genome analysis platform that includes variant calling (Blankenberg et al. 2010) and provides an automated workflow. PATRIC (https://www.patricbrc.org) provides a genome analysis environment exclusively for microbes (Wattam et al. 2017). It provides the user with multiple options for alignment and variant calling, which enables the user to frame their own simple SNP calling pipeline. Apart from these non-commercial open source tools, the commercial sequencing vendors offer all in one package that include variant listing along with the sequencing results, for example, the NGS pipeline CASAVA from Illumina. But still the experimentalists should have a basic idea on the advantages and limitations of the various alignment and SNP calling algorithms in order to obtain reliable and meaningful results.

7.3 SNP Data Resources

The Human Sequencing Project spurred research activities in various directions as a result of in-depth genomic analysis and resulted in enormous variation data. There was a need to regulate the unambiguous and uniform sharing and documentation of variation data. The Human Genome Variation Society established the nomenclature to be followed by the researchers to facilitate the effective utilization of such data (http://www.HGVS.org/varnomen) (den Dunnen et al. 2016). Several databases were established to document SNPs. The NCBI's short genetic variation database, commonly referred to as dbSNP (Sherry et al. 2001) (https://www.ncbi.nlm.nih.gov/snp) is the primary and one of the earliest established databases for SNPs in 1999 as a collaboration between NCBI and the National Human Genome Research Institute (NHGRI). Other short variations included in dbSNP are small scale insertions, deletions and microsatellites. As of June 2020, dbSNP houses nearly 0.73 billion reference SNP records that are archived, curated and annotated from approximately 2 billion submissions. Moreover the data at dbSNP forms the basis of variant resources for other databases within NCBI such as the OMIM (Hamosh et al. 2005) (https://www.omim.org/), ClinVar (Landrum et al. 2014) (https://www.ncbi.nlm.nih.gov/clinvar), dbVar (Lappalainen et al. 2013) (http://www.ncbi.nlm.nih.gov/dbvar) and as well as for platforms beyond NCBI such as the Variant Annotation Integrator at UCSC (Hinrichs et al. 2016); Ensembl Variant Effect Predictor (McLaren et al. 2016) (https://grch37.ensembl.org/info/docs/tools/vep). As a result of the above-mentioned data integration feature in NCBI, all SNPs associated with any gene in NCBI search can be easily viewed through the "Gene view in dbSNP" link. Allele frequency data of the SNP derived from studies such as 1000 Genomes project; clinical disease association status obtained through ClinVar; publications related to the SNP are some of the prominent information associated with each SNP record in dbSNP. Further, the variation viewer provides an interactive display to correlate and examine a variation in the genomic context. The data from dbSNP can be downloaded in variant call format (VCF format) (Danecek et al. 2011) (https://vcftools.github.io), a generic format exclusively developed for storing sequence variation. Each SNP is given a unique accession id called the reference SNP cluster ID or rsID. SNPs are identified using rsID not just within dbSNP but across different databases. Thus, dbSNP is undoubtedly the single largest and the most useful resource for SNP data analysis. However, since 2017 dbSNP no longer accepts any non-human SNP data submissions. Therefore, researchers interested in non-human variation data may utilize the European Variation Archive (https://www.ebi.ac.uk/eva) resource under European Bioinformatics Institute, which is also supported by its own variant browser.

The 1000 GENOMES project (https://www.internationalgenome.org/) (Auton et al. 2015; Sudmant et al. 2015) is an important initiative to catalogue human genetic variation on a massive scale by sequencing a large number of people from different ethnicities. The data from this project is freely available for scientific community through various databases. As mentioned earlier, this project data is integrated with dbSNP as well. The different ethnic population selection for this project is based on the samples of the HapMap project, which is a depository of human haplotype data. (haplotype refers to the cluster of SNPs occurring in a chromosome). Since 1000 Genome project has gained momentum in human population genetics and genomics research, the HapMap project has been discontinued.

Apart from the general SNP databases like dbSNP, many researchers created specialized databases to suit their needs and applications. The Human Gene Mutation Database (HGMD) (Stenson et al. 2020) (http://www.hgmd.cf.ac.uk) specifically curates germline mutations in nuclear genes that may cause human diseases from journal publications. It is available as a free public version for registered academic and non-profit institutions, which is updated only twice annually. On the other hand, the HGMD Professional version can be availed through a subscription from QIAGEN, which is updated quarterly.

Human genetic variation database (HGVD) is a reference database for genetic variations observed in a Japanese sample population (http://www.hgvd.genome.med.kyoto-u.ac.jp) (Higasa et al. 2016). The data is collated from exome sequencing of 1208 Japanese individuals performed at five different research institutes in Japan.

Ethnic National Database Operating software (ETHNOS) is a useful tool that provides a platform for establishing National and Ethnic Mutation Databases (NEMDBs) (van Baal et al. 2010). Using this tool NEMDBs of Israeli, Tunisian, Egyptian populations were established. Further these NEMDBs along with core databases like OMIM provide opportunities to create more specialized databases like FINDbase. FINDbase (Frequency of INherited Disorders database) is an online resource for collating information on frequencies of genomic variations that are pharmacogenomic biomarkers and cause inherited disorders (http://www.findbase.org/) (Kounelis et al. 2020).

Genome wide association study (GWAS) is the best approach utilized to detect common SNPs between disease and healthy individuals for different diseases, but with some challenges in its execution depending on disease complexity. GWAS studies are thus instrumental in identifying and correlating the role of SNPs in various diseases. The database of Genotypes and Phenotypes (dGaP) in NCBI is a repository of data on the interaction of genotypes and phenotypes, collated from various genome wide association studies (https://www.ncbi.nlm.nih.gov/gap). GWAS catalog is another central repository of information curated from the literature of various GWAS studies, which was originally started in 2008 by NHGRI and later maintained in collaboration with the EBI since 2010 (https://www.ebi.ac.uk/gwas/) (Buniello et al. 2019). GWAS central is yet another repository that provides summary of GWAS study findings that are curated from literature as well as documented from public domain projects (www.gwascentral.org*)* (Beck et al. 2020).

Ethical concerns pertaining to sharing of human data was an important area of research that was initiated during the human genome project execution. But today with sequencing and variation data being handled and shared by a large number of databases worldwide, there is larger concern over ethical integrity. In this regard, the Global Alliance for Genomics and Health (GA4GH) was established in 2013 to set standards for the secure and responsible sharing of human genomics data, which paves way for uninterrupted human health research (https://www.ga4gh.org/). This is a non-profit global organization comprising of members from 500 research organizations spread across 71 countries. While GA4GH focuses on the needs of researchers, the Human Variome Project (HVP) (https://www.humanvariomeproject.org/) (Cotton et al. 2008) established much earlier is a larger consortium; an official partner of UNESCO; spanning 81 countries is focused on the clinical needs and lays out guidelines for genomic variation data handling, which has direct impact on disease management. Adherence to such initiatives will sustain the trust and fruitful collaboration of researchers across the globe for human genetics research. While waiting for such initiatives to gain momentum and support from the concerned community, researchers who wish to share their data in a secure environment may utilize the services of Café Variome (https://www.cafevariome.org/). Café variome creates a network of clinicians and researchers who mutually trust each other to share and receive inputs on their genetic data.

The Leiden Open (source) Variation Database (LOVD) software was developed to support hassle free creation of variation databases (https://www.lovd.nl/) (Fokkema et al. 2011). Those researchers interested in creating and maintaining locus specific sequence variation database can utilize this freely available, platform independent software. It is essential that the researchers follow the nomenclature suggested by HGVS while documenting the sequence variations. Mutalyzer tool offers a quick option to crosscheck the sequence variance descriptions as per the HGVS nomenclature (https://www.mutalyzer.nl) (Wildeman et al. 2008).

The human sequencing project parallelly amplified the sequencing efforts in various other organisms that were important to humans, which again led to humungous sequence variation data and in turn multitude of non-human variation databases. The HGVS website provides a list of such non-human variation databases such as OMIA (Online Mendelian Inheritance in Animals) (https://omia.org/). The list at HGVS is just few examples, as the list is as exhaustive as for human variation databases. Genome based *Mycobacterium tuberculosis* variation database (GMTV)

(Chernyaeva et al. 2014)-(http://mtb.dobzhanskycenter.org); CropSNPdb http://snpdb.appliedbioinformatics.com.au/ (Scheben et al. 2019); variations in Chick pea https://cegresources.icrisat.org/cicarvardb/ (Doddamani et al. 2015); Arabidopsis https://arageno.gmi.oeaw.ac.at/ (Pisupati et al. 2017) are some more examples of non-human polymorphism repositories.

7.4 Functional Interpretation of SNPs

Though sequencing revolution led to an exponential increase in the identification of SNPs, however, the downstream process of characterization, annotation and functional interpretation of SNPs is still challenging. From the millions of SNPs in the human genome, only few are functional. Hence, it is a daunting task to identify such functional polymorphisms irrespective of their location in the genome. Experimental validation is the ultimate step to delineate the functional significance of SNPs. But, however, some *in silico* prefiltering step will be desirable to minimize the candidate SNP numbers from millions to a handful. Among the types of SNPs, nsSNPs are the most characterized due to their location in coding regions. In the following section, different computational approaches to annotate and characterize nsSNPs are discussed.

7.4.1 Sequence-Based Analysis

Sequence-based analysis of SNPs for functional prediction is the preferred method for prefiltering of large number of SNPs as there is no requirement for 3-dimensional structure information. There are numerous such tools available for the research community (Table 7.1).

SIFT (Sorting Intolerant Fr om Tolerant) uses sequence homology to predict the functional effect of nsSNPs and also frameshifting indels (insertion/deletion) on proteins. SIFT was one of the earliest tools used to analyse SNPs in human genome, but is also useful to study SNPs in other organisms such as Arabidopsis, *Mycobacterium tuberculosis* and model organisms like rat. SIFT calculates the probability of tolerance of an amino acid substitution and holds the substitution to be deleterious if the normalized value is less than a cut off (Ng and Henikoff 2003).

SNAP (Screening for Non-Acceptable Polymorphisms) tool classifies SNPs into neutral or non-neutral using a neural network model, which is derived using sequence-based features such as evolutionary conserved residue information and secondary structural attributes. The unique aspect of SNAP is that it provides a reliability index for every prediction, which in turn is a reflection of the level of confidence. This aspect uplifted the accuracy of SNAP when compared to the previous tools (Bromberg and Rost 2007).

dbNSFP is a database dedicated to functional predictions and annotation of all potential nsSNPs in the human genome. Predictions are combined from 37 prediction tools (Liu et al. 2011).

Table 7.1 Sequence-based SNP prediction tools

S. No	Tool	Significance	URL
1	SIFT (Ng and Henikoff 2003)	Uses sequence homology and calculates the probability of amino substitution and gives the score.	https://sift.bii.a-star.edu.sg/
2	SNAP (Bromberg and Rost 2007)	Uses neural network model, using sequence conservation and secondary structure attributes.	http://www.bio-sof.com/snap
3	dbNSFP (Liu et al. 2011)	A database that provides functional prediction and annotation of the SNPs in the human genome.	https://sites.google.com/site/jpopgen/dbNSFP
4	ANNOVAR (Wang et al. 2010)	A heuristic tool that provides functional annotation for novel SNPs	https://doc-openbio.readthedocs.io/projects/annovar
5	VAAST (Yandell et al. 2011)	Uses aggregative approach along with amino acid substitution data.	https://www.hufflab.org/software/vaast
6	PROVEAN (Choi et al. 2012)	Uses amino acid substitution and sequence homology to predict the effect of mutation on the protein.	http://provean.jcvi.org/index.php
7	SNPeff (Cingolani et al. 2012)	Provides annotation to SNPs across organisms from both coding and non-coding regions.	https://pcingola.github.io/SnpEff
8	VEST (Carter et al. 2013)	Uses random forest for predicting disease-causing variants.	http://www.cravat.us/CRAVAT/
9	FATHMM (Shihab et al. 2013)	Uses Hidden Markov Model for prediction.	http://fathmm.biocompute.org.uk/
9	MutationTaster2 (Schwarz et al. 2014)	Uses Bayes classification method for prediction.	http://mutationtaster.org/
10	PANTHER-PSEP (Tang and Thomas 2016)	Quantifies evolutionary preservation of site-specific amino acids for prediction.	http://pantherdb.org/tools/csnpScoreForm.jsp
11	MutPred2 (Pejaver et al. 2017)	Uses a neural network model for prediction	http://mutpred.mutdb.org/
12.	MyVariant.info	Provides variant annotation information collated from SNP databases.	http://myvariant.info/
13	VariO (Vihinen 2014)	Provides ontology and annotation for variation effects.	http://variationontology.org/
14	CADD (Rentzsch et al. 2019)	An integrated variant/indel annotation tool	https://cadd.gs.washington.edu/
15	Meta-SNP (Capriotti et al. 2013)	A meta predictor combining the results of 4 missense SNP predicting tools.	https://snps.biofold.org/meta-snp/

(continued)

Table 7.1 (continued)

S. No	Tool	Significance	URL
16	REVEL (Ioannidis et al. 2016)	A meta predictor combining the results of 13 pathogenic missense SNP predicting tools.	https://sites.google.com/site/revelgenomics/

ANNOVAR (Annotate Variation) is a heuristic tool that provides functional annotation for novel SNPs of human as well as non-human genomes obtained from cross-platform sequencing technologies. It provides functional effect annotation along with genomic region-specific annotations such as transcription binding sites, predicted microRNA targeted sites and stable RNA secondary structures. While many tools are developed rapidly to analyse SNPs, only few are continually updated and supported. ANNOVAR is one such reliable tool to annotate SNPs with updated access to newest version of SNP databases (Wang et al. 2010).

VAAST (the Variant Annotation, Analysis & Search Tool) uses an aggregative approach to identify both common and disease-causing variants. Similar to SIFT, VAAST also uses amino acid substitution data. VAAST is a fast and flexible tool with options to include new scoring methods in future (Yandell et al. 2011).

PROVEAN (Protein Variation Effect Analyzer) is a tool similar to SIFT in predicting the effect of amino acid substitutions on protein function using sequence homology, and which is executed using BLAST algorithm. PROVEAN is suitable for all organism predictions, whereas the PROVEAN BATCH PROTEIN execution option is available only for human and mouse (Choi et al. 2012).

SNPeff offers SNP annotations for multi-organisms and for both coding and non-coding regions of the genome. The speed of SNPeff is comparable to the earlier tools such as VAAST and ANNOVAR. SNPeff provides coverage of over 320 versions of genomes from multiple organisms. SNPeff can be easily integrated into GALAXY genome analysis software and as well as with Broad Institute's GATK variant calling tool (Cingolani et al. 2012).

VEST (Variant Effect Scoring Tool) uses random forest, a supervised machine learning approach for prioritizing rare functional SNPs that are disease causing. The training data set comprises of the HGMD and the Exome Sequencing Project population. VEST uses aggregated p-values to rank disease-causing variants across genomes and outperforms its counterparts like SIFT and PolyPhen in accuracy (Carter et al. 2013).

FATHMM (Functional Analysis Through Hidden Markov Models) uses a Hidden Markov Model (HMM) to analyse the amino acid substitution data in order to predict the deleteriousness of the variants (Shihab et al. 2013). FATHMM is available as species independent tool as well as species dependent tool with weightings for human mutations. FATHMM performs better than the conventional tools like SIFT, PANTHER when adapted with weightage. FATHMM is also able to predict disease associations with considerable accuracy and is applicable to high throughput genome predictions. The new extended tool FATHMM-XF is applicable for prediction of variants in non-coding regions of the genome (Rogers et al. 2018).

MutationTaster2 can predict the functional significance of nsSNPs but also, sSNPs and ncSNPs. MutationTaster2 incorporates Bayes classification method in three different classification models to suit the prediction of single amino acid substitutions, complex amino acid substitutions and sSNPs or ncSNPs, respectively. MutationTaster2 is trained and validated with SNP data from the 1000 Genome Project and HGMD Professional version. MutationTaster showed higher accuracy of prediction when compared to SIFT and PROVEAN (Schwarz et al. 2014).

PANTHER-PSEP (Protein analysis Through Evolutionary Relationships-Position Specific Evolutionary Preservation) is a simple tool that quantifies the concept of evolutionary preservation by measuring the time for which an amino acid has been preserved at a site, where an amino acid retained for a longer time implies a functional effect. This is the major difference for PANTHER-PSEP to perform better than other tools, which predominantly use evolutionary conservation of amino acids. Another advantage is the access to the large number of reference genomes of different organisms in PANTHER enabling multiple organism predictions (Tang and Thomas 2016).

MutPred2 is yet another amino acid substitution data using tool that can predict the pathogenicity of a variant as well as provide a list molecular alteration using probability scores. MutPred2 uses a neural network model with features extracted from amino acid sequence-based properties and trained using variant data from dbSNP, SwissVar (a portal to Swiss-Prot variants is currently discontinued and archived) and HGMD. MutPred2 performed substantially better than other tools such as MutationTaster2, Polyphen2, SIFT, etc. (Pejaver et al. 2017).

MyVariant.info is a simple to use web interface to search for variant annotation information collated from many different variation databases such as dbSNP, dbNSFP, GWS catalog, etc. One can also obtain all possible variants associated with a single gene. MyVariant.info can be integrated into other web applications thereby eliminating the need to store variant data locally.

VariO provides ontology for standard descriptions of SNPs as well as offers annotation for variation effects (Vihinen 2014).

CADD (Combined Annotation-Dependent Deletion) is a widely used integrated variant/indel annotation tool. It scores variants and ranks them based on a machine learning model comprising of 60 genomic features. The unique feature of CADD is that the model is trained using an unbiased and extensive dataset comprising of a set of simulated *de novo* variant list spanning the entire human genome in addition to all the variants that have arisen in the human genome since the split of human-chimpanzee. This is so unlike the other tools that are dependent only on the known list of variants deposited in various databases (Rentzsch et al. 2019).

As discussed so far there are multiple tools to annotate and functionally delineate variations based on amino acid substitution data. But when conflict arises between the results of these tools, meta prediction algorithm is the only solution, which has been found successful in several other applications such as DNA prediction and protein prediction methods.

Meta-SNP is an integration of four SNP prediction tools such as SIFT, PANTHER, Phd-SNP and SNAP. The results of all four tools are combined and used to run the random forest machine learning method in order to differentiate disease causing and disease-free variants. Meta-SNP indeed performed with better accuracy than the individual tools (Capriotti et al. 2013).

REVEL (rare exome variant ensemble learner) is another meta predictor that combines the results of 13 pathogenic missense SNP predicting tools. REVEL is exclusively developed for rare exome variants. REVEL incorporates random forest machine learning model that is trained with disease variants from HGMD and various population-based SNPs from dbSNP. The features for the model are the results from 13 individual prediction tools. REVEL performed better than the 13 individual predictor tools and as well as than other meta predictors like MetaSVM (Ioannidis et al. 2016).

7.4.2 Structure-Based Analysis

Tools that can discern the impact of nsSNPs on protein stability and its effect on interactions with other molecules are required to understand the role of such SNPs for the whole cell be it a cancer cell or an infectious bacterium or even a virus. Hence, advanced tools that utilize the complete structural features are inevitable to comprehend the mechanism by which the SNPs cause phenotypic changes. Structure-based methods either use potential energy functions for quantifying the mutational effects or machine learning approach (Table 7.2).

Polyphen-2 (Polymorphism phenotyping v2), a successor of Polyphen is a combination tool that uses sequence homology and 3-dimensional structural features for predicting the impact of amino acid substitutions on protein structure and function. Specifically, the functional significance is predicted using Naïve Bayes, a supervised machine learning method. The results are presented as benign or possibly damaging or probably damaging (Adzhubei et al. 2010).

SDM (Site Directed Mutator) uses statistical potential energy function to calculate the score for the effect of SNPs on protein stability and it also predicts the disease-causing propensity of the SNPs. SDM specifically uses environment-specific substitution tables (ESST) that stores probability values for amino acid substitution data (Worth et al. 2011).

PoPMuSiC 2.1 is a free webserver that predicts protein thermodynamic stability changes upon protein mutations. PoPMuSiC can predict the stability changes that a protein may undergo due to all possible amino acid mutations with a high computational speed that is unique to this tool when compared to its counterparts and with better accuracy as well (Dehouck et al. 2011).

mCSM (mutation Cutoff Scanning Matrix) utilizes specific signature such as graph-based interatomic distances, pharmacophore changes and experimental conditions to predict the impact of single point mutations on protein stability as well as the affinity changes towards other proteins and nucleic acid complexes. mCSM has been elaborated into several individual modules such mCSM-lig to study

Table 7.2 Structure-based SNP prediction tools

S. No.	Tool	Significance	URL
1	PoyPhen2 (Adzhubei et al. 2010)	Combined sequence homology and protein structural features for prediction	http://genetics.bwh.harvard.edu/pph2/
2	SDM (Worth et al. 2011)	Uses statistical approach to score the effect of SNP on protein stability.	http://marid.bioc.cam.ac.uk/sdm2/prediction
3	PoPMuSiC (Dehouck et al. 2011)	Predicts optimality of all amino acids along with thermodynamic stability changes	http://babylone.ulb.ac.be/popmusic
4	mCSM (Pires et al. 2014a)	Uses graph-based interatomic atomic distances, pharmacophore changes and experimental conditions to predict on protein stability	http://biosig.unimelb.edu.au/mcsm
5	DUET (Pires et al. 2014b)	A consensus tool combining mCSM and SDM.	http://biosig.unimelb.edu.au/duet
6	MAESTRO (Laimer et al. 2015)	Uses ensemble machine learning model	https://pbwww.che.sbg.ac.at
7	Dynamut (Rodrigues et al. 2018)	Incorporates Normal mode analysis (NMA) and graph-based signatures to analyse the changes in protein dynamics.	http://biosig.unimelb.edu.au/dynamut

protein affinity to small molecules apart from mCSM-NA for studying nucleic acid interactions (Pires et al. 2014a).

DUET is a consensus tool that combines the results of mCSM and SDM to predict the effect of nsSNPs on protein stability. DUET is claimed to be powerful than the individual tools as it optimizes the prediction by collating both the results and then used for predicting using support vector machine (Pires et al. 2014b).

MAESTRO is a versatile tool that can predict stability changes upon protein mutations and also several other applications. MAESTRO uses an ensemble machine learning model comprising of multiple linear regression method, neural network and SVM to compute the ΔΔG values (difference in folding free energy between wild type and mutant) along with a confidence metric for the predictions made. The Protherm, a thermodynamics database for proteins and mutants, was used to train and validate the model. The results of MAESTRO were comparable to that of PoPMuSiC and mCSM (Laimer et al. 2015).

DynaMut analyses changes in protein dynamics that are caused due to mutations. A consensus prediction is obtained by incorporating normal mode analysis (NMA) approaches and graph-based signatures to assess the effect of mutations on protein stability (Rodrigues et al. 2018).

7.5 Future Perspectives

As the sequencing cost decreased, the number of genomes sequenced in humans as well as other organisms increased exponentially opening up vast challenges and opportunities for bioinformaticians and computational biologists. One such opportunity as well as a challenge is the study of variations, especially SNPs. As discussed through the various sections of this chapter, there are countless number of bioinformatic resources to identify, detect, store and functionally analyse SNPs. Even in this chapter many resources could not be included either because they were not updated and abandoned or archived if it is a database. The reasons being lack of funding, manpower, redundant due to replacement with newer resources or majorly incompatible with the latest sequencing technology. Thus, a researcher with minimal expertise in computational aspects and more inclined towards the biological context of SNPs is spoilt for choice and simultaneously ambiguous in using the reliable tools. This issue holds good for any bioinformatic analysis. Especially tools for functional interpretation of SNP effects are found to disagree with their results for the same set of SNPs, leaving the researcher perplexed. This is mainly due to differences in training datasets, as most of the tools incorporate SVMs. The logical solution would be to use a standardized data set. VariBench (Sasidharan Nair and Vihinen 2013; Sarkar et al. 2020) (http://structure.bmc.lu.se/VariBench is a new initiative to address this issue.) is one such effort that offers training/testing datasets with experimentally verified variants curated from literature and other published resources. This is a good opportunity for research community to assess the performance for existing predictor tools as well as to develop new prediction tools. VariSNP (Schaafsma and Vihinen 2015) (http://structure.bmc.lu.se/VariSNP) is also a similar resource, consisting of exclusively disease-causing variants curated from the dbSNP, which can be used for testing the performance of existing tools. There are many tools to predict functional perturbation and disease association of SNPs as discussed in the section of 'sequence-based analysis', on the other hand tools to understand the structural impact of such variants is limited. Hence, there is an unmet need to develop more structure-based analysis tools, which will enable to obtain a better understanding of the disease context and lead to even drug target characterization. This is in turn dependent on the availability of 3D structure of proteins and interactome data. Another area of scope is that currently, for many non-human genome variant analysis users are dependent on tools built exclusively on human variant data. Thus, tools should also be extended to support the analysis of wide range of organisms, apart from humans. In conclusion, the current explosion of bioinformatic tools in SNP analysis, only makes the future bright for better comprehension of the functional significance of the SNPs, provided there is more unified and consistently updated interrogation offered by the current and upcoming tools. To this end initiatives such as the CAGI (Critical Assessment of Genomic Interpretation) (https://genomeinterpretation.org/) offer better hope for the future of SNP analysis using various bioinformatics resources.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249. NIH Public Access. https://doi.org/10.1038/nmeth0410-248

Ahmed S, Zhou Z, Zhou J, Chen SQ (2016) Pharmacogenomics of drug metabolizing enzymes and transporters: relevance to precision medicine. Genomics Proteomics Bioinf 14(5):298–313). Beijing Genomics Institute. https://doi.org/10.1016/j.gpb.2016.03.008

Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet 131(10):1541–1554. https://doi.org/10.1007/s00439-012-1213-z

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Schloss JA (2015) A global reference for human genetic variation. Nature 526(7571):68–74). Nature Publishing Group. https://doi.org/10.1038/nature15393

Baniecki ML, Faust AL, Schaffner SF, Park DJ, Galinsky K, Daniels RF, Hamilton E, Ferreira MU, Karunaweera ND, Serre D, Zimmerman PA, Sá JM, Wellems TE, Musset L, Legrand E, Melnikov A, Neafsey DE, Volkman SK, Wirth DF, Sabeti PC (2015) Development of a single nucleotide polymorphism barcode to genotype plasmodium vivax infections. PLOS Neglected Trop Dis 9(3):e0003539. https://doi.org/10.1371/journal.pntd.0003539

Beck T, Shorter T, Brookes AJ (2020) GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. Nucl Acids Res 48(D1):D933–D940. https://doi.org/10.1093/nar/gkz895

Beres SB, Richter EW, Nagiec MJ, Sumby P, Porcella SF, DeLeo FR, Musser JM (2006) Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group A Streptococcus. Proc Natl Acad Sci USA 103(18):7059–7064. https://doi.org/10.1073/pnas.0510279103

Berger J, Suzuki T, Senti KA, Stubbs J, Schaffner G, Dickson BJ (2001) Genetic mapping with SNP markers in Drosophila. Nat Genet 29(4):475–481. https://doi.org/10.1038/ng773

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protocols Mol Biol 0 19(SUPPL. 89, p. Unit). NIH Public Access. https://doi.org/10.1002/0471142727.mb1910s89

Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, Mari B, Barbry P, Mosnier JF, Hébuterne X, Harel-Bellan A, Mograbi B, Darfeuille-Michaud A, Hofman P (2011) A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet 43(3):242–245. https://doi.org/10.1038/ng.762

Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucl Acids Res 35(11):3823–3835. https://doi.org/10.1093/nar/gkm238

Brookes AJ (1999) The essence of SNPs. Gene 234(2):177–186. https://doi.org/10.1016/S0378-1119(99)00219-X

Budowle B, Van Daal A (2008) Forensically relevant SNP classes. BioTechniques 44(5):603–610. Future Science Ltd London, UK. https://doi.org/10.2144/000112806

Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Parkinson H (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucl Acids Res 47(D1):D1005–D1012. https://doi.org/10.1093/nar/gky1120

Capriotti E, Altman RB, Bromberg Y (2013) Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics 14 Suppl 3(Suppl 3):S2. https://doi.org/10.1186/1471-2164-14-s3-s2

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R (2013) Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3(Suppl 3):S3. https://doi.org/10.1186/1471-2164-14-s3-s3

Chernyaeva EN, Shulgina MV, Rotkevich MS, Dobrynin PV, Simonov SA, Shitikov EA, Ischenko DS, Karpova IY, Kostryukova ES, Ilina EN, Govorun VM, Zhuravlev VY, Manicheva OA, Yablonsky PK, Isaeva YD, Nosova EY, Mokrousov IV, Vyazovaya AA, Narvskaya OV, O'Brien SJ (2014) Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. BMC Genomics 15(1):308. https://doi.org/10.1186/1471-2164-15-308

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 7(10):e46688. https://doi.org/10.1371/journal.pone.0046688

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6(2):80–92. https://doi.org/10.4161/fly.19695

Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG (2014) A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun 5. https://doi.org/10.1038/ncomms5812

Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8(12):1229–1231. https://doi.org/10.1101/gr.8.12.1229

Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. Nature 422(6934):835–847. https://doi.org/10.1038/nature01626

Cotton RGH, Auerbach AD, Axton M, Barash CI, Berkovic SF, Brookes AJ, Burn J, Cutting G, Den Dunnen JT, Flicek P, Freimer N, Greenblatt MS, Howard HJ, Katz M, Macrae FA, Maglott D, Möslein G, Povey S, Ramesar RS, Watson M (2008) Genetics: The human variome project. Science 322(5903):861–862. https://doi.org/10.1126/science.1167363

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dehouck Y, Kwasigroch JM, Gilis D, Rooman M (2011) PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinf 12:151. https://doi.org/10.1186/1471-2105-12-151

Dekkers CMJ (2012) Application of genomics tools to animal breeding. Curr Genomics 13(3):207–212. https://doi.org/10.2174/138920212800543057

den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, Mcgowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PEM (2016) HGVS recommendations for the description of sequence variants: 2016 update. Human Mutat 37(6):564–569. https://doi.org/10.1002/humu.22981

Doddamani D, Khan AW, Katta MAVSK, Agarwal G, Thudi M, Ruperao P, Edwards D, Varshney RK (2015) CicArVarDB: SNP and InDel database for advancing genetics research and breeding applications in chickpea. Database 2015:78. https://doi.org/10.1093/database/bav078

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT (2011) LOVD v.2.0: The next generation in gene variant databases. Human Mutat 32(5):557–563. https://doi.org/10.1002/humu.21438

Garrison E, Marth G (2016) Haplotype-based variant detection from short-read sequencing. Nat Genet 48(6):593–599

Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, Graviss EA, Musser JM (2006) Single-nucleotide polymorphism–based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J Infect Dis 193(1):121–128. https://doi.org/10.1086/498574

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucl Acids Res 33(Database Iss):D514–D517. https://doi.org/10.1093/nar/gki033

Higasa K, Miyake N, Yoshimura J, Okamura K, Niihori T, Saitsu H, Doi K, Shimizu M, Nakabayashi K, Aoki Y, Tsurusaki Y, Morishita S, Kawaguchi T, Migita O, Nakayama K, Nakashima M, Mitsui J, Narahara M, Hayashi K, Matsuda F (2016) Human genetic variation database, a reference database of genetic variations in the Japanese population. J Human Genet 61(6):547–553. https://doi.org/10.1038/jhg.2016.12

Hinrichs AS, Raney BJ, Speir ML, Rhead B, Casper J, Karolchik D, Kuhn RM, Rosenbloom KR, Zweig AS, Haussler D, Kent WJ (2016) UCSC data integrator and variant annotation integrator. Bioinformatics 32(9):1430–1432. https://doi.org/10.1093/bioinformatics/btv766

Hodgkinson A, Eyre-Walker A (2010) Human triallelic sites: evidence for a new mutational mechanism? Genetics 184(1):233–241. https://doi.org/10.1534/genetics.109.110510

Hou Y, Zhao J, Martin W, Kallianpur A, Chung MK, Jehi L, Sharifi N, Erzurum S, Eng C, Cheng F (2020) New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis. BMC Med 18(1):216. https://doi.org/10.1186/s12916-020-01673-z

Hunt R, Sauna ZE, Ambudkar SV, Gottesman MM, Kimchi-Sarfaty C (2009) Silent (synonymous) SNPs: should we care about them? Methods Mol Biol (Clifton, N.J.) 578:23–39. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-60327-411-1_2

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Sieh W (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Human Genet 99(4):877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25(17):2283–2285. https://doi.org/10.1093/bioinformatics/btp373

Kounelis F, Kanterakis A, Kanavos A, Pandi M, Kordou Z, Manusama O, Vonitsanos G, Katsila T, Tsermpini E, Lauschke VM, Koromina M, Spek PJ, Patrinos GP (2020) Documentation of clinically relevant genomic biomarker allele frequencies in the next-generation FINDbase worldwide database. Human Mutat 41(6):1112–1122. https://doi.org/10.1002/humu.24018

Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, Cooper DN (2000) Human gene mutation database - a biomedical information and research resource. Human Mutat 15 (1):45–51. https://doi.org/10.1002/(SICI)1098-1004(200001)15:1<45::AID-HUMU10>3.0.CO;2-T

Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Human Genet 80 (4):727–739. https://doi.org/10.1086/513473

Kwok P-Y, Chen X (2003) Detection of single nucleotide polymorphisms 43 detection of single nucleotide polymorphisms. Issues Mol Biol 5:43–60

Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, Hayden MJ, Forrest KL, Fleury D, Baumann U, Zander M, Mason AS, Batley J, Edwards D (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant Biotechnol J 10(6):743–749. https://doi.org/10.1111/j.1467-7652.2012.00718.x

Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P (2015) MAESTRO - multi agent stability prediction upon point mutations. BMC Bioinf 16(1):116. https://doi.org/10.1186/s12859-015-0548-6

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, Levine R, McEwan P, Morgan MJ (2001) Initial sequencing and analysis of the human genome. Nature 409(6822):860–921. https://doi.org/10.1038/35057062

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucl Acids Res 42(D1):D980. https://doi.org/10.1093/nar/gkt1113

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9 (4):357–359. https://doi.org/10.1038/nmeth.1923

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):25. https://doi.org/10.1186/gb-2009-10-3-r25

Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, Paschall J, Ananiev V, Flicek P, Church DM (2013) DbVar and DGVa: Public archives for genomic structural variation. Nucl Acids Res 41(D1):D936. https://doi.org/10.1093/nar/gks1213

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18(11):1851–1858. https://doi.org/10.1101/gr.078212.108

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009a) The sequence alignment/map format and SAMtools. Bioinformatics 25 (16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li W, Raoult D, Fournier PE (2009b) Bacterial strain typing in the genomic era. FEMS Microbiol Rev 33(5):892–916. Oxford Academic. https://doi.org/10.1111/j.1574-6976.2009.00182.x

Liu X, Jian X, Boerwinkle E (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human Mutat 32(8):894–899. https://doi.org/10.1002/humu.21517

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21(6):936–939. https://doi.org/10.1101/gr.111120.110

Mansur YA, Rojano E, Ranea JAG, Perkins JR (2018) Analyzing the effects of genetic variation in noncoding genomic regions. In: Precision medicine: tools and quantitative approaches. Elsevier Inc, pp 119–144. https://doi.org/10.1016/B978-0-12-805364-5.00007-X

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20 (9):1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F (2016) The ensembl variant effect predictor. Genome Biol 17(1):122. https://doi.org/10.1186/s13059-016-0974-4

Md V, Misra S, Li H, Aluru S (2019) Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: Proceedings - 2019 IEEE 33rd International Parallel and Distributed Processing Symposium, IPDPS 2019, pp 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Moran JL, Bolton AD, Tran PV, Brown A, Dwyer ND, Manning DK, Bjork BC, Li C, Montgomery K, Siepka SM, Vitaterna MH, Takahashi JS, Wiltshire T, Kwiatkowski DJ, Kucherlapati R, Beier DR (2006) Utilization of a whole genome SNP panel for efficient genetic mapping in the mouse. Genome Res 16(3):436–440. https://doi.org/10.1101/gr.4563306

Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science 314(5807):1930–1933. https://doi.org/10.1126/science.1131262

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucl Acids Res 31(13):3812–3814. https://doi.org/10.1093/nar/gkg509

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12(6):443–451. Nature Publishing Group. https://doi.org/10.1038/nrg2986

Nogales A, Dediego ML (2019) Host single nucleotide polymorphisms modulating influenza a virus disease in humans. Pathogens 8(4). MDPI AG. https://doi.org/10.3390/pathogens8040168

Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet 6:1–15. https://doi.org/10.3389/fgene.2015.00235

Pejaver V, Urresti J, Lugo-Martinez J, Pagel K, Lin GN, Nam H-J, Mort M, Cooper D, Sebat J, Iakoucheva L, Mooney S, Radivojac P (2017) MutPred2: inferring the molecular and phenotypic impact of amino acid variants. BioRxiv 134981. https://doi.org/10.1101/134981

Peterson TA, Doughty E, Kann MG (2013) Towards precision medicine: Advances in computational approaches for the analysis of human variants. J Mol Biol 425(21):4047–4063). Academic Press. https://doi.org/10.1016/j.jmb.2013.08.008

Petkov PM, Ding Y, Cassell MA, Zhang W, Wagner G, Sargent EE, Asquith S, Crew V, Johnson KA, Robinson P, Scott VE, Wiles MV (2004) An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res 14(9):1806–1811. https://doi.org/10.1101/gr.2825804

Phillips C, Amigo J, Carracedo Á, Lareu MV (2015) Tetra-allelic SNPs: informative forensic markers compiled from public whole-genome sequence data. Forensic Sci Int Genet 19:100–106. https://doi.org/10.1016/j.fsigen.2015.06.011

Pires DEV, Ascher DB, Blundell TL (2014a) MCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. https://doi.org/10.1093/bioinformatics/btt691

Pires DEV, Ascher DB, Blundell TL (2014b) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucl Acids Res 42(W1):W314. https://doi.org/10.1093/nar/gku411

Pisupati R, Reichardt I, Seren Ü, Korte P, Nizhynska V, Kerdaffrec E, Uzunova K, Rabanal FA, Filiault DL, Nordborg M (2017) Verification of arabidopsis stock collections using SNPmatch, a tool for genotyping high-plexed samples. Sci Data 4(1):1–9. https://doi.org/10.1038/sdata.2017.184

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucl Acids Res 47(D1):D886–D894. https://doi.org/10.1093/nar/gky1016

Rodrigues CHM, Pires DEV, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucl Acids Res 46(W1):W350–W355. https://doi.org/10.1093/nar/gky300

Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics 34 (3):511–513. https://doi.org/10.1093/bioinformatics/btx536

Roos DS (2001) Bioinformatics--trying to swim in a sea of data. Science 291(5507):1260–1261

Sarkar A, Yang Y, Vihinen M (2020) Variation benchmark datasets: update, criteria, quality and applications. Datab J Biolog Databases Curation 2020:117. https://doi.org/10.1093/database/baz117

Sasidharan Nair P, Vihinen M (2013) VariBench: a benchmark database for variations. Human Mutat 34(1):42–49. https://doi.org/10.1002/humu.22204

Schaafsma GC, Vihinen M (2015) VariSNP, a benchmark database for variations from dbSNP. Hum Mutat 36:161–166. https://doi.org/10.1002/humu.22727

Scheben A, Verpaalen B, Lawley CT, Chan CK, Bayer PE, Batley J, Edwards D (2019) CropSNPdb: a database of SNP array data for Brassica crops and hexaploid bread wheat. Plant J 98(1):142–152. https://doi.org/10.1111/tpj.14194

Schwarz JM, Cooper DN, Schuelke M, Seelow D (2014) Mutationtaster2: mutation prediction for the deep-sequencing age. Nat Methods 11(4):361–362. https://doi.org/10.1038/nmeth.2890

Shen YJ, Jiang H, Jin JP, Zhang ZB, Xi B, He YY, Wang G, Wang C, Qian L, Li X, Yu QB, Liu HJ, Chen DH, Gao JH, Huang H, Shi TL, Yang ZN (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol 135(3):1198–1205. https://doi.org/10.1104/pp.103.038463

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) DbSNP: the NCBI database of genetic variation. Nucl Acids Res 29(1):308–311. https://doi.org/10.1093/nar/29.1.308

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM, Gaunt TR (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. Human Mutat 34(1):57–65. https://doi.org/10.1002/humu.22225

Sobrino B, Brión M, Carracedo A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int 154(2–3):181–194. https://doi.org/10.1016/j.forsciint.2004.10.020

Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, Hayden M, Heywood S, Millar DS, Phillips AD, Cooper DN (2020) The human gene mutation database (HGMD®): optimizing its use in a clinical diagnostic or research setting. Human Genet. Springer. https://doi.org/10.1007/s00439-020-02199-3

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Korbel JO (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526(7571):75–81. https://doi.org/10.1038/nature15394

Taillon-Miller P, Gu Z, Li Q, Hillier LD, Kwok PY (1998) Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Res 8(7):748–754. https://doi.org/10.1101/gr.8.7.748

Tan K-K, Tan Y-C, Chang L-Y, Lee KW, Nore SS, Yee W-Y, Mat Isa MN, Jafar FL, Hoh C-C, AbuBakar S (2015) Full genome SNP-based phylogenetic analysis reveals the origin and global spread of Brucella melitensis. BMC Genomics 16(1):93. https://doi.org/10.1186/s12864-015-1294-x

Tang H, Thomas PD (2016) PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. Bioinformatics 32(14):2230–2232. https://doi.org/10.1093/bioinformatics/btw222

van Baal S, Zlotogora J, Lagoumintzis G, Gkantouna V, Tzimas I, Poulas K, Tsakalidis A, Romeo G, Patrinos GP (2010) ETHNOS: a versatile electronic tool for the development and curation of national genetic databases. Human Genomics 4(5):361–368. https://doi.org/10.1186/1479-7364-4-5-361

Vihinen M (2014) Variation Ontology for annotation of variation effects and mechanisms. Genome Res 24(2):356–364. https://doi.org/10.1101/gr.157495.113

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Lander ES (1998) Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. Science 280(5366):1077–1082. https://doi.org/10.1126/science.280.5366.1077

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucl Acids Res 38(16):e164. https://doi.org/10.1093/nar/gkq603

Wang Y, Zhang M-M, Huang W-W, Wu S-Q, Wang M-G, Tang X-Y, Sandford AJ, He J-Q (2018) Polymorphisms in toll-like receptor 10 and tuberculosis susceptibility: evidence from three independent series. Front Immunol 9:309. https://doi.org/10.3389/fimmu.2018.00309

Wang Z, Moult J (2001) SNPs, protein structure, and disease. Human Mutat 17(4):263–270. https://doi.org/10.1002/humu.22

Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, Gerdes S, Henry CS, Kenyon RW, Machi D, Mao C, Nordberg EK, Olsen GJ, Murphy-Olson DE, Olson R, Stevens RL (2017) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucl Acids Res 45(D1):D535–D542. https://doi.org/10.1093/nar/gkw1017

Wildeman M, Van Ophuizen E, Den Dunnen JT, Taschner PEM (2008) Improving sequence variant descriptions in mutation databases and literature using the mutalyzer sequence variation nomenclature checker. Human Mutat 29(1):6–13. https://doi.org/10.1002/humu.20654

Wilkening S, Tekkedil MM, Lin G, Fritsch ES, Wei W, Gagneur J, Lazinski DW, Camilli A, Steinmetz LM (2013) Genotyping 1000 yeast strains by next-generation sequencing. BMC Genomics 14(1):90. https://doi.org/10.1186/1471-2164-14-90

Worth CL, Preissner R, Blundell TL (2011) SDM - A server for predicting effects of mutations on protein stability and malfunction. Nucl Acids Res 39(SUPPL. 2):W215–W222. https://doi.org/10.1093/nar/gkr363

Xu C (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J 16:15–24). Elsevier B.V. https://doi.org/10.1016/j.csbj.2018.01.003

Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG (2011) A probabilistic disease-gene finder for personal genomes. Genome Res 21(9):1529–1542. https://doi.org/10.1101/gr.123158.111

Zhang F, Lupski JR (2015) Non-coding genetic variants in human disease. Human Mol Genet 24 (R1):R102–R110. Oxford University Press. https://doi.org/10.1093/hmg/ddv259

8 Vaccine Design and Immunoinformatics

Fariya Khan and Ajay Kumar

Abstract

The emanation use of vaccines has shown tremendous applications of computational algorithms that can be used for amelioration of health globally. Vaccine Research has become a center area of research that embarks its applications to save several lives, reduced cost of treatment, and potential inhibitor of infectious diseases. The stimulating progress of immunoinformatics approach with the concept of peptide vaccines has proven to be productive way to target unknown antigenic proteins, complex life-cycle of infectious diseases, variability of immune system response, and long term protection. This Chapter reviews the comprehensive database analysis for the construction of vaccine design targeting epitope based approach which has proven to be a very robust method for the characterization of vaccine targets for systemic models of vaccine. The design of vaccine from traditional to computational methods enables to understand the complexity of disease causing organisms and their hyper variable nature. The investigations of vaccine include rigorous methods that validate the designed vaccine to be antigenic, immunogenic, and non-allergenic and higher solubility and furthermore predicted designed vaccine should have the capability to trigger high immune responses. The docking and simulation of the predicted peptides provide insight information of the binding energy and the stability of vaccine candidates for a better accuracy.

Keywords

Immunoinformatics · Algorithm · Epitopes · T cell · B cell · Vaccine candidate · Molecular docking and simulation

F. Khan · A. Kumar (✉)
Department of Biotechnology, Faculty of Engineering and Technology, Rama University, Uttar Pradesh Kanpur, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_8

8.1 Introduction

Immunoinformatics has emerged as a new way for vaccine development as it enables to study infectious diseases whose treatment is not obtainable. Vaccination is the most powerful method to eradicate chronic infections and drug resistant pathogens and therefore substantial growth of vaccine development has increased (Angus et al. 2020). Immunoinformatics is a combination of computational software along with immunology for developing more robust immunogenic vaccine that can provide protection worldwide against diseases. Several literatures reported the targeting of conserved known antigens for vaccine prediction but it is established that most antigens are not conserved and showed hypervariability nature. The complete genomic information of pathogen is not enough to predict all the potential proteins that cause diseases thus proteomics can be helpful in identifying and differentiating all the immunogenic proteins that can be used as important inhibitors in vaccine designing. Therefore, to target complete proteome can be the best choice to predict epitopes that has salient role in disease diagnosis (Arafat et al. 2017).

An ideal vaccine should trigger an effective immune response as immunity has been categorized into cell mediated and humoral immunity. So, it should be important to recognize the antigens by specific receptors that can induce signals needed for the stimulating of immune system. B and T cells triggers adaptive immunity as pathogens cannot be recognized by B and T cell alone as antigen recognition for these cells differ greatly (Huber Sietske et al. 2014). T cells recognize antigens when they are bound on the surface of Major Histocompatibility Complex (MHC) molecules. T cell epitopes can be presented by MHC class I and MHC class II molecules. MHC-I molecules can predict peptides ranging from 9 to 11 residues therefore most peptide ligands have fixed 9 residues due to its deep binding groove. The extent of the peptide for MHC class II molecules ranges from 9 to 22 residues as the binding groove of the peptide is open (Lippolis et al. 2002).

B cells recognize pathogens through B cell receptors (BCR) by releasing antibodies that mediate humoral immunity and the antibodies which they released play different function in order to destroy the pathogens or toxins. Studies suggested that B and T cell epitope prediction portray wide applications in designing a vaccine that can be effective in triggering a better immunity (Clarisa et al. 2018).

B cell epitopes can be characterized by several approaches but traditional epitope identification method has been completely dependent on experimental methodology that is costly as well as time-consuming. Therefore, *in silico* approach facilitates epitope prediction and has decreased the cost of experimental analysis along with better accuracy of prediction (Khan et al. 2017).

Immunoinformatics approach is the most investigated approach in the last few years as it is a combination of high throughput screening, homology modeling, molecular docking, and molecular simulation. The use of computational to retrieve the genomic data has made a remarkable progress in the era of vaccines development and was termed as Reverse Vaccinology (Alessandro and Rino 2010). The accuracy of prediction to determine the immunogenic antigens that has the utmost potential to robust protective immunity was inserted for vaccine development. It has been challenging to discover unknown proteins that have a role in pathogenesis of any diseases but with the help of this approach novel antigens being predicted and studied (Khan et al.). The most important complication in epitope based prediction method is polymorphism of MHC alleles as there are thousands of different alleles that are associated with human leukocyte antigens (HLAs) and each alleles bind with different peptides (Backert and Kohlbacher 2015). Therefore, synthesis and prediction of all peptides are practically not possible so scientists have developed advanced MHC prediction methods that have the capability to predict peptide that binds to non-characterized HLA alleles (Hamrouni et al. 2020). Thus, characterization of peptides binding to HLA alleles led to the construction of T cell epitope based vaccine with maximum coverage population.

This review chapter describes the use of immunoinformatics vaccine designing approach by using different tools that scrutinize and predict most potential epitopes. Recent studies suggest that T cell epitopes have been a key contributor in vaccine development for stimulating a protective immunity (Jensen et al. 2018). The use of computational vaccinology allows studying the complex nature of pathogens and retrieving the data of genomic information through databases thus led to the designing of vaccine more easily, precise, and accurate. There are different web based tools that can be used to screen immunogenic antigens, predict MHC binding peptides, determine population coverage, binding energy evaluation, mapping of epitopes, three-dimensional structural modeling and simulation of peptides over a period of time (Monterrubio-López and Ribas-Aparicio 2015). So, here is an outline of some of the most used online computational software that aid in the vaccine development for epitope based vaccines.

8.2 Immunoinformatics in Vaccine Discovery and Infectious Diseases

The rapid emergence of infectious diseases spreading in all geographical range leads to the disease outbreak thus urges demand of advanced research and development in public health sector. Development of vaccines has become utmost need globally due to the factors of resistant of drugs for infectious diseases by immune system and therefore required advanced method to save public life (Pahil et al. 2017). Thus, immunoinformatics examines infectious diseases by understanding its mode of pathogenesis, genome information, outbreak of widely spread pathogenic strain, and structural modeling (Li et al. 2014). Therefore it was found that immunoinformatics helps to find the best vaccine candidates using different computational model ruling out traditional analysis that involves isolating and cultivating infectious pathogens (Fig. 8.1).

This chapter on immunoinformatics will help the researchers to illustrate the use of advanced computer based technology in the development of vaccine that could be effective against hypervariable pathogens. Epitope based vaccine has become a key contributor in the change towards traditional vaccine design concept and strong long term immunity (Tomar and De 2010). The application of new and updated tools

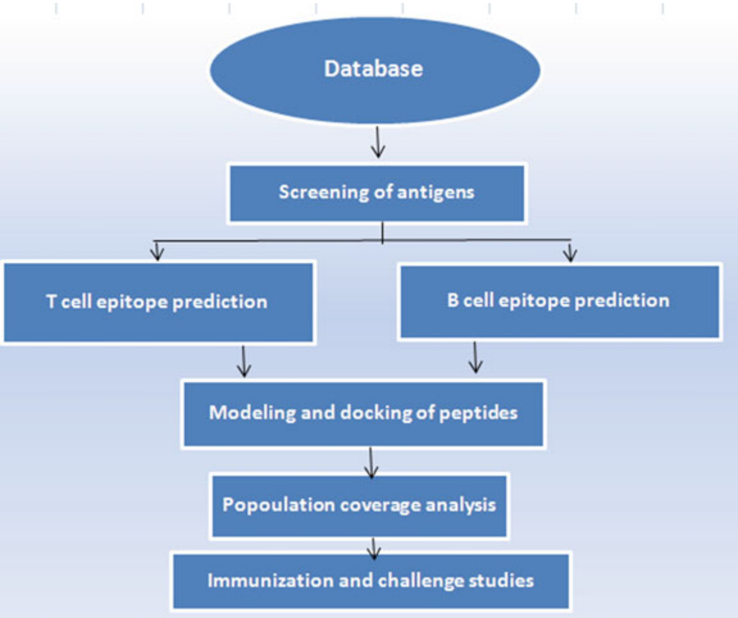

Database
Screening of antigens
T cell epitope prediction
B cell epitope prediction
Modeling and docking of peptides
Popoulation coverage analysis
Immunization and challenge studies

Fig. 8.1 Flowchart representing the immunoinformatics approaches for vaccine designing

screens the set of immunogenic antigens from pool of genes thus only promising epitopes can be forwarded to rigorous steps for binding capability (Khan et al. 2018). Screening of the epitopes is also a very important and first parameter in designing vaccine and hence different tools are available for it (Xiang and He 2009; Irini and Darren 2007). As epitopes should not be allergic in nature, they should be toxic, it should be highly antigenic and adhesive, antigen should be stable as well as soluble and localization of all potential proteins (Table 8.1).

8.3 Computational Databases for Prediction of T Cell Epitopes

The first step in vaccine designing is the selection of immunodominant epitope from a large pool of antigens using different immunoinformatics tools. Since, T cell epitopes can only be identified when it is linked to MHC molecules and therefore tools based on these parameters have been employed to predict putative T cell epitopes (Oyarzun and Kobe 2015). The accuracy rate of MHC-I binding epitopes is estimated to be 90–95% that is very effectual for wide coverage of alleles and therefore different servers were used. NetCTL 1.2 server is used to find cytotoxic T-lymphocyte (CTL) epitopes from pathogenic proteins (Larsen et al. 2007). It can characterize epitopes that bind to 12 supertypes A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, and B62 and work on the parameters of artificial neural networks, SMM, and scoring matrices. NetMHC pan 4.1 servers are the most updated version

Table 8.1 List of different Algorithms for screening of epitopes

Tools name	Link	Prediction mode
Vaxign Server	http://www.violinet.org/vaxign/index.php	Screen complete proteome
VaxiJen tool	http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html	Predicts antigens
AllergenFP	http://www.ddg-pharmfac.net/AllergenFP/index.html	Predicts allergen and non-allergens
SPAAN	http://www.violinet.org/vaxign/docs/index.php	Predicts adhesive antigens
ProtParam	https://web.expasy.org/protparam/	Predicts physiochemical properties of protein
ToxinPred	http://crdd.osdd.net/raghava/toxinpred/	Predicts toxicity
AlgPred	http://crdd.osdd.net/raghava/algpred/	Predicts allergen and non-allergens

Table 8.2 Bioinformatics Tools for prediction of T cell epitopes

Server name	Link	Prediction mode
ProPred	http://crdd.osdd.net/raghava/propred/	Predicts MHC-II binding epitopes
ProPred I	http://crdd.osdd.net/raghava/propred1/	Predicts MHC-I binding epitopes
NetCTL 1.2	http://www.cbs.dtu.dk/services/NetCTL/	Predicts CTL and MHC-I binding epitopes
NetMHC pan 4.1	http://www.cbs.dtu.dk/services/NetMHCpan/	Predicts MHC- I binding epitopes
IEDB tool	http://tools.iedb.org/mhci/	Predicts MHC-I and MHC-II binding epitopes
MHCPred 2.0	http://www.ddg-pharmfac.net/mhcpred/MHCPred/	Predicts MHC-I and MHC-II binding epitopes
CTLPred	http://crdd.osdd.net/raghava/ctlpred/	Predicts CTL epitopes
MHC2Pred	http://crdd.osdd.net/raghava/mhc2pred/	Predicts MHC-II epitopes

that is also employed for the quantitative prediction of peptides that binds to MHC alleles using artificial neural networks (ANNs). This method predicts epitopes on the basis of quantitative Binding Affinity (BA) along with Mass Spectrometry Eluted Ligands (EL) peptides (Birkir et al. 2020). It predicts the peptides based on the threshold value that indicates 0.5 for strong binding epitopes and 2 for weak binding epitopes.

The epitopes can also be analyzed by IEDB analysis tool that uses stabilized matrix method (SMM) for identification of MHC-I binding alleles (Vita et al. 2018). Here, peptides having IC50 less than 22 nm were considered to be strong binders with higher affinity and it works on the parameter of MHC-I binding score, proteasomal cleavage, and TAP score. Based on these criteria, the best vaccine candidate was selected for conservancy analysis (Table 8.2).

ProPred I server is widely used for MHC class I binding epitopes based on matrix based method that facilitate prediction of the binding regions in an antigen for a total of 47 MHC-I molecules (Singh and Raghava 2003). This method uses quantitative analysis to investigate the interaction of antigens with alleles belonging to MHC class I with great binding affinity. Although the greater prediction rate has been obtained from MHC-I prediction method but there are databases for identifying the epitopes that binds to MHC class II alleles also. MHC2Pred tool is used for the prediction of dominant epitopes that bind to Class II MHC alleles as this tool is based on SVM method therefore the accuracy was estimated as ~80% for 42 alleles. MHCPred 2.0 is used to predict epitopes that bind to both MHC class I and II alleles and generates Quantitative Structure Activity Relationship (QSAR) models (Guan et al. 2003). It is also used to predict Transporter associated with Processing (TAP) scores of the peptide. The performance of this tool was approved by 5-fold cross validation and the epitopes having value of IC50 less than 500 nm were considered to be as binders and vice versa.

Another immunoinformatics server ProPred is used to find out the antigen zone in protein sequence and is placed on quantitative matrix based method (Singh and Raghava 2001). SVMHC server also allows the prediction of T cell epitopes for both class I and class II MHC alleles but mostly it is used for MHC-II binding prediction along with the effects of single nucleotide polymorphisms (SNPs) (Pierre and Oliver 2006).

8.4 Computational Databases for Prediction of B Cell Epitopes

B cell epitopes provide long term immunity and can be used against several diseases but the mostly B cell epitopes are not continuous in sequence therefore prediction completely varies from T cell epitope prediction method (Zobayer et al. 2019). Previous literature depicts that continuous epitopes are easier to predict in comparison to discontinuous one and based on these parameters different tools have been designed (Krawczyk et al. 2014). The identification of continuous B cell epitopes is entirely based on physiochemical characteristics like charge, hydrophilicity, polarity, flexibility, and secondary structure. BcePred tool was used to evaluate B cell epitopes on the basis of their physiochemical properties and the accuracy of this tool was retrieved as 58.70% at threshold value of 2.38 (Saha and Raghava 2004). The ABCPred server is used to identify potential epitopes based on artificial neural network with accuracy of 65.93% and portray the result in tabular as well as graphical form (Saha and Raghava 2006). BepiPred is also used to predict B cell epitopes from a sequence of protein applying random forest algorithm method and is a very valuable tool in bioinformatics for analysis (Jespersen et al. 2017).

To predict the discontinuous B cell epitopes, DiscoTope is also considered to be a novel method to evaluate surface accessibility and propensity score of amino acid using protein three-dimensional structures (Jens Vindahl et al. 2012). B cell epitopes has a significant role in disease treatment as well as other immune therapy and thus ElliPro tool is used to identify discontinuous epitopes on the basis of 3D

Table 8.3 Bioinformatics Tools for prediction of B cell epitopes

Server name	Link	Prediction mode
BepiPred	http://www.cbs.dtu.dk/services/BepiPred/	Predicts continuous B cell epitopes
BcePred	http://www.imtech.res.in/raghava/bcepred/	Predicts continuous B cell epitopes
ABCPred	http://www.imtech.res.in/raghava/abcpred/	Predicts continuous B cell epitopes
ElliPro	http://tools.immuneepitope.org/tools/ElliPro/iedb_input	Predicts discontinuous B cell epitopes
Epi Search	http://curie.utmb.edu/episearch.html	Predicts discontinuous B cell epitopes
DiscoTope	http://www.cbs.dtu.dk/services/DiscoTope/	Predicts discontinuous B cell epitopes
SEPPA	http://lifecenter.sgst.cn/seppa/index.php	Predicts discontinuous B cell epitopes
PEPITO	http://pepito.proteomics.ics.uci.edu/	Predicts discontinuous B cell epitopes

structure of proteins. It calculates the score of epitope residues in form of Protrusion Index (PI) and epitopes that scores higher are considered to be of great solvent accessibility (Ponomarenko et al. 2008). Episearch tool is used for mapping of discontinuous epitopes from phage display sequences. It works on patch analysis that characterizes similarity between residues on the antigen surface with physio-chemical properties of phage display sequences (Negi and Braun 2009). The highest score of Episearch tool was found to be more than 50% in all the test cases.

Thus, these different tools are very helpful in prediction continuous and discontinuous B cell epitopes and therefore can be a needful algorithm in the development of vaccine designing (Table 8.3).

8.5 Prediction of T Cell Epitope Modeling

The major obstacle in designing vaccine construct is to understand the functional and insight properties of proteins. Thus in order to overcome this issue, different computational servers have been designed to homology modeling of epitopes (Kaur et al. 2007). Several bioinformatics tools have been employed to construct three-dimensional modeling of vaccine targets to retrieve the complete information of proteins (Aurelien et al. 2011). Modeler is the most popular tool for comparative based modeling of proteins and generates 5 models by comparing the target sequence alignment with template structure (Fiser et al. 2002). Phyre 2 is used to build 3-dimensional model using HMM (Hidden Markov Models) as well as identify binding site of ligand (Kelley et al. 2015). All the 3d coordinates files of HLA alleles were taken from IPD-IMGT/HLA Database (Robinson et al. 2015). RaptorX is used

Table 8.4 List of different computational tools for modeling of epitopes and alleles

Tools	Link	Prediction mode
Modeler 9.19	https://salilab.org/modeller/9.19/release.html	3 D modeling of Proteins
PEPstrMOD	https://webs.iiitd.edu.in/raghava/pepstrmod/	3 D modeling of Epitopes
Raptor X	http://raptorx.uchicago.edu/	3 D modeling of Proteins
SWISS-MODEL	https://swissmodel.expasy.org/	3 D modeling of Proteins
I-TASSER	https://zhanglab.ccmb.med.umich.edu/I-TASSER/	Predicts protein structure and functions
PyMOL	https://pymol.org/2/	Molecular visualization of proteins and epitopes
Chimera 1.12	http://www.rbvi.ucsf.edu/chimera/download.html	Molecular visualization of proteins and epitopes
Phyre2	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index	3 D modeling of Proteins
ERRAT	https://servicesn.mbi.ucla.edu/ERRAT/	Validation of 3D modeled structure
RAMPAGE	http://services.mbi.ucla.edu/SAVES/Ramachandran/	Validation of 3D modeled structure
PROCHECK	https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/	Validation of 3D modeled structure
IPD-IMGT/HLA	https://www.ebi.ac.uk/ipd/imgt/hla/	Search sequence of alleles

to analyze the tertiary and secondary structure and functional information of the binding region of proteins (Morten et al. 2012).

The designed model can be validated by using different servers that provide validation score and approves the quality of designed model. The list of servers that can be used for validation include RAMPAGE, PROCHECK, and ERRAT. RAMPAGE is referred as Ramachandran Plot analysis which is used to read insight structural information of proteins by calculating torsional angles. PROCHECK is also used to analyze the quality of model on the basis of their stereochemistry and it compares the model with known define model of high resolution (Laskowski et al. 1993) (Table 8.4).

8.6 Multi-Epitope Vaccine Design as a Promising Approach

Multi-epitope vaccine is a better approach compared to single epitope vaccines as it has the distinctive features. Compared to the properties of single B and T cell epitope vaccine design, it consists of a combination of B cell epitopes, CTL and HTL epitopes that provides a powerful immune response. It also has adjuvant that improves the immunity level and provides a longer immune response as well as vaccine candidates that should be immunogenic and antigenic (Kaur et al. 2020).

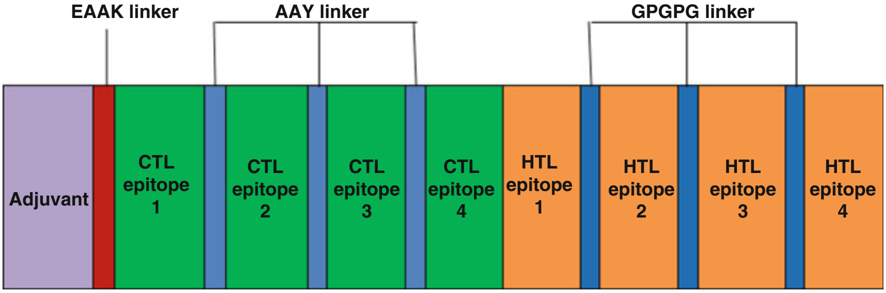


Fig. 8.2 Schematic diagram of Multi-epitope vaccine design construct

The binding capability of epitopes to HLA alleles should be strong binder and highly promiscuous in nature. Therefore, the combination of multiple epitopes with the help of linkers and adjuvant will make a tremendous effect on immunity of an individual (Fig. 8.2).

To make an effective multi-epitope vaccine, it is very crucial to characterize the key antigens and epitopes that have the great ability to trigger immunity. Thus, immunoinformatics tools can predict B and T cell epitopes of the targeted strain based on survey and then these epitopes will be merge together with the help of linkers and unique adjuvant to increase immunity level (Lifhang Zhang 2018). Therefore, the use of computational algorithm predicts peptide with higher antigenic score, MHC binding score, solubility, epitope conservancy, and covering large number of population (Nezafat et al. 2014). Linkers play a very unique role in combining all the CTL, HTL, and B cell epitopes with an adjuvant and provide better flexibility and protein folding thus making protein more stable. The linkers that are mostly used are EAAAK, AAY, and GPGPG linker based on the review of previous papers (Pandey et al. 2018).

8.7 Docking and Simulation of Peptides to Enhance Vaccine Design Approach

Docking is the most eminent step in designing of vaccine as it is a powerful way to determine the binding interaction between the peptides and alleles. It indicates the position of amino acid and bond between ligand and receptor thus higher the binding interaction energy, the more effective vaccine is. In immunoinformatics approach, docking in vaccine design has proven to be the best algorithm to predict the strong binders on the basis of their binding energy value (Schneidman-Duhovny et al. 2005). There are several tools that are used to evaluate the binding energy calculation and here will discuss some of the algorithms (Duhovny et al. 2002). Autodock 4 is a very popular tool used for docking calculation which is based on linear regression analysis (Goodsell et al. 1998; Morris et al. 2009). This tool provides high quality result and good binding affinity and therefore widely used in vaccine as

Table 8.5 List of different tools for docking and molecular simulation study

Server name	Link	Prediction mode
Autodock 4.2	http://autodock.scripps.edu/downloads	Docking of epitope and alleles
Autodock Vina	http://vina.scripps.edu/	Docking of epitope and alleles
PatchDock	https://bioinfo3d.cs.tau.ac.il/PatchDock/php.php	Docking of epitope and alleles
SwissDock	http://www.swissdock.ch/docking	Docking of epitope and alleles
NAMD tool	https://www.ks.uiuc.edu/Research/namd/	Molecular Simulation
CHARMM	https://www.charmm.org/	Molecular Simulation
MDWeb	https://mmb.irbbarcelona.org/MDWeb/	Molecular Simulation

well as drug design also. Autodock Vina is recent and updated version for predicting molecular docking (Trott and Olson 2010). It uses structural file of molecule in PDBQT format and generates the PDBQT file with the help of MGL tools. Autodock Vina is advanced tool with better performance and is faster compared to Autodock 4 version. Epidock is also used for molecular docking mainly for MHC class II predictions and it includes 23 alleles for generating the output. It accepts input sequence of protein in a single letter code and translates into nonamers for which scores are generated in output file (Atanasova et al. 2013; Patronov et al. 2011).

Along with the docking, molecular dynamics simulation allows to understand the complex stability of the designed vaccine candidates. As it is important to know the molecular details of the molecules and their stability in water over a certain period of time (Xu and Zhang 2012). NAMD tool is used for simulation of molecules along with VMD (Visual Molecular Dynamics) especially designed for high performance rate (James et al. 2005; Humphrey et al. 1996). It generates the results in trajectory DCD files by retrieving the information from PDB file.

MDWeb built a friendly environment for molecular dynamic simulation by analyzing trajectory input file of molecule with the help of Gromacs package (Adam et al. 2012). It generates result in the graphical form of RMSD plot and Radius of Gyration which allows the user to identify which have the maximum number of mobility and regions that has shown stability during simulation. Thus, with the help of these plots the stability of the selected vaccine candidate can be accessed and can be suggested further for experimental confirmation (Table 8.5).

8.8 Conclusion

Thus, *in silico* analytical methods have a powerful impact for vaccine discovery for novel antigens for precise immune response. This method has allowed the researchers to strengthen broad spectrum research in epitope based vaccine that has overshadowed traditional methods in the last few decades. Immunoinformatics technology has become a boon in medical science research to study infectious diseases, cancer, personalized medicine, and allergy. Although there are certain limitations also related to immunoinformatics approaches that also cannot be neglected in terms of handling real data and therefore predicted vaccine epitopes should undergo experimental analysis in animal model.

References

Adam H, Andrio P, Fenollosa C, Cicin-Sain D, Orozco M, Gelpi JL (2012) MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. Bioinformatics 28 (9):1278–1279

Alessandro S, Rino R (2010) Reverse vaccinology: developing vaccines in the era of genomics. Immunity 33(4):530–541

Angus NO, Obialor WO, Ifeanyichukwu MO, Odimegwu DC, Okoyeh JN, Emechebe GO, Adejumo SA, Ibeanu GC (2020) Immunoinformatics and vaccine development: an overview. Immunotargets Ther 9:13–30

Arafat RO, Pervin T, Mia M, Hossain M, Shahnaij M, Mahmud S, Kaderi Kibria KM (2017) Vaccinomics approach for designing potential peptide vaccine by targeting Shigella spp. Serine protease autotransporter subfamily protein SigA. J Immunol Res. https://doi.org/10.1155/2017/6412353

Atanasova M, Dimitrov I, Flower DR, Doytchinova I (2013) EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. Protein Eng Des Sel 26(10):631–634

Aurelien G, Zoete V, Michielin O (2011) SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucl Acids Res 39:W270–W277

Backert L, Kohlbacher O (2015) Immunoinformatics and epitope prediction in the age of genomic medicine. Genome Med 7(1):119

Birkir R, Alvarez B, Paul S, Peters B, Nielsen M (2020) NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucl Acids Res. https://doi.org/10.1093/nar/gkaa379

Clarisa B P-d-S, Soares I d S, Rosa DS (2018) Editorial: epitope discovery and synthetic vaccine design. Front Immunol 9:826

Duhovny D, Nussinov R, Wolfson HJ (2002) Efficient unbound docking of rigid molecules. proceedings of the 2'nd workshop on algorithms in bioinformatics (WABI) Rome, Italy. Lecture Notes in Computer Science, vol 2452. Springer, pp 185–200

Fiser A, Feig M, Brooks CL, Sali A (2002) Evolution and physics in comparative protein structure modeling. Acc Chem Res. 35:413–421. https://doi.org/10.1021/ar010061h

Goodsell DS, Morris GM, Halliday RS, Huey R, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. J Comp Chem 19:1639–1662

Guan P, Doytchinova IA, Zygouri C, Flower DR (2003) MHCPred: bringing a quantitative dimension to the online prediction of MHC binding. Appl Bioinf 2:63–66

Hamrouni S, Bras-Gonçalves R, Kidar A, Aoun K, Chamakh-Ayari R, Petitdidier E, Messaoudi Y, Pagniez J, Lemesre JL, Meddeb-Garnaoui A (2020) Design of multi-epitope peptides containing HLA class-I and class-II-restricted epitopes derived from immunogenic Leishmania proteins, and evaluation of CD4+ and CD8+ T cell responses induced in cured cutaneous leishmaniasis subjects. PLoS Negl Trop Dis 14(3):e0008093

Huber Sietske R, van Beek J, de Jonge J, Luytjes W, van Baarle D (2014) T cell responses to viral infections – opportunities for peptide vaccination. Front Immunol 5:171

Humphrey W, Dalke A, Schulten K (1996) VMD—visual molecular dynamics. J Mol Graphics 14:33–38

Irini AD, Darren RF (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinf 8:4

James CP, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K (2005) Scalable molecular dynamics with NAMD. J Comp Chem 26 (16):1781–1802

Jens Vindahl K, Lundegaard C, Lund O, Nielsen M (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comp Biol 8(12): e1002829

Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, Sette A, Peters B, Nielsen M (2018) Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology 154(3):394–406

Jespersen MC, Peters B, Nielsen M, Marcatili P (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucl Acids Res. https://doi.org/10.1093/nar/gkx352

Kaur H, Garg A, Raghava GPS (2007) PEPstr: A de novo method for tertiary structure prediction of small bioactive peptides. Protein Pept Lett 14(7):626–630

Kaur R, Arora N, Jamakhani MA, Malik S, Kumar P, Anjum F, Tripathi S, Mishra A, Prasad A (2020) Development of multi-epitope chimeric vaccine against Taenia solium by exploring its proteome: an in silico approach. Exp Rev Vaccines 19(1):105–114

Kelley L, Mezulis S, Yates C et al (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protocol 10(6):845–858

Khan F, Srivastava V, Kumar A (2017) Epitope-based peptides prediction from proteome of *Enterotoxigenic E coli*. Int J Peptide Res Ther 24(2):323–336

Khan F, Srivastava V, Kumar A (2018) Computational identification and characterization of potential T-Cell epitope for the utility of vaccine design against Enterotoxigenic *Escherichia coli*. Int J Peptide Res Ther (Springer) 25:289–302

Krawczyk K, Liu X, Baker T, Shi J, Deane CM (2014) Improving B-cell epitope prediction and its application to global antibody-antigen docking. Bioinformatics 30(16):2288–2294

Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinf 8:424

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK - a program to check the stereochemical quality of protein structures. J App Cryst 26:283–291

Li W, Joshi MD, Singhania S, Ramsey KH, Murthy AK (2014) Peptide vaccine: progress and challenges. Vaccine 2(3):515–536

Lippolis JD et al (2002) Analysis of MHC class II antigen processing by quantitation of peptides that constitute nested sets. J Immunol 169:5089–5097

Monterrubio-López GP, Ribas-Aparicio RM (2015) Identification of novel potential vaccine candidates against tuberculosis based on reverse vaccinology. Biomed Res Int 12:1–16

Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 16:2785–2791

Morten K, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protocols 7:1511–1522

Negi SS, Braun W (2009) Automated detection of conformational epitopes using phage display peptide sequences. Bioinform Biol Insights 3:71–81

Nezafat N, Ghasemi Y, Javadi G, Khoshnoud MJ, Omidinia E (2014) A novel multi-epitope peptide vaccine against cancer: an *in silico* approach. Theor Biol 349:121–134

Oyarzun P, Kobe B (2015) Computer-aided design of T-cell epitope-based vaccines: addressing population coverage. Int J Immunogenet 42(5):313–321

Pahil S, Taneja N, Ansari HR, Raghava GPS (2017) In silico analysis to identify vaccine candidates common to multiple serotypes of Shigella and evaluation of their immunogenicity. PLoS One 12:8

Pandey RK, Ojha R, Aathmanathan VS, Krishnan M, Prajapati VK (2018) Immunoinformatics approaches to design a novel multiepitope subunit vaccine against HIV infection. Vaccine 36:2262–2272. https://doi.org/10.1016/j.vaccine.2018.03.042

Patronov A, Dimitrov I, Flower DR, Doytchinova I (2011) Peptide binding prediction for the human class II MHC allele HLA-DP2: a molecular docking approach. BMC Str Biol 11:32

Pierre D, Oliver K (2006) SVMHC: a server for prediction of MHC-binding peptides. Nucl Acids Res 34:W194–W197

Ponomarenko JV, Bui H, Li W, Fusseder N, Bourne PE, Sette A, Peters B (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinf 9:514

Robinson J, Halliwell JA, Hayhurst JH, Flicek P, Parham P, Marsh SGE (2015) The IPD and IMGT/HLA database: allele variant databases. Nucl Acids Res 43:D423–D431

Saha S, Raghava GPS (2004) BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. In: Nicosia G, Cutello V, Bentley PJ, Timmis J (eds) Artificial immune systems. ICARIS 2004. Lecture Notes in Computer Science, vol 3239. Springer, Berlin. https://doi.org/10.1007/978-3-540-30220-9_16.

Saha S, Raghava GPS (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins Struct Funct Bioinf 65:40–48

Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucl Acids Res 33:W363–W367

Singh H, Raghava GPS (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17 (12):1236–1237

Singh H, Raghava GPS (2003) ProPred I: prediction of HLA class-I binding sites. Bioinformatics 19:1009–1014

Tomar N, De RK (2010) Immunoinformatics: an integrated scenario. Immunology 131(2):153–168

Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Comput Chem:455–461

Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B (2018) The Immune epitope database (IEDB). Nucl Acids Res. https://doi.org/10.1093/nar/gky1006

Xiang Z, He Y (2009) Vaxign: a web-based vaccine target design program for reverse vaccinology. Proc Vaccinol 1(1):23–29

Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80(7):1715–1735

Zhang L (2018) Multi-epitope vaccines: a promising strategy against tumors and viral infections. Cell Mol Immunol 15:182–184

Zobayer N, Hossain AA, Rahman MA (2019) A combined view of B-cell epitope features in antigens. Bioinformation 15(7):530–534

Computer-Aided Drug Designing

9

Thakor Rajkishan, Ailani Rachana, Surani Shruti, Patel Bhumi, and Dhaval Patel

Abstract

The long-used traditional methodology for novel drug discovery and drug development is an immensely challenging, multifaceted, and prolonged process. To overcome this limitation, a new advanced approach was developed and adopted over time which is referred to as computer-aided drug discovery (CADD). Over the course, CADD has accelerated the overall traditional time-consuming process of new drug entity development with the advancement of computational tools and methods. Recently CADD methodologies have become a fundamental and indispensable tool in different junctures of drug discovery and development. Additionally, with the increasing availability and knowledge of biological/biomacromolecule structures, as well as an exponential increase in computing power, it is now plausible to use these methods effectively without the significant loss of accuracy. CADD has also paved paths for the screening of selected compounds and synthesis of those entities for better therapeutics. Therefore, CADD is continuously employed with the collective biological and chemical knowledge to rationalize lead optimization, design, and thus can be effectively used in different stages of the discovery and development pipeline. Over the past decades, various CADD techniques such as homology modeling, docking, pharmacophore modeling based virtual screening, conformation generation, ab initio design, toxicity profile, quantitative structure–activity relationship (QSAR), and quantitative free-energy calculation have been greatly improved. The current methods of CADD are routinely utilized in academic and commercial research, as it has been now an emerging field of interest in drug design and developments.

T. Rajkishan · A. Rachana · S. Shruti · P. Bhumi · D. Patel (✉)
Department of Biological Sciences and Biotechnology, School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India
e-mail: dhaval.patel@iar.ac.in

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_9

This chapter aims to illustrate some crucial CADD techniques also commonly referred to as in silico methods in a diverse arena of drug discovery and focusing some of the modern advancements.

Keywords

CADD (Computational aided drug design) · SBDD · LBDD · Docking · Virtual screening · Pharmacophore · QSAR · ADMET

9.1 Introduction

9.1.1 Drug and Drug Designing

The English word “Drug” originated from the French word “Drogue” that means Dry Herb, strongly recommends that most primitive drugs were extracted out from various plant sources (Rekker 1992; Wadud et al. 2007). A drug or medicine shows a physiological effect when introduced to the human body. The drug may be a natural or synthetic substance that affects the structure and functioning of a living body and is used for the prevention, treatment, and diagnosis of a specific disease and results in the relief of discomfort. In the field of pharmacology, a drug molecule is a chemical entity other than an essential dietary ingredient that develops a biological effect on a living system after administration. A medicine or pharmaceutical drug is used to cure or prevent disease. So far, traditional drugs are acquired through medicinal plants, however, also obtained by organic synthesis. Moreover, various pharmaceutical drugs are categorized into different drug classes. A group of related medicine or drug that possess similar chemical structures shows the same binding to the common biological target (reveal the same mechanism of action), correlated mode of action and that drugs are used to cure the same disease (Izzo and Ernst 2001). The Anatomical Therapeutic Chemical Classification System (ATC) is the most commonly used classification system in which a specific ATC code is designated to the drugs. One more broad classification system is the Biopharmaceutics Classification System; it categorizes the drugs based on their absorption or permeability properties and aqueous solubility (Lennernäs and Abrahamsson 2005). Broadly, any substance administered orally, or injected subcutaneously, intramuscularly, or intravenously, or applied topically or to a body cavity to treat or prevent a disease or condition is termed as “Drug.”

Drug design is a brilliant and magnificent inventive process in the development of novel therapeutics according to the biological target. Generally, it is also termed as rational drug design. Drug designing is a powerful invention in medicinal chemistry or biological history to produce an important and noteworthy beneficial or therapeutic reaction. A drug is an organic substance, once it binds to the particular target site it may either stimulate or inhibit the function of a biological molecule or macromolecule that outcomes as therapeutic benefits. The success of drug design is influenced by the accurate information of the 3D structure of biological targets (Phoebe Chen and Chen 2008).

9.1.2 Computer-Aided Drug Discovery

The discovery of a novel drug or identifying novel drug/drug-like entity is a costly, multifaceted, extremely risky, and time-consuming process which encompasses a wide range of modern tools/techniques and various scientific disciplines. It is estimated to take approximately 1.0 billion USD (Myers and Baker 2001) and 10–15 years (Moses et al. 2005) to complete a traditional drug discovery and development phase, from concept to approval of a novel drug into the market. This cost and time majorly contribute towards the lead synthesis and the testing of the lead compounds/analogues (Basak 2012). Research, development, and innovations achieved in medicinal chemistry resulted in an increase in compound databases spanning large and diverse chemical spaces aided in the development of high-throughput screening methods (HTS) as well drug discovery (Lavecchia and Giovanni 2013; Jhoti et al. 2013). To circumvent the challenges faced by traditional drug discovery approaches, pharmaceutical companies, academia, and other research organization employed computer-aided drug discovery (CADD) techniques. CADD has now become an essential tool for minimizing failures right from the preliminary screening to final phase of drug discovery and development. The overall workflow of CADD is summarized in Fig. 9.1.

The present approaches of the traditional drug discovery use several inter-disciplines such as protein biochemistry/biophysics, structural biology, computational and medicinal chemistry, synthetic chemistry, and pharmacy. Briefly, the different stages involved can be summarized as follows:

(a) **Target identification:** This phase includes identification of drug targets which are further inspected and correlated for their functions, biology associated with a specific disease/disorder (Anderson 2003).
(b) **Target validation:** It is the phase where the association and correlation of identified drug target with the disease are confirmed by biological assays which involve assessing the capacity to regulate biological functions (Chen and Chen 2008).
(c) **Lead identification:** It brings about the identification or discovery of a synthetic chemical moiety with a higher and significant degree of specificity, sensitivity, and potency against an above-identified drug target. The lead identified in this stage can be a drug-like candidate or precursor moiety to a drug-like entity (Kalyaanamoorthy and Chen 2011).
(d) **Lead optimization:** This is an important stage as it includes the improvement of the identified lead molecule through iterative cycles of lead compound (s) evaluation. Combined *in vitro* and *in vivo* experiments/validations are carried out to sort and filter the potential candidates to develop a safe, specific, and efficient drug. Further structure–activity relationships (SARs) strategies are

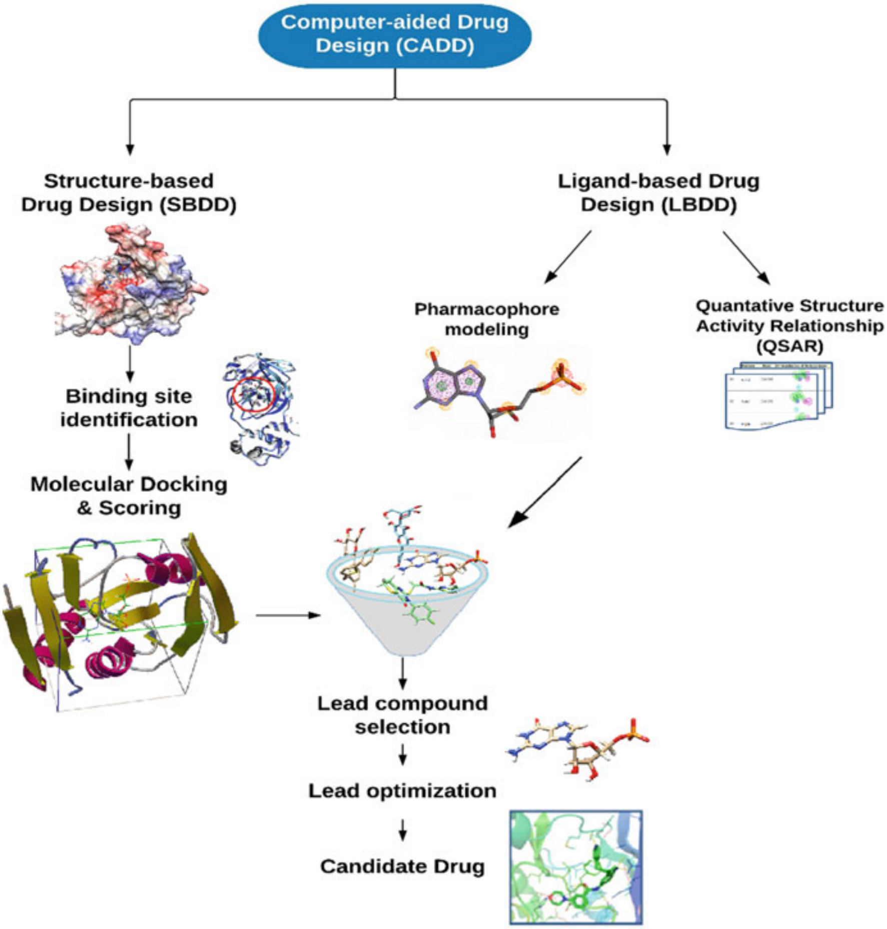

Computer-aided Drug Design (CADD)
Structure-based Drug Design (SBDD)
Ligand-based Drug Design (LBDD)
Pharmacophore modeling
Quantative Structure Activity Relationship (QSAR)
Binding site identification
Molecular Docking & Scoring
Lead compound selection
Lead optimization
Candidate Drug

Fig. 9.1 Overview of typical CADD workflow

developed to find out the pharmacodynamics and pharmacokinetic characteristics that can be extrapolated to lead compound(s) analogues (Andricopulo et al. 2009).

(e) **Preclinical stage:** This phase includes synthesis and formulation of a drug(s) on laboratory scale which is followed by in vivo animal studies for drug's toxicity and potency (Cavagnaro 2002; Silverman and Holladay 2015).

(f) **Clinical trials:** This is the final stage of drug discovery and development cycle which includes three clinical trials/phases that scrutinize the safety, evaluating any potential side effects, predetermination of dosage and its efficacy, and overall pharmacological profile of the candidate drug on human volunteers (Silverman and Holladay 2015).

The high budget of drug development is often ascribed to the fact that approximately 90% of the lead molecules entering clinical trials fail to get regulatory approvals to reach the consumer end-market. An estimated 75% of the total cost of drug development occurs due to the failures met during the drug discovery and development pipeline (Leelananda and Lindert 2016). Nowadays with rapid developments in high-throughput screening (HTS) experiments and combinatorial chemistry, which can screen thousands of probes with robotics have accelerated the drug discovery expedition. However, HTS is still expensive in terms of cost and experimental resources, which are not easily available in academic settings. Eventually, various computational methodologies have been developed to reduce the series of the drug development cycle and drastically diminish the expenses and threat of failure in drug discovery and development.

This drug development procedure can be separated into two distinct phases. The preliminary drug discovery and development phases comprise the drug target selection and simultaneously the potential hit-to-lead compounds identification by employing *in silico* screening and/or high-throughput (virtual) screening approaches. This is followed by lead optimization procedure to search for a best clinical candidate with increased affinity and optimized pharmacokinetics properties. The development of a new drug also includes the second phase, which is dedicated to determining the clinical utilization of candidate leads (Bleicher et al. 2003).

In the current post-genomic era, the CADD tools have been utilized as a “virtual shortcut” to analyze, develop, and discover/identify potent drug-like molecules. CADD is an interdisciplinary area where numerous aspects of basic and applied research combine to cover most phases of the drug discovery practice starting from drug target identification and its validation to the optimization of the identified lead molecule. During the initial phase of the drug development procedure, investigators/researchers may face problems without any prior information of structure–activity relationship (Kore et al. 2012). It also includes evaluation and optimization of the pharmacokinetic parameters of drug-like molecules such as ADMET (absorption, distribution, metabolism, excretion, and toxicity) for the safety issues. Noteworthy, the utilization of these computational tools and techniques in the arena of drug discovery and development is promptly attaining popularity and is implemented frequently. Numerous terms are being adapted to this in silico domain, such as computational drug designing (CAD), computer-aided molecular modeling (CAMM), computer-aided drug design (CADD), computer-aided molecular design (CAMD), computer-aided rational drug discovery, rational drug design, and *in silico* drug design. The use of computational tools along with the experimental findings has a significant role in the novel drug development process and also has trimmed down the total cost of drug discovery and development without significant trade-off to overall accuracy.

Computer-aided drug design and development necessitate (Kapetanovic 2008):

- The practice of computing environment to rationalize/streamline the entire procedure of drug-like identification and optimization.

- The advantage of entire biological, pharmacological, and chemical knowledge about targets/ligands to recognize and optimize novel compounds (drug).
- Design of *in silico* screening to remove unwanted molecules with poor pharmacokinetic constraints such as ADMET (absorption, distribution, metabolism, excretion, and toxicity)/poor activity and choose the furthermost excellent, potent, and specific candidate molecules.

In the light of the current scenario with an empirical increase in biological macromolecule and small molecule data along with technological advancements in computational power, further advancements are plausible for forthcoming drug innovation tools. The most commonly used CADD approaches are structure-based drug design (SBDD), ligand-based drug design (LBDD), and sequence-based approaches. All of the above methods perform better when combinational and hierarchical strategies are applied employing multiple computational approaches rather than a single approach (Macalino et al. 2015).

The chapter includes a concise summary of the current strategy of computational drug designing and developments including methodologies, i.e. structure-based and ligand-based and application of *in silico* approach. It is also intended to deliver a brief overview of the *in silico* applications in the drug design procedure from the initial stage of target selection to the final stage of lead compounds development and evaluation.

9.2 Approaches to Drug Designing

Generally, approaches utilized in CADD are classified into two categories: structure-based and ligand-based methods. The structure-based approach relies on the availability of the 3D structure of the target protein for the screening and identification of promising ligand molecules by calculating the interaction energies between the target and compound. In contrast, the latter approach utilizes the information/knowledge of active and inactive molecules with diverse chemical structures as well as the development of predictive models such as QSAR (quantitative structure–activity relation) (Kalyaanamoorthy and Chen 2011). These models are further utilized for screening and identification of additional newer chemical entities through a large chemical database search, a process called virtual screening. To summarize, the structure-based method is preferred when high-resolution 3D structural data of drug target is accessible and the ligand-based approach is ideal when no significant structural information through experimental approaches are available.

9.2.1 Structure-Based Drug Design (SBDD)

Structure-based drug design (SBDD) approach depends on the availability and knowledge of 3D structural information about the target protein. The knowledge of the binding/active site in the target protein structure is utilized to identify, design, and evaluate ligands based on their interactions with the residues present in the binding/active site of target protein (Grinter and Zou 2014; Lavecchia and Giovanni 2013). Thus the acquisition of structural information is a core part of the hypothesis of SBDD approach which exploits the molecule's potential to have energetically favorable interaction with the target protein's binding site and its consequent biological outcome. Thus, 3D structural information of the drug target protein is quintessential in SBDD which is now routinely available through state-of-the-art technologies like X-ray diffraction (XRD), in solution nuclear magnetic resonance (NMR), Cryo-electron microscopy (Cryo-EM) along with computational modeling techniques such as homology modeling and all-atom molecular dynamics (MD) simulations (Kalyaanamoorthy and Chen 2011; Lin et al. 2020; Patel et al. 2019). Various packages used in binding site prediction, modeling packages, and homology modeling tools are summarized in Table 9.1. Further, SBDD can be divided into molecular docking and de novo ligand design approaches. Molecular docking approach used the physiochemical properties of binding sites on proteins such as hydrogen bond, hydrophobic and electrostatic fields, key residues and then a compound is exploited to identify whether it is a suitable molecule whose molecular shapes match the binding sites of the receptor with high binding affinity. If this approach is used for screening large databases or compound libraries, it is also referred to as high-throughput virtual screening (VS) approach (Lavecchia and Giovanni 2013; Andricopulo et al. 2009). Another approach known as de novo ligand design utilizes information from protein binding site to identify small fragments that align well with the molecular shape of the binding site followed by linking these small fragments based on chemistry connection rules identify a structurally novel ligand moiety (Kutchukian and Shakhnovich 2010; Reker et al. 2014). The successful screening of potential lead molecule from any of the above approaches is further synthesized followed by evaluation of their biological activities.

Structure-based CADD (SB-CADD) approach depends on the capability to conclude and examine 3D structures of biological entities. The underlying fundamental assumption of this method is that compounds affinity and ability to bind with an exact receptor/target and maintain preferred biological interactions is based on its competence to favorably bind to a specific interacting pocket on the same target. Compounds that contribute to those complementary interactions would reveal similar biological effects. Consequently, novel molecules can be identified through the accurate and attentive investigation of the binding pocket of a specific protein. The SB-CADD project relies on prerequisite structural information about the target of interest. The SB-CADD has emerged as a frequently used approach in the area of drug discovery, thanks to improvements in -omics era that have enabled researchers with finding a huge repertoire of candidate/potential drug targets (Bambini and Rappuoli 2009).

So far, many scientists and research group all over the world have shown enthusiastic efforts in the direction of virtual high-throughput ligand screening (VLS) and structure-based drug discovery (SBDD), as the furthermost scientifically challenging and promising methodologies to identify the best lead for

Table 9.1 Tools and software packages used in various computational drug design

Function	Program/ Server	Free/ Commercial	Description	Websites
Binding/ interaction sites prediction	CASTp	Free	Utilizes the alpha complex and weighted Delaunay triangulation for shape measurements	sts.bioe.uic.edu/castp/
	Cavitator	Free	Grid-based geometric analysis for pocket prediction	sites.gatech.edu/cssb/cavitator/
	ConCavity	Free	Uses evolutionary sequence and 3D structures	compbio.cs.princeton.edu/concavity/
	eFindSite	Free	Uses a set of evolutionarily related proteins for predicting common ligand-binding site	www.cct.lsu.edu/resources
	SiteComp	Free	Uses molecular interaction fields for binding site comparison	sitecomp.sanchezlab.org
	PocketFinder	Free (PyMOL plugin)	Utilizes shape descriptors for pocket identification	www.modeling.leeds.ac.uk/pocketfinder/
	fpocket	Free as a standalone program	Uses alpha sphere theory	fpocket.sourceforge.net/
	ProBis	Free	Local structural alignments	probis.cmm.ki.si/index.php
	3DLigandSite	Free	Based on homologous structure for ligand-binding site prediction	www.sbg.bio.ic.ac.uk/B3dligandsite/
	ConSurf	Free	Surface-mapping	consurf.tau.ac.il/2016/
Docking	AutoDock/ AutoDock Vina	Free	Flexible side chains (genetic algorithm)	autodock.scripps.edu/
	Adaptive BP-Dock	NA	Integrates perturbation response scanning (PRS) with the flexible docking protocol	(Bolia et al. 2014)
	cDocker	Commercial	Uses side-chain flexibility at the atomic level with grid-based docking	accelrys.com/
	Docking server	Free/commercial	Integration of several computational chemistry tools	www.dockingserver.com/web

(continued)

Table 9.1 (continued)

Function	Program/ Server	Free/ Commercial	Description	Websites
	FLIPDock	Free for academic usage	Flexible LIgand-Protein Docking with receptor conformational change	flipdock.scripps.edu/
	GOLD	Commercial	Protein flexibility	www.ccdc.cam.ac.uk/Solutions/GoldSuite/Pages/GOLD.aspx
	Glide	Commercial	Flexible protein and ligand docking	www.schrodinger.com/Glide
	idock	Free	Flexible ligand docking	www.schrodinger.com/Glide
	SwissDock	Free	Grid-based rigid online docking server	www.swissdock.ch/
	VLifeDock	Commercial	A stochastic method based on GRID, genetic algorithm, and GRIP	www.vlifesciences.com/products/Functional_products/VLifeDock.php
	PatchDock	Free	Uses shape complementarity method of rigid docking	bioinfo3d.cs.tau.ac.il/PatchDock/
	GEMDOCK	Free	Docking using generic evolutionary method	gemdock.life.nctu.edu.tw/dock/
	PLANTS	Free for academic	Uses ant colony optimization (ACO) method	http://www.tcd.uni-konstanz.de/plants_download/
	DOCK	Free for academic	Based on Delphi electrostatics, ligand conformational entropy corrections, and desolvation of ligand and receptor	dock.compbio.ucsf.edu/
	FRED	Free for academic	Based on systematic and non-stochastic evaluations of all possible protein–ligand poses, shape complementarity, and chemical feature alignment	www.eyesopen.com/oedocking

(continued)

Table 9.1 (continued)

Function	Program/ Server	Free/ Commercial	Description	Websites
	HADDOCK	Free for academic	Information driven flexible protein–protein docking	https://wenmr.science.uu.nl/haddock2.4/
	ICM	Commercial	Flexible protein–ligand/peptide/protein docking	www.molsoft.com/docking.html
	FlexX	Commercial	Uses an incremental construction approach for flexible ligand docking	www.biosolveit.de/FlexX/
	LigandFit	Commercial		accelrys.com
	LibDock	Commercial		accelrys.com
Modeling and molecular dynamics packages	Amber	Free	Molecular Dynamics Simulation package	ambermd.org/
	BioSolveIT	Free/commercial	Molecular modeling packages	www.biosolveit.de/
	Desmond	Free/commercial	Molecular Dynamics Simulation package	www.deshawresearch.com/resources_desmond.html
	Discovery studio	Commercial	Molecular mechanics simulation program based on CHARMM force fields	accelrys.com/products/discovery-studio/simulations.html
	GROMACS	Free	Molecular Dynamics Simulation package	www.gromacs.org
	Molecular operating environment (MOE)	Commercial	Molecular modeling and simulation package	www.chemcomp.com/Products.htm
	NAMD	Free	Molecular Dynamics Simulation package	www.ks.uiuc.edu/Research/namd/
	SYBYL-X	Commercial	Molecular modeling and simulation package	https://www.certara.com/
	Yasara dynamics	Commercial	Molecular modeling and simulation package	www.yasara.org/md.htm
Databases	DrugBank	Free	A comprehensive database of target and drug	www.drugbank.ca/
	GLIDA	Free	GPCR ligand database	pharminfo.pharm.kyoto-u.ac.jp/services/glida/

(continued)

Table 9.1 (continued)

Function	Program/ Server	Free/ Commercial	Description	Websites
	PubChem	Free	Database of small compounds	pubchem.ncbi.nlm.nih.gov/
	ZINC	Free	Collection of commercially available compounds for virtual screening	zinc.docking.org/
	ChemSpider	Free/commercial	Database of chemical structures	www.chemspider.com/
	ChEMBL	Free/commercial	Manually curated database of bioactive compounds	www.ebi.ac.uk/chembl/
	Cambridge Structural Database (CSD), CCDC	Free/commercial	Database of small molecules from structures solved using X-ray and NMR	www.ccdc.cam.ac.uk/
	BindingDB	Free/commercial	Binding affinities of drug targets	www.bindingdb.org/bind/index.jsp
	NCI	Free/commercial	Large database of curated compounds	cactus.nci.nih.gov/download/nci/index.html
	HMDB	Free/commercial	Metabolites in human body	www.hmdb.ca/
Homology modeling	RaptorX	Free for non-commercial use	Distance-based protein folding prediction using deep learning	raptorx.uchicago.edu/
	Biskit	Free	Collection of python-based libraries for structural bioinformatics	biskit.pasteur.fr/
	Phyre2	Free	HMM-based web server for structure prediction, analysis of functions and mutations	http://www.sbg.bio.ic.ac.uk/~phyre2/
	EsyPred3D	Free	MODELLER based automated web server	www.unamur.be/sciences/biologie/urbm/bioinfo/esypred/
	Modeller	Free	Homology/comparative modeling tool based on spatial restraints and de novo loop modeling	salilab.org/modeller/

(continued)

Table 9.1 (continued)

Function	Program/ Server	Free/ Commercial	Description	Websites
	Robetta	Free	Structure prediction using combined approaches such as deep learning, comparative modeling, and *ab initio* modeling	robetta.bakerlab.org/
	I-TASSER	Free	Iterative Threading ASSEmbly Refinement based on hierarchical approach for structure prediction and structure-based function annotation	zhanglab.ccmb.med. umich.edu/I-TASSER/
	Bhageerath-H	Free	Web-based hybrid protein structure prediction approaches	www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp

pharmaceutical objectives (Bleicher et al. 2003; Foloppe et al. 2006; Klebe 2006; Miteva 2008). This structure-based approach of ligand identification offers an insight into the molecular interaction of protein–ligand complexes as well, permitting medicinal chemists to formulate extremely precise and exact chemical alterations or alterations around the skeleton/scaffold of ligand (Kitchen 2017; Kitchen et al. 2004). Overview of types of computational tools used in drug discovery, precision medicine, and chemical biology is shown in Fig. 9.2.

9.2.1.1 Structure-Based Virtual Screening

The structure-based virtual screening approach of a ligand is a computational method which comprises fast searching of huge libraries of chemical structures to identify and screen more potential drug-like hits (candidates) that are most probable to interact or bind to a specified target like protein, enzymes, and receptors. This is followed by docking of the screened hits into the active/binding site of the protein drug target and the calculated scoring function evaluates the probability whether the candidate will bind to the target with maximum affinity or not (Cheng et al. 2012). The most vital application of this approach is that it augments the hit rate frequency by remarkably reducing the number of molecules for experimental evaluation of their respective biological activity and in that way raises the realization rate of in vitro experiments. Various packages used for molecular docking and databases for virtual screening are summarized in Table 9.1. This approach has been pragmatic

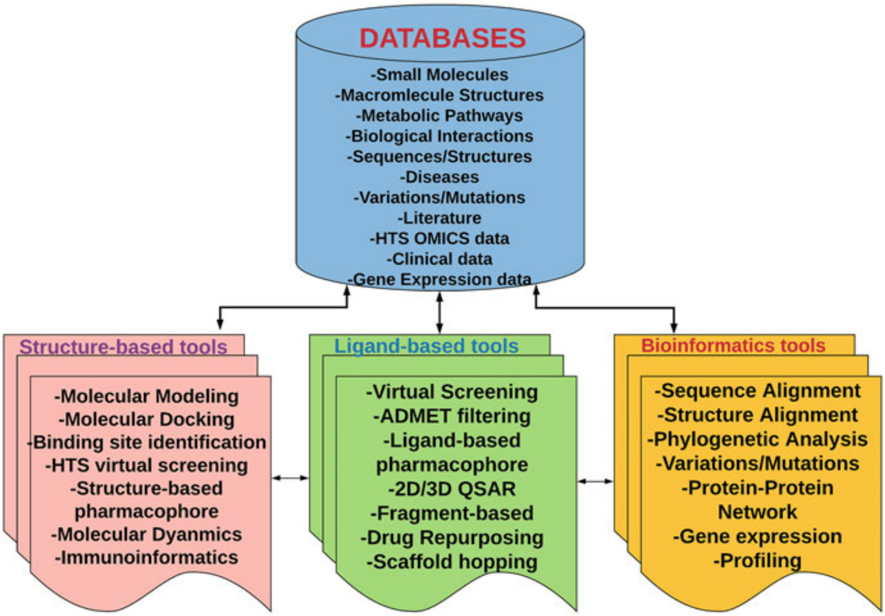


Fig. 9.2 Overview of bioinformatics resources in drug designing

consistently in biotechnology industries, pharmaceutics, and academic area for the early phase of the drug discovery process.

9.2.1.2 Structure-Based Lead Optimization (*In silico*)

After the initial hits obtained from VS, the preferred ligands are optimized using these techniques. Screening of best hit (lead) is based on its high affinity for its particular target/receptor/protein mostly based on free energy binding estimations. The assessment of its pharmacokinetics parameters, namely absorption, distribution, metabolism, excretion, and toxicity with its physiochemical description increases the probability of success in a later clinical phase trial/research. This lead optimization can be accomplished using different computational methods, which comprise quantitative–structure activity relationships, similarity search, databases, homology modeling, pharmacophores, etc. (Ekins et al. 2007).

9.2.2 Ligand-Based Drug Design

In case, where drug target structure is unavailable or the structure prediction using approaches such as ab initio structure prediction or homology modeling is challenging, the alternative to SBDD is ligand-based drug design (LBDD). Tools for modeling packages and homology modeling are summarized in Table 9.1. It depends on the information of compounds that interact with the biological protein target of curiosity. The three-dimensional quantitative structure–activity relationship

(3D-QSAR) and pharmacophore modeling along with 2D/3D molecular similarity assessment are the most crucial and frequently utilized outfits in the ligand-based drug design process. These both together can yield predictive models appropriate for lead identification and optimization (Acharya et al. 2010). Due to the lack of 3D experimental structure, the well-known ligand compounds that interact with drug target are taken into consideration by understanding the physicochemical and structural properties of the inhibitors/ligands that build the relation with preferred and anticipated pharmacological activity/properties of those compounds/ligands (Guner et al. 2004). Above and beyond the well-established ligand compounds, the LBDD approach may also involve natural derivatives/products or other substrate analogues that bind/interact with the specific target protein revealing the chosen biological outcome (Guner et al. 2004; Koehn and Carter 2005; Kuntz 1992; Lee 1999). Conversely, where the target protein information is not present, the biological knowledge of the collection of ligands activity against an appropriate receptor or enzyme (drug target) can be utilized to identify crucial structural and physicochemical relationship regarding properties and molecular descriptors accountable for the predicted biological activity. Now, here is a hypothesis that ligands with structural similarity might exhibit similar biological response and interactions with the specific drug target (Prathipati et al. 2007). Commonly applied approaches for LBDD are the QSAR (quantitative structure–activity relationships) and pharmacophore-based methodology (Tintori et al. 2010).

9.3 Introduction and Principal of the QSAR

The most standard and validated methods for LBDD are (2D/3D) QSAR and pharmacophore modeling. Overall, the QSAR is a computational technique to quantify the relationship among the specific biological process or activity and chemical structural properties for a series of molecules. The underlying postulates behind QSAR approach are that related physicochemical or structural properties result in similar biological activity (Akamatsu 2002; Verma and Hansch 2009). Primarily a set of a chemical compound or lead compounds are screened which indicates the appropriate and relevant biological activity. A quantitative-structural relationship is well settled between the physicochemical properties of the potential leads and the specialized biological activity. The generated QSAR model is further utilized for optimization of the best active molecules to enhance the significant biological activity. These predicted hits are then further directed to experimental test for the essential and desired activity. To conclude, the QSAR approaches thus can be utilized as a supervisory strategy/tool for recognition of novel hits along with their modified features and improved biological activity.

General components for developing QSAR models are as follows:

1. Primarily, categorization, screening, and identification of compounds with measured experimental values of their respective biological activity. Preferably

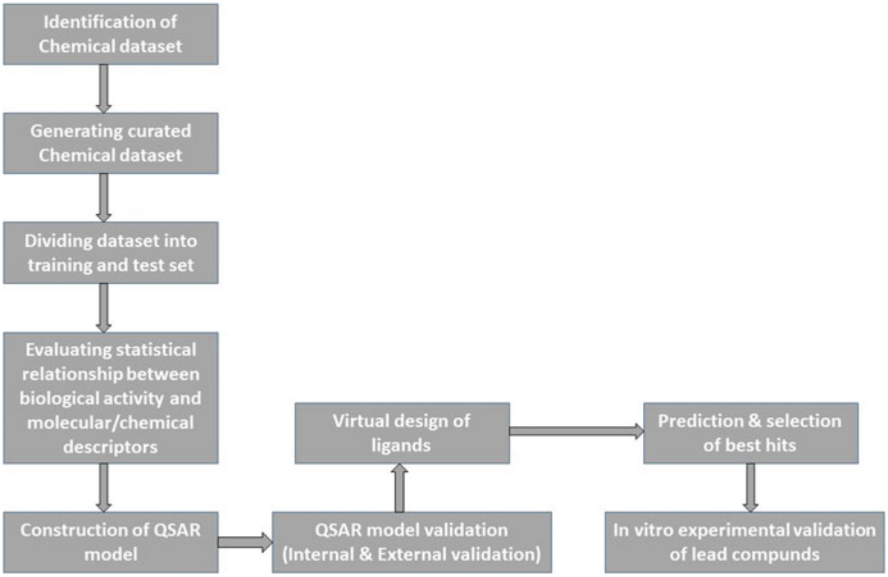

Identification of Chemical dataset
Generating curated Chemical dataset
Dividing dataset into training and test set
Evaluating statistical relationship between biological activity and molecular/chemical descriptors
Construction of QSAR model
QSAR model validation (Internal & External validation)
Virtual design of ligands
Prediction & selection of best hits
In vitro experimental validation of lead compunds

Fig. 9.3 Methodology for developing QSAR model

all the compounds are of a congeneric chain but should be sufficiently chemical diverse to have a great dissimilarity in biological activity.

2. Secondly, the identification and determination of molecular descriptors associated with diverse physicochemical and structural properties of the compounds and examination.
3. Identifying the correlation between 2D or 3D molecular descriptors and with their corresponding biological activity that can elucidate the difference in activities in the biological dataset.
4. Investigation of the mathematical reliability and analytical power of the derived QSAR model (Fig. 9.3).

9.3.1 Historical Progress and Development of QSAR

Over the past two decades, the logical or intellectual focal point (the core of gravity) of the arena of medicinal chemistry has reallocated significantly and dramatically from how to build the best compound, to what compound needs to build or make. The challenge at present is the collection of knowledge to make choices about the use of various assets in drug design. The information contributing to the drug design attempt and practice are progressively quantitative, constructing upon modern improvements in molecular structure explanation, statistics, combinatorial mathematics, and computer simulations. Generally, these fields have directed to an innovative paradigm in drug design and development that has been denoted as "*quantitative structure–activity relationships*." For above 40 years, the QSAR hypothesis first initiated its mode into the tradition of pharmaceutical chemistry (Tropsha and Golbraikh 2007).

In 1865, Crum-Brown and Fraser described the idea and knowledge that there was a statistical relationship among the chemical/molecular structure and their corresponding biological activity. First, one has to formulate the "physiological activity," i.e. ϕ, a function of the molecular/chemical structure C, which is signified below in Eq. (9.1); (Rekker 1992)

$$\phi = f(C) \tag{9.1}$$

Furthermore, Richet, Meyer, and Overton after few decades in 1893 and 1900 independently created a correlation between simple organic compounds and their water solubility, which additionally established a linear relationship denoting as oil–water partition or solubility, i.e. lipophilicity. In the 1930s L. Hammett stated that there is a correlation between electronic characteristics of organic bases and acids with their equilibrium reactivity and constants (Hansch et al. 1991).

In 1969, Corwin Herman Hansch published a model related to free energy to show a relationship between biological activities and physicochemical properties (Hansch 1978). Moreover, Taft formulated a mode for extracting polar resonance and steric effects and leading to the first steric parameters (Hansch et al. 1991). Taft and Hammett together contributed to formulating the mechanistic base for the generation of the QSAR prototype given by Fujita and Hansch (Lombardo et al. 2000). They mixed the hydrophobic constants along with Hammett's electronic constants to acquire the linear Hansch equation and numerous extended forms (Leo and Hansch 1999).

$$\mathrm{Log}\, 1/C = a\sigma + b\text{л} + ck \quad \text{Linear form} \tag{9.2}$$

$$\mathrm{Log}\, 1/C = a\,\mathrm{Log}P - b(\log P)_2 + c\sigma + k \quad \text{Non} - \text{linear form} \tag{9.3}$$

where C = Concentration vital to yield a standard response; Log P = Partition coefficient between water and 1-octanol; σ = Hammett substituent parameter; л = Relative hydrophobicity of substituents; a, b, c, k = Model coefficient

Moreover, along with the Hansch method, other approaches were also established to address the structure–activity problems. The Free-Wilson methodology states the structure–activity analysis in a congeneric sequence as shown in Eq. (9.4)

$$\mathrm{BA} = \sum aiXi + u \tag{9.4}$$

where BA = Biological activity; u = Average contribution to the parent compound; ai = Involvement of each structural features; Xi = 1 signifies presence; Xi = 0 signifies absence of a specific structural fragment

The outcomes of this method directed to the much complicated Fujita-ban equation which used the activity logarithm that draws the biological activity parameters in with other free energy-related relations (Myint and Xie 2010).

$$\mathrm{Log\,BA} = \sum GIXi + u \tag{9.5}$$

where u = Computed biological activity value of the unsubstituted parent molecule of a specific series; Gi = Biological activity involvement of the substituent; Xi = Value of one when the substituent is present or absent

Klopman and co-workers have prolonged differences in this activities based method (Benigni 1991). The topological approach has also been utilized to state the correlation between biological activity and chemical structure. Simon's Minimum Topological Differences (MTD) methodology and the elaborated analysis on molecular connectivity by Hall and Kier have done a significant contribution in the generation of QSAR relationships (Hall et al. 1991).

Currently, other improvements in QSAR comprise methods such as hologram QSAR (QASR), binary QSAR, and inverse QSAR (Prathipati et al. 2007).

9.3.2 Statistical Tools Applied for QSAR Model Development and Validation

The accomplishment of any developed model of QSAR significantly and importantly depends on the selection of the molecular descriptors and the capability to develop a significant statistical correlation between the relevant biological activity (specific) and molecular descriptors. Ever since the starting time of the QSAR study, it is confirmed that the depiction of molecular descriptors is the central and essential part of the methodology (Akamatsu 2002). Advanced software nowadays permits the creation of a huge number of molecular descriptors which can be refined for the development of QSAR model. The three main statistical methodologies conventionally executed in the linear model of the QSAR approach to elect molecular features crucial for activity are:

1. Principal component analysis (PCA)
2. Partial least square (PLS)
3. Multivariable linear regression analysis (MLR)

The principal component analysis (PCA) was accomplished to resolve the issues of MLR study by taking out statistics/knowledge from the multiple, probably redundant variables into a less significant number of non-associated variables. Hence, PCA offers a competent and effective mode designed for diminution of the number of independent variables applied in QSAR model generation. This approach has shown great help for schemes with several molecular descriptors comparable to the number of predictions. Though, outcomes achieved from PCA are frequently challenging and complex to evaluate the identification of special physicochemical and structural features significant for biological activity.

The multivariable linear analysis (MLR) is the straightforward and simple technique to enumerate the various molecular descriptors containing better correlation with the deviation in biological activity. QSAR model generation by MLR approach can consist of forward or backward stepwise regression based on a statistical test to retrieve the robust model (i.e. analytically adding or removing molecular descriptors, to establish the ideal model). Nonetheless, for high numbers of molecular descriptors, the MLR procedure may be inefficient and the researcher needs to be attentive to eliminate the variable associated with high internal correlation. Even though this issue can be resolved by applying statistical tools/software (Langer and Hoffmann 2005; Scior et al. 2009) wherever the user can mechanize the MLR approach with suitable conditions.

Moreover, the PCA and MLR methods are combined to form the partial least square (PLS), where the biological activity (dependent variable) is also exported into a fresh new component to optimize the corresponding relationship (Geladi and Kowalski 1986). Where more than one dependent variable is present, the PLS method is useful and advantageous. Many additional variable selection techniques are available like the Bayesian method and genetic algorithms applied in linear QSAR models (Bajorath 2001; Zheng and Tropsha 2000). Once a primary QSAR model has been computed, it must be followed with exhaustive validation. Derived QSAR model is validated through two ways of validation methods:

1. Internal validation
2. External validation

The most familiar and adopted inner validation process is the leave-one-out cross-validation method (Kohavi and Kohavi 1995). Briefly, in this approach, one set of predictions is tagged as validation data with the remaining of the data tagged as the training set to approximate the coefficients of the developed QSAR model. The relevant biological activity of the test set is further estimated using the derived model consisting of the training set compounds. Further, the method is reiterated for additional rest of the remaining compounds unless each one of them is considered once as a test set of compounds. Afterwards, the analytical ability and strength of the derived model are evaluated by analyzing and evaluating the cross-validated r2 or q2 computed from the following mathematical equation:

$$Q^2 = 1 - \frac{\sum \left(y_{\text{pred}} - y_{\text{obs}}\right)^2}{\sum \left(y_{\text{obs}} - y_{\text{mean}}\right)^2}$$

The major drawback of this procedure is the amount of time required to accomplish the calculation rises with double the size of the training set compounds. Conversely, another deviation of this procedure is k-fold cross-validation. As an alternative to leaving one molecule out or away, this approach generated the training set by parting a subset of several compounds out of it at the moment. Nevertheless, the k-fold way is objected to the selection of the value *k*. Moreover, these two cross-validation approaches fail to utilize entire existing data at the same for validation of the derived model as well. Alternatively, the external validation technique includes anticipating the test set compound's biological activity that is not employed for a model generation (Cherkasov et al. 2014). This approach is very accurate and rigorous as it can be easily compiled and used for an entire available.

9.3.3 Molecular Descriptors applied in QSAR

In any of the QSAR study, molecular descriptors are exemplified as a mathematical demonstration of chemical information coded with a molecular structure by using a mathematical approach as shown in Table 9.2 (Karelson 2000). Various programs used for calculations of molecular descriptors and QSAR model analysis are summarized in Table 9.3.

Table 9.2 Molecular descriptors applied in QSAR

Type	Descriptors
Electronic parameters	Hammett constant (σ, σ +, σ -)
	Taft's inductive (polar) constant (σ*)
	Swain and Lupton field parameter
	Ionization constant (pKa and ΔpKa)
	Chemical shifts (IR and NMR)
Steric parameters	Taft's steric parameter (Es)
	Molar volume (MV)
	Van der Waals radius
	Van der Waals volume
	Molar refractivity (MR)
	Paracord
	Sterimol
Hydrophobic parameters	Partition coefficient (log P)
	Hansch's substitution constant (π)
	Hydrophobic fragmental constant (f, f')
	Distribution coefficient (log D)
	Apparent log P
	Capacity factor in HPLC (log k', log key)
	Solubility parameter (log S)
Quantum chemical descriptors	Atomic net charge (Qσ, Qπ)
	Superdelocalizability
	EHOMO—The energy of highest occupied molecular orbital
	ELUMO—The energy of lowest unoccupied molecular orbital
Spatial descriptor J	Jurs descriptors, shadow indices, radius of gyration
	Principle moment of inertia

Table 9.3 Tools/programs/software for calculating the molecular descriptor or QSAR model analysis

Function	Program/ Server	Free/ Commercial	Description	Websites
QSAR	McQSAR	Free	Generates multi-conformational QSAR using genetic function approximation paradigm	users.abo.fi/mivainio/mcqsar/index.php
	SYBYL-X	Commercial	Collection of QSAR packages (CoMFA, HQSAR, and Topomer CoMFA)	www.certara.com/pressrelease/certara-enhances-sybyl-x-drug-design-and-discovery-software-suite/
	MOLFEAT	Free	Calculates molecular fingerprints and descriptors derived from curated QSAR models	jing.cz3.nus.edu.sg/cgi-bin/molfeat/molfeat.cgi
	Open3DQSAR	Free	High-throughput chemometric analysis of molecular interaction fields (MIFs)	open3dqsar.sourceforge.net/
	E-Dragon	Free	Calculates molecular descriptors to evaluate structure–activity or structure–property relationship studies for HTS screening	www.vcclab.org/lab/edragon/

The statistics of structure descriptors are based on two main factors:

1. The molecular representation of molecules.
2. The geometric algorithm that is utilized for the descriptor calculation.

Three key types of parameters primarily recommended are:

i. Electronic
ii. Steric
iii. Hydrophobic

9.3.4 Approaches of QSAR

Since the Hansch's seminal efforts, various diverse methodologies of the QSAR have been generated. QSAR techniques may be examined to view the fact:

1. The different kinds of structural factors that are employed to distinguish molecular identities consisting of numerous demonstrations of compounds, from simple chemical principles to 3D conformations.
2. The scientific approach which is used to achieve the statistical association between these biological activities and structural parameters (Gonzalez et al. 2008).

9.3.4.1 2D-QSAR Techniques

For correlating the associations involving chemical structure and experimental observations, 2D QSAR is a preferred choice of method. The critical elements of 2D QSAR are the numerical descriptors utilized for translation of a chemical structure into mathematical variables, the quality of experimental observations, and the choice of statistical methods engaged to derive the relationships among the observations and numerical descriptors. The appropriate descriptors of each molecular structure are crucial constraints for an efficient QSAR model (Hansch 1978). The valid descriptors range from the very basic (element number) to the extremely complex (electrostatic field maps). 2D descriptors generally rely on the connection table for the chemical structure and are deterministic. These 2D descriptors can be the amount of the acids in a molecule which are designated as unweighted, or weighted, like molecular weight (the mass of each element multiplied by the number of atoms of that element). Sometimes the QSAR models may also yield 2D descriptors, e.g. several different octanol/water partition (log P) predictors used in QSAR analysis developed from the analysis of calculated log P. Other QSAR model-based descriptors include the Abraham descriptors for hydrogen bonding strength (Abraham 1993), the pKa and ionization state calculations (Milletti et al. 2007), and the topological polar surface area (Ertl et al. 2000). Apart from this molecular fingerprints (binary or integer units) representing the occurrence and absence of a set of molecular substructures in a chemical molecule are also employed.

Hansch Analysis

Corwin Hansch in 1969 summarized the idea of linear free energy (LFER) relationships to illustrate the efficiency of a biologically active and potential compound. This is one of the best promising methods to the quantification and assessment of the drug compounds and interaction with the biological system. It also describes as an extra thermodynamic procedure which concludes an additive consequence of numerous substituents in steric, electronic, hydrophobic, and scattering data of macromolecules interaction. Also, it supported in the scattering data of the non-covalent interaction of macromolecules like protein/enzyme/receptor and a drug compound. This approach creates the relation of biological activity within a homologous chain of molecules to a set of hypothetical and theoretical chemical parameters which define significant properties of the drug compound. Hansch demonstrated a linear and non-linear dependence of biological activity on various parameters.

Hansch suggested that the action of drug molecule relies on two practices:

1. Starting from the point of entrance in the body to the site of action of a drug that comprises the path of a sequence of membranes and consequently, it is associated with partition coefficient log P (lipophilic) and can be elucidated by random walk theory.
2. Interaction with the target site that in turn based on Steric bulk substitution groups and the electron density on the attached group

Free-Wilson Analysis

The Free-Wilson method is a structure–activity centered approach since it integrates the contribution created by numerous structural fragments to the entire biological activity (Schaper 1999). Indicator variables are utilized to indicate the absence or presence of a specific structural feature.

9.3.4.2 3D-QSAR

The 3D-QSAR model includes the study of the quantifiable relationship involving the three-dimensional properties of a set of molecules and their respective biological activity by employing statistical correlation approaches. 3D-QSAR practices three-dimensional characteristics which are mostly steric and electrostatic properties of compounds that depend on probe-based sampling inside the molecular lattice and that can further establish a correlation between the 3d descriptors with their respective biological activity (Deora et al. 2013).

Advancement in medicinal chemistry and drug development depends on the user's skill to decipher the molecular interaction of drug/lead molecule compounds with their relevant biological targets/receptors. The conventional QSAR analysis defines biological activity about physicochemical properties of molecules at the definite site of the drug compounds. This 3D-QSAR analysis utilizes the application of different force field functions that need the three-dimensional structures of a particular selected set of small compounds, i.e. training set with their known reported biological activity. The selected training set requires to be aligned or superimposed by using either experimental data of crystallographic protein–ligand complex or most active compound alignment approach. It employs the Lennard-Jones potential, a computed potential that is concerned with the complete compound rather than a single substituent. Cramer et al. named the first 3D-QSAR as CoMFA, i.e. comparative molecular field analysis (Cramer et al. 1988). Recently other 3D QSAR strategies have also been in use, for example, spectral structure–activity relationship (S-SAR) (Putz and Lacrămă 2007), an adaptation of the fields for molecular comparison (AFMoC) (Gohlke and Klebe 2002), Topomer CoMFA (Cramer 2003), and comparative residue interaction analysis (CoRIA) (Datar et al. 2006).

Comparative Molecular Field Analysis (COMFA)

The CoMFA study reported in 1988 is a grid-based approach, most frequently used technique for the establishment of three-dimensional structure–activity relationships. This method relies on the hypothesis that the molecular interaction between drug and target is non-covalent. The alterations in biological activities of binding affinities of selected molecules are associated with the modifications within the electrostatic and steric fields of these molecules. Following that the field values are further linked with biological activity by PLS analysis (Chilton et al. 2017). Comparative molecular field analysis (CoMFA) is a typical and conventional 3D-QSAR technique that covers the complete procedure of drug discovery. However, CoMFA is significant and remarkable for high predictive power, the basic data-based characteristics still used by this approach are useless by noise. As far as, various endeavors have been practiced to meliorate the robustness of CoMFA model and calculate the predictive accuracy by involving numerous factors, comprising molecular alignment and confirmation along with grid spacing (Kubinyi et al. 1998).

Drawbacks and Limitations of CoMFA

Despite various benefits and its characteristics to overcome classical QSAR and its best performance in numerous practical routines and applications, CoMFA has shown quite a few pitfalls and limitations which are mentioned below (Lokendra et al. 2013).

- Improbability and uncertainty in the selection of dataset compounds of interest.
- Many amendable parameters like lattice placement, probe atom, step size, overall orientation, etc.
- Cut-off ranges applied.
- Several practical issues with PLS.
- Failures in potential energy functions.
- Not well-measured hydrophobicity.
- Truncated signal-to-noise ratio because of many inappropriate field variables.
- Useful only in vitro experimental data.

Application

Since the origin CoMFA, various studies/applications of the methodology in many disciplines have been published. Several successful activities of CoMFA method in the area of geochemistry (herbicide, pesticide, or insecticide), physio-chemistry (capacity factors, partition coefficients, and chemical shift), pharmacokinetics and toxicity kinetics, and thermodynamics have also been comprehensively appraised in many reviews (Bordás et al. 2003).

Comparative Molecular Similarity Indices (COMSIA)

CoMSIA suggests the Gaussian function for the distances based between the probe atoms and molecular atoms, to overcome a few of the intrinsic insufficiencies arising through Lennard-Jones and Coulomb potential useful forms. In this method, similarity fields such as electrostatic, hydrogen bond donor and acceptor, and steric are computed. These different fields were taken to cover the most important aid to ligand binding and have few applications over CoMFA method like the generation of most robust and reliable 3D-QSAR model, without any cut-offs and much spontaneously interpretable 3D-contour maps. It is useful to elucidate a structure–activity relationship to offer more useful knowledge for the development of novel and potent drug-like compounds/derivatives/analogues (Klebe and Abraham 1999).

9.4 The Concept of Pharmacophore Mapping

In drug designing, pharmacophore methods have emerged as one of the significant tools afterwards the past time's development. Since its introduction, various structure-based and ligand-based approaches have been developed for advanced and refined pharmacophore modeling. This evolved and enhanced pharmacophore methodologies have now been utilized successfully in the virtual screening process, de novo design, and additionally in lead optimization. Instead of this success, pharmacophore tools have not attained their predictable, full ability, especially in need of decreasing the recent expensive overall price related to drug design and development.

A pharmacophore is a set of essential features on a compound that interacts with protein and which are responsible for the biological phenomenon are collectively well-known as a pharmacophore. Ehrlich introduced the pharmacophore concept for first time in 1909 (Ehrlich 1909) by demonstrating the pharmacophore model as "molecular structure/framework that represents the significant features (snapshots) for the biological activity of the drug (pharmacon)". As per recent portrayal, a pharmacophore model is an "altogether collection of steric and electronic characters that are essential to assure the optimum supramolecular interactions with a particular biological drug target and to stimulate/inhibit its biological functions". Since 1998, the IUPAC (International Union of Pure and Applied Chemistry) officially addressed the pharmacophore as "the collection of electronic and steric characteristics that is essential to assure the optimum supramolecular communications/interactions with a particular biological target/receptor and to induce or inhibit its biological reaction" (Wermuth 2006).

For pharmacophore modeling two approaches are employed: first by ligand-based mode through alignment or superimposing a set of active compounds and identifying the common chemical characters/features that are necessary for respective biological activity and second by structure-based mode through probing probable interaction points among ligands and macromolecular target/receptors. Till date, various pharmacophore methods have been applied broadly in HTS virtual screening, lead optimization, and multi-target drug design. Some automated techniques were constantly appearing after the improvements in computational chemistry pharmacophore modeling and their applications. The pharmacophore method, nevertheless, still faces many problems which make it less capable to attain its maximum potential particularly concerning the high cost related with the identification of novel drug compound (Wermuth et al. 1998).

9.4.1 Pharmacophore-Model-Based Virtual Screening

Once when either structure-based or ligand-based methodology generates the pharmacophore model, it can be further utilized for searching the chemical database with 3D structures to identify probable ligands. This approach is known as the "pharmacophore-based virtual screening." Virtual screening of ligands based on pharmacophore and molecular docking signifies the mainstream tool for virtual screening in recent time. Conversely, docking-based and pharmacophore-based virtual screening decreases the complications arising from insufficient deliberation of protein elasticity or the application of inadequately designed or improved and optimized scoring functions by presenting acceptance radius for all pharmacophore characteristics.

9.4.1.1 Ligand-Based Pharmacophore Modeling

This is a significant computational approach used in the absence of a biological target structure. It is generally implemented by collecting common features of the chemical 3D structures of a set of well-known ligands/drugs illustrating the necessary molecular interactions among a specified molecular target and ligands. Usually, pharmacophore formation from various ligands of the training set molecules is comprised of two crucial steps: First, generating the conformational space for each ligand flexibility in the training set and to align the multiple ligands in the training set, and second, identification of the crucial common chemical features to generate the pharmacophore model (Poptodorov et al. 2006).

9.4.1.2 Structure-Based Pharmacophore Modeling

When the experimental 3D structure of a protein/target–ligand complex is available, the structure-based modeling approach is often employed for pharmacophore modeling. It relies on the corresponding chemical features of the active site, its microenvironment and their spatial relationship. The structure-based pharmacophore modeling approaches are further categorized into two independent subcategories:

1. Target/macromolecule based (without ligand/substrate)
2. Protein–ligand complex based

The major drawback of this approach is the availability of protein–ligand complex 3D structure. This method is not useful in the scenario where no molecules targeting the active/binding site of interest are known (Wolber and Langer 2005).

The outline of pharmacophore modeling is shown in Fig. 9.4 (Yang 2010).

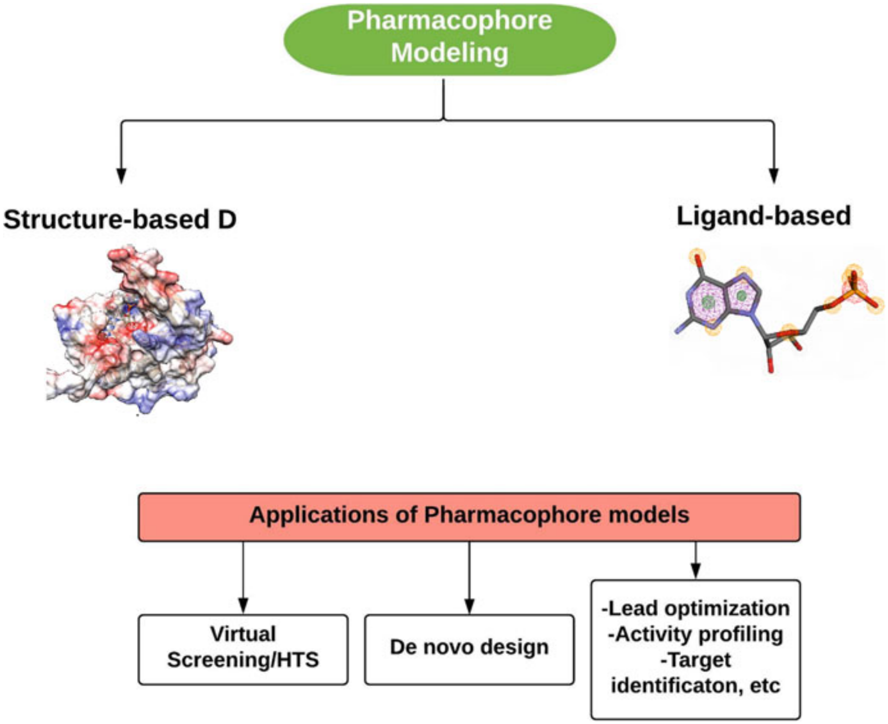


Fig. 9.4 The pharmacophore modeling workflow

9.5 ADME-Tox (Absorption, Distribution, Metabolism, Excretion-Toxicity)

ADMET includes absorption, distribution, metabolism, excretion, and toxicity of the identified lead/drug candidate. The studies of ADME-Tox are a very significant part of the early drug discovery procedure. If utilized in the early stage, it can reduce time, overall development cost, and decreases the likelihood of failure at the later stage. This analysis is conducted during lead optimization, discovery, and preclinical trial/development phases to deliver crucial information for characterization and classification of the compounds based on their properties to forecast their consequence after administration into the living system. This information associated with pharmacokinetics, metabolism, and toxicity is compiled to qualify and approve the safe and effective use of the drugs [https://www.admescope.com/about-us/flexible-adme-tox-services.html].

These are vital and crucial phenomenon taking place when chemical entities are administrated to transport and transform inside the living beings. ADMET analysis and modeling are complex in developing novel drugs and estimating the danger and side effects of a chemical entity like food additives, environmental pollutants, and

Table 9.4 Software/programs/tools used in prediction of ADMET properties

Function	Program/ server	Free/ commercial	Description	Websites
ADMET properties	QikProp	Commercial	ADME properties of drug candidates	www.schrodinger.com/
	ADMET predictor	Commercial	Estimates ADMET properties from query structure	www.simulations-plus.com/software/admetpredictor/
	ADMET and predictive toxicology	Commercial	Identifies ADMET properties	www.3ds.com/products-services
	FAF-Drugs2	Free	Uses in silico ADMET filters for candidate compound screening	www.mti.univ-paris-diderotfr/recherche/plateformes/logiciels
	SwissADME	Free as a web server	Evaluates overall pharmacokinetics and drug-likeness properties of small compounds	www.swissadme.ch/

pesticides that may communicate or enter the body of humans or other life forms. This is very commonly and frequently used in the area of *in silico* assessment of pharmacokinetic and metabolic-related points and molecular modeling (Balani et al. 2005; Tetko et al. 2006). Several commercial software and open-source are available at ease for ADMET analysis as mentioned in Table 9.4. ADMET profiling is governed by molecular structure knowledge, molecular descriptors, resulted from molecular graphs or another molecular presentation, between the non-candidate and candidate compounds. It incorporates into the early phase of drug design and development to speed up the practice of drug discovery and to decrease the number of molecules and removing them in the late phase of drug designing.

9.6 Applications of CADD Approach in Lead Discovery

The complete procedure of the drug discovery is extremely costly, lengthy, and time-consuming and a crucial problem for the biotech and pharmaceutical industry. Hence, the computational/in silico approaches have applied very smartly and widely in the drug-like lead development through compiling the knowledge of medicinal, chemical, and pharmacology biology to make it successful. At present, hundreds to millions of compounds have to be examined and tested within a very short period to find out novel hits, thus, exceedingly effective methodologies are essential for today's users/researchers. Therefore, *in silico* methods, involving computational tools and techniques are utilized in experimental works as it is cost-effective, favorable, and less complex to perform reliable virtual screening for lead identification. These various *in silico* methods involve several databases, similarity search, quantitative structure–activity relationships, pharmacophore modeling, data mining, data and network analysis tools, and other machine learning and molecular modeling tools. Such approaches have also been used in further optimization of the lead molecules with high affinity to a particular target, the elucidation of ADMET (absorption, distribution, metabolism, excretion, and toxicity) characteristics along with other physicochemical characterization.

The foremost application of the CADD or *in silico* approach provides an opulent array of prospects/chances that will shed light on the discovery of novel targets and at the end identification of lead to molecules with their predicted biological activity and physicochemical properties for defined novel targets.

9.7 Conclusion

More than the past decades, computer-supported/aided drug design and the property calculation of lead molecules have emerged as an extensively used and well-established field to support the research and development procedure in biotechnology/life science. Modern and advanced drug discovery and development cycle includes the identification of best hits, optimization of the drug-like lead compound to enhance their affinity, specificity to decrease the possible side effects along with effectiveness or potency, efficacy, metabolic constancy and stability to maximize the half-life and oral bioavailability. Once an initial lead is identified that satisfies all of these essential requirements, then further it would commence the drug discovery procedures before entering the clinical trials. These computational approaches of lead identification are favorable in the terms of cost, time and less tedious and eco-friendly than earlier lengthy, expensive, challenging and insufficient that resulted in low rate discovery of novel therapeutic.. Advanced *in silico* approaches, namely QSAR, pharmacophore modeling, and molecular docking and pharmacokinetics portray a significant role in the identification of novel drug-like compounds.

References

Abraham MH (1993) Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. Chem Soc Rev. https://doi.org/10.1039/CS9932200073

Acharya C, Coop A, E Polli J, D MacKerell A (2010) Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Curr Comput Aided-Drug Des 7:10–22. https://doi.org/10.2174/157340911793743547

Akamatsu M (2002) Current state and perspectives of 3D-QSAR. Curr Top Med Chem 2:1381–1394. https://doi.org/10.2174/1568026023392887

Anderson AC (2003) The process of structure-based drug design. Chem Biol. https://doi.org/10.1016/j.chembiol.2003.09.002

Andricopulo A, Salum L, Abraham D (2009) Structure-based drug design strategies in medicinal chemistry. Curr Top Med Chem 9:771–790. https://doi.org/10.2174/156802609789207127

Bajorath J (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. J Chem Inf Comput Sci 41:233–245. https://doi.org/10.1021/ci0001482

Balani S, Miwa G, Gan L-S, Wu J-T, Lee F (2005) Strategy of utilizing in vitro and in vivo ADME tools for lead optimization and drug candidate selection. Curr Top Med Chem 5:1033–1038. https://doi.org/10.2174/156802605774297038

Bambini S, Rappuoli R (2009) The use of genomics in microbial vaccine development. Drug Discov Today. https://doi.org/10.1016/j.drudis.2008.12.007

Basak SC (2012) Chemobioinformatics: the advancing frontier of computer-aided drug design in the post-genomic era. Curr Comput Aided-Drug Des 8:1–2. https://doi.org/10.2174/157340912799218507

Benigni R (1991) QSAR prediction of rodent carcinogenicity for a set of chemicals currently bioassayed by the US national toxicology program. Mutagenesis 6:423–425. https://doi.org/10.1093/mutage/6.5.423

Bleicher KH, Böhm HJ, Müller K, Alanine AI (2003) Hit and lead generation: beyond high-throughput screening. Nat Rev Drug Discov. https://doi.org/10.1038/nrd1086

Bolia A, Gerek ZN, Ozkan SB (2014) BP-dock: a flexible docking scheme for exploring protein-ligand interactions based on unbound structures. J Chem Inf Model 54:913–925. https://doi.org/10.1021/ci4004927

Bordás B, Kömíves T, Lopata A (2003) Ligand-based computer-aided pesticide design. A review of applications of the CoMFA and CoMSIA methodologies. Pest Manag Sci. https://doi.org/10.1002/ps.614

Cavagnaro JA (2002) Preclinical safety evaluation of biotechnology-derived pharmaceuticals. Nat Rev Drug Discov 1:469–475. https://doi.org/10.1038/nrd822

Chen YPP, Chen F (2008) Identifying targets for drug discovery using bioinformatics. Expert Opin Ther Targets. https://doi.org/10.1517/14728222.12.4.383

Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH (2012) Structure-based virtual screening for drug discovery: a problem-centric review. AAPS J. https://doi.org/10.1208/s12248-012-9322-0

Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'Min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem. https://doi.org/10.1021/jm4004285

Chilton SS, Falbel TG, Hromada S, Burton BM (2017) A conserved metal binding motif in the Bacillus subtilis competence protein ComFA enhances transformation. J Bacteriol 199. https://doi.org/10.1128/JB.00272-17

Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967. https://doi.org/10.1021/ja00226a005

Cramer RD (2003) Topomer CoMFA: a design methodology for rapid lead optimization. J Med Chem 46:374–388. https://doi.org/10.1021/jm020194o

Datar PA, Khedkar SA, Malde AK, Coutinho EC (2006) Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. J Comput Aided Mol Des 20:343–360. https://doi.org/10.1007/s10822-006-9051-5

Deora GS, Joshi P, Rathore V, Kumar KL, Ohlyan R, Kandale A (2013) Pharmacophore modeling and 3D QSAR analysis of isothiazolidinedione derivatives as PTP1B inhibitors. Med Chem Res 22:3478–3484. https://doi.org/10.1007/s00044-012-0349-7

Ehrlich P (1909) Über den jetzigen Stand der Chemotherapie. Ber Dtsch Chem Ges 42:17–47. https://doi.org/10.1002/cber.19090420105

Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: applications to targets and beyond. Br J Pharmacol. https://doi.org/10.1038/sj.bjp.0707306

Ertl P, Rohde B, Selzer P (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem 43:3714–3717. https://doi.org/10.1021/jm000942e

Foloppe N, Fisher LM, Howes R, Potter A, Robertson AGS, Surgenor AE (2006) Identification of chemically diverse Chk1 inhibitors by receptor-based virtual screening. Bioorg Med Chem 14:4792–4802. https://doi.org/10.1016/j.bmc.2006.03.021

Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 185:1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Gohlke H, Klebe G (2002) Drugscore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. J Med Chem 45:4153–4170. https://doi.org/10.1021/jm020808p

Gonzalez M, Teran C, Saiz-Urra L, Teijeira M (2008) Variable selection methods in QSAR: an overview. Curr Top Med Chem 8:1606–1627. https://doi.org/10.2174/156802608786786552

Grinter SZ, Zou X (2014) Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design. Molecules. https://doi.org/10.3390/molecules190710150

Guner O, Clement O, Kurogi Y (2004) Pharmacophore modeling and three dimensional database searching for drug design using catalyst: recent advances. Curr Med Chem 11:2991–3005. https://doi.org/10.2174/0929867043364036

Hall LH, Mohney B, Kier LB (1991) The electrotopological state: an atom index for QSAR. Quant Struct Relatsh 10:43–51. https://doi.org/10.1002/qsar.19910100108

Hansch C (1978) Recent advances in biochemical QSAR. In: Correlation analysis in chemistry. Springer US, pp 397–438. https://doi.org/10.1007/978-1-4615-8831-3_9

Hansch C, Leo A, Taft RW (1991) A survey of hammett substituent constants and resonance and field parameters. Chem Rev 91:165–195. https://doi.org/10.1021/cr00002a004

Izzo AA, Ernst E (2001) Interactions between herbal medicines and prescribed drugs: a systematic review. Drugs 61:2163–2175

Jhoti H, Rees S, Solari R (2013) High-throughput screening and structure-based approaches to hit discovery: is there a clear winner? Expert Opin Drug Discovery. https://doi.org/10.1517/17460441.2013.857654

Kapetanovic IM (2008) Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. Chem Biol Interact 171:165–176. https://doi.org/10.1016/j.cbi.2006.12.006

Kalyaanamoorthy S, Chen YPP (2011) Structure-based drug design to augment hit discovery. Drug Discov Today. https://doi.org/10.1016/j.drudis.2011.07.006

Karelson M (2000) Molecular descriptors in QSAR/QSPR, p 35168

Kitchen DB (2017) Computer-aided drug discovery research at a global contract research organization. J Comput Aided Mol Des 31:309–318. https://doi.org/10.1007/s10822-016-9991-3

Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. https://doi.org/10.1038/nrd1549

Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. Drug Discov Today. https://doi.org/10.1016/j.drudis.2006.05.012

Klebe G, Abraham U (1999) Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. J Comput Aided Mol Des 13:1–10. https://doi.org/10.1023/A:1008047919606

Koehn FE, Carter GT (2005) The evolving role of natural products in drug discovery. Nat Rev Drug Discov. https://doi.org/10.1038/nrd1657

Kohavi R, Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection, pp 1137–1143

Kore PP, Mutha MM, Antre RV, Oswal RJ, Kshirsagar SS (2012) Computer-aided drug design: an innovative tool for modeling. Open J Med Chem 02:139–148. https://doi.org/10.4236/ojmc.2012.24017

Kubinyi H, Hamprecht FA, Mietzner T (1998) Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. J Med Chem 41:2553–2564. https://doi.org/10.1021/jm970732a

Kuntz ID (1992) Structure-based strategies for drug design and discovery. Science (80-) 257:1078–1082. https://doi.org/10.1126/science.257.5073.1078

Kutchukian PS, Shakhnovich EI (2010) De novo design: balancing novelty and confined chemical space. Expert Opin Drug Discovery. https://doi.org/10.1517/17460441.2010.497534

Langer T, Hoffmann R (2005) Virtual screening an effective tool for lead structure discovery. Curr Pharm Des 7:509–527. https://doi.org/10.2174/1381612013397861

Lavecchia A, Giovanni C (2013) Virtual screening strategies in drug discovery: a critical review. Curr Med Chem 20:2839–2860. https://doi.org/10.2174/09298673113209990001

Lee KH (1999) Anticancer drug design based on plant-derived natural products. J Biomed Sci. https://doi.org/10.1007/BF02253565

Leelananda SP, Lindert S (2016) Computational methods in drug discovery. Beilstein J Org Chem. https://doi.org/10.3762/bjoc.12.267

Lennernäs H, Abrahamsson B (2005) The use of biopharmaceutic classification of drugs in drug discovery and development: current status and future extension. J Pharm Pharmacol 57:273–285. https://doi.org/10.1211/0022357055263

Leo AJ, Hansch C (1999) Role of hydrophobic effects in mechanistic QSAR. Perspect Drug Discov Des. https://doi.org/10.1023/A:1008762321231

Lin X, Li X, Lin X (2020) A review on applications of computational methods in drug screening and design. Molecules. https://doi.org/10.3390/molecules25061375

Lokendra OK, Rachana S, Mukta Rani B (2013) Modern drug design with advancement in QSAR: a review. Int J Res Biosci

Lombardo F, Shalaeva MY, Tupper KA, Gao F, Abraham MH (2000) ElogP (oct): a tool for lipophilicity determination in drug discovery. J Med Chem 43:2922–2928. https://doi.org/10.1021/jm0000822

Macalino SJY, Gosu V, Hong S, Choi S (2015) Role of computer-aided drug design in modern drug discovery. Arch Pharm Res. https://doi.org/10.1007/s12272-015-0640-5

Milletti F, Storchi L, Sforna G, Cruciani G (2007) New and original pKa prediction method using grid molecular interaction fields. J Chem Inf Model 47:2172–2181. https://doi.org/10.1021/ci700018y

Miteva M (2008) Hierarchical structure-based virtual screening for drug design. Biotechnol Biotechnol Equip 22:634–638. https://doi.org/10.1080/13102818.2008.10817525

Moses H, Dorsey ER, Matheson DHM, Thier SO (2005) Financial anatomy of biomedical research. J Am Med Assoc 294:1333–1342. https://doi.org/10.1001/jama.294.11.1333

Myers S, Baker A (2001) Drug discovery – an operating model for a new era. Despite the advent of new science and technologies, drug developers will need to make radical changes in their operations if they are to remain competitive and innovative. Nat Biotechnol. https://doi.org/10.1038/90765

Myint KZ, Xie XQ (2010) Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. Int J Mol Sci 11:3846–3866. https://doi.org/10.3390/ijms11103846

Patel B, Singh V, Patel D (2019) Structural bioinformatics. In: Essentials of bioinformatics, vol I. Springer International Publishing, Cham, pp 169–199. https://doi.org/10.1007/978-3-030-02634-9_9

Phoebe Chen Y-P, Chen F (2008) Identifying targets for drug discovery using bioinformatics. Expert Opin Ther Targets 12:383–389. https://doi.org/10.1517/14728222.12.4.383

Poptodorov K, Luu T, Hoffmann RD (2006) Pharmacophore model generation software tools. In: Pharmacophores and pharmacophore searches. Wiley, pp 15–47. https://doi.org/10.1002/3527609164.ch2

Prathipati P, Dixit A, Saxena A (2007) Computer-aided drug design: integration of structure-based and ligand-based approaches in drug design. Curr Comput Aided-Drug Des 3:133–148. https://doi.org/10.2174/157340907780809516

Putz M, Lacrămă A-M (2007) Introducing spectral structure activity relationship (S-SAR) analysis. Application to ecotoxicology. Int J Mol Sci 8:363–391. https://doi.org/10.3390/i8050363

Rekker RF (1992) The history of drug research: from Overton to Hansch. Quant Struct Relatsh 11:195–199. https://doi.org/10.1002/qsar.19920110214

Reker D, Rodrigues T, Schneider P, Schneider G (2014) Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. Proc Natl Acad Sci U S A 111:4067–4072. https://doi.org/10.1073/pnas.1320001111

Schaper KJ (1999) Free-Wilson-type analysis of non-additive substituent effects on THPB dopamine receptor affinity using artificial neural networks. Quant Struct Relatsh 18:354–360. https://doi.org/10.1002/(SICI)1521-3838(199910)18:4<354::AID-QSAR354>3.0.CO;2-2

Scior T, Medina-Franco J, Do Q-T, Martinez-Mayorga K, Yunes Rojas J, Bernard P (2009) How to recognize and workaround pitfalls in QSAR studies: a critical review. Curr Med Chem 16:4297–4313. https://doi.org/10.2174/092986709789578213

Silverman RB, Holladay MW (2015) The organic chemistry of drug design and drug action, 3rd edn. Elsevier Inc. https://doi.org/10.1016/C2009-0-64537-2

Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI (2006) Can we estimate the accuracy of ADME-Tox predictions? Drug Discov Today. https://doi.org/10.1016/j.drudis.2006.06.013

Tintori C, Manetti F, Botta M (2010) Pharmacophoric models and 3D QSAR studies of the adenosine receptor ligands. Curr Top Med Chem 10:1019–1035. https://doi.org/10.2174/156802610791293118

Tropsha A, Golbraikh A (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharm Des 13:3494–3504. https://doi.org/10.2174/138161207782794257

Verma RP, Hansch C (2009) Camptothecins: a SAR/QSAR study. Chem Rev 109:213–235. https://doi.org/10.1021/cr0780210

Wadud A, Prasad PVV, Rao MM, Narayana A (2007) Evolution of drug: a historical perspective. Bull Indian Inst Hist Med Hyderabad 37:69–80

Wermuth CG (2006) Pharmacophores: historical perspective and viewpoint from a medicinal chemist. In: Pharmacophores and pharmacophore searches. Wiley, pp 1–13. https://doi.org/10.1002/3527609164.ch1

Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA (1998) Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). Pure Appl Chem 70:1129–1143. https://doi.org/10.1351/pac199870051129

Wolber G, Langer T (2005) Ligand scout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 45:160–169. https://doi.org/10.1021/ci049885e

Yang SY (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug Discov Today. https://doi.org/10.1016/j.drudis.2010.03.013

Zheng W, Tropsha A (2000) Novel variable selection quantitative structure-property relationship approach based on the k-Nearest-Neighbor principle. J Chem Inf Comput Sci 40:185–194. https://doi.org/10.1021/ci980033m

Chemoinformatics and QSAR

10

Vivek Srivastava, Chandrabose Selvaraj, and Sanjeev Kumar Singh

Abstract

In recent years, constant increase in the performance of computer-based tools and several mathematical algorithms to solve chemistry-related problems. In recent years, screening of potent lead molecules using computational approaches has been gaining more attention as alternate approaches for high-throughput screening. Several cheminformatics tools are used in research, but integrating it with statistical methods are said to reflect the development of new algorithms and applications. These molecular modeling or cheminformatics methods strongly depend on the quantitative structure–activity relationship (QSAR) analysis. This QSAR technique is extensively applied to predict the pharmacokinetics property through the reference biological activity and it is one sound technique in the medicinal chemistry. Through this chapter, the basic principle of computational methods that relies on QSAR models, their descriptors, statistical phenomenon towards the molecular structures are discussed. At the same time, we also highlight the important components of QSAR models and their types to describe the molecular structure of lead molecules and discuss future limitations and perspectives to guide future research in the field of QSAR.

Keywords

Cheminformatics · QSAR · Descriptors · CoMFA · CoMSIA · Pharmacokinetics

V. Srivastava
Department of Biotechnology, Faculty of Engineering and Technology, Rama University Uttar Pradesh, Kanpur, Uttar Pradesh, India

C. Selvaraj (✉) · S. K. Singh (✉)
Computer Aided Drug Design and Molecular Modeling Lab, Department of Bioinformatics, Science Block, Alagappa University, Karaikudi, Tamilnadu, India
e-mail: selnikraj@bioclues.org

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_10

10.1 Introduction

Cheminformatics is a broad interdisciplinary field that encompasses chemistry and computing, primarily focusing on the extraction, processing, and extrapolation of data from chemical structures. In the field of drug discovery also cheminformatics plays a vital role to explore the chemical space, small molecule library design, pharmacophore, and scaffold analysis. Several data mining or machine learning algorithms are involved in converting chemical structure to chemical information. For that, it requires multilayer computational procession including descriptor generation, fingerprint construction, and similarity analysis to develop potential lead molecules and often used in several machine learning approaches to analyze the quality of the chemical data (Gasteiger 2003; Varnek and Baskin 2011a, b; Bajorath 2011; Kapetanovic 2008). The classical drug discovery process comprises seven major phases such as disease selection, prediction of target hypothesis, small molecule identification and its optimization, preclinical trial, clinical, and pharmacogenomics optimization. These steps are executed sequentially, and if an interruption may result in a slowing down of the entire process (Augen 2002). Previously, the drug discovery process was a time and cost consuming one to test new chemical entities. Since 1980, high-throughput screening approaches have been employed to predict the hit molecules, and approximately $7500 was calculated to develop a potent molecule against a disease (Xu and Hagler 2002). This has resulted in the replacement of new technologies to reduce the time and cost of synthesizing and testing new lead molecules. Many numbers of lead molecules against different target drugs are reported annually (Hecht 2002). Although hundreds to thousands of compounds are being assessed against drug targets, there is an increasing demand from biologists for the development of potent new compounds.As a result, nowadays the application of combinatorial methods has been carried out to develop several potential lead molecules in shorter periods. Combinatorial chemistry significantly produces a massive collection of compounds from a set of various types of chemical molecules called building blocks. Since 2000, several solutions and solid-phase combinatorial chemistry strategies have been developed (Hall et al. 2001). Parallel synthetic approaches are used in several pharmaceutical companies to increase the efficiency of the manufacture and testing of lead molecules in the drug design approach. Recently, the pathway approach has been used to address the above challenges and improve the chemical diversity of libraries. Through these emerging methods, several compounds have been discovered by combinatorial technology, leading to powerful drug candidates. To prevent the waste of chemical combinatorial determinations, it was assumed that it can act as a superlative to develop the different and similar structure function of lead molecule libraries. This can be subject to wide variety of chemical compounds and along with the structure processing technology, for biodiversity analysis. Several chemical-diversity-related methods were set up and utilized to find the more hit molecules from the chemically diverse libraries which take both drugs-like and non-drug like compounds. Thus, these methods have been applied to distinguish the potential lead molecules from a different compound library (Xu and Stevenson 2000; Clark and Pickett 2000; Matter et al. 2001). These filters have significantly solved the problems associated with the selection process, although they have not been sufficient to overcome this problem. It has been assumed that more sophistication of these screening methods must be trained for the identification of potent lead molecules (Oprea et al. 2001; Proudfoot 2002).

QSAR is a widely used method in the drug development process over the last few decades to develop mathematical models to find statistically significant chemical structures and its relationship between structure and activity using regressive analysis (Cherkasov et al. 2014). These days, QSAR modeling has expanded, and there are concerns about building a predictive model of biological activity. The QSAR concept has been extensively used for various applications including drug discovery processes for relating both the biological and physicochemical properties of lead molecules; therefore, it was known as quantitative structure–property relationship (QSPR). QSAR is used in the diagnostic process to find the relations among compound configuration and biological activity. QSAR has developed to satisfy the medicinal chemist's desire to predict the biological response. QSAR models are often used for the comparative analysis with molecular descriptors including various representatives (1D, 2D, and 2D) which results of final computational processes to describe the structure and behavior of the molecules in the system. Then the final outputs of QSAR have mainly computed the set of the mathematical equation relating the chemical structure to biological activity (Eriksson et al. 2003; Golbraikh et al. 2003; Wedebye et al. 2015). The current review discusses the predictive QSAR modeling and development procedures for validation and application in the drug discovery process. It also discusses the various molecular descriptors in QSAR methods, successful QSAR based screening of compounds, and its applications in computer aided drug design and environmental chemical risk valuation and future perspectives.

10.2 Overview of QSAR

In the year 1868 Crum-Brown and Fraser first proposed an equation and considered the first generation of formulation of QSAR in their investigations of various alkaloids (Crum-Brown 1868). Initially QSAR was used for working on the narcotic activity of various drugs. Later, Hammett introduced the new method for the account for the effect of the substituent on the reaction mechanism (Hammet 1935). QSAR studies correlated the affinities of the ligands with their respective receptor, rate of constant, inhibition constant, biological endpoints with the atomic group, and other properties including lipophilicity, the polarizability nature of the molecule. Though, this approach has limited utility for the development of novel lead molecules owing to the nonexistence of the three-dimensional structure of the molecules. To overcome this problem, 3D-QSAR has been used as the extent of the classical Hansch and Free-Wilson methods to exploit biological activity of the molecules via chemometric techniques like PLS, ANN. The techniques have aided as a valuable implement in the drug development process, especially in pharmaceutical and agrochemical industries. Though the trial and error issue complicated in the design of novel lead molecules cannot be unnoticed, QSAR undeniably reduced the quantity of the compounds and enhanced the selection of the potent lead molecules and its synthesis (verma et al. 2010; Dessalew et al. 2007). Topological techniques proposed by Enslein et al. to evaluate the structure and biological activity of molecules with minimum topological difference (MTD) method of Simon. Recently, electro topological indices have been used to encode the significance of structural statistics on the topological state of the atom and fragments and also their electron valence has applied to biological data (Tong et al. 1998), the other developments in QSAR include inverse QSAR, HQSAR, and binary QSAR (Gao et al. 1999).

10.3 QSAR Methods

To build the target-specific structure–activity representations of known chemical structure can guide HTS by rapid screening against small molecule libraries for most promising candidates. The screening significantly reduces the sum of trials and allows for the practice of more complexes with low throughput assays (Bajorath 2002; Singh et al. 2006). This model offers understanding closeness of the chemistry of the lead molecules and the sequential screening approach allows more rational improvement concerning the high quality of the lead molecules (Rusinko III et al. 2002). The following methods are widely used in the classical methods of QSAR.

10.3.1 Forward Selection (FS) Method

This method enhances the descriptors to the QSAR model at a time, which includes the regression approach that provides the maximum fitness function, the selected descriptor designated the force into all the QSAR models and novel descriptors gradually supplemented to the regression and each descriptor contributes the highest fitness function during the addition of previously preferred one (Guyon et al. 2002). Though it has some disadvantages, a set of descriptors communally considered as a good predictor, but it gives poor prediction when it is alone, hence the forward selection methods have been in several QSAR studies (Wegner et al. 2004).

10.3.2 Backward Elimination (BE) Method

This method starts with all descriptors that are selected and are screened one by one to reduce the quantity of the descriptors based on their impact on decline of an error benchmark like a sum of the squares. This process is completed when all the descriptors are removed; to date only some degree of reports has used this method as the flexible selection approach (Yang et al. 2003).

10.3.3 Stepwise Selection (SS) Method

This method is the well-known and extensively used method in QSAR and has a stepwise descriptor which has a step-step procedure. The assortment time of this model development starts deprived of any descriptors in the regression equation. Each step has introduced the descriptors which offer the highest fitness function applied (for example, prediction of biological activity with correlation coefficient), but it also analyzes the importance of the descriptors in the regression QSAR model. The process is ended when the important descriptors are not present in the pool and satisfy the collection benchmark. Hence it is an unpretentious and an authoritative method to acquire a subset of a descriptor, but it does not interpret the artificial neural network methods.

10.3.4 Variable Selection and Modeling Method

Liu et al. (2003) proposed this technique based on the estimation of two statistics such as the interrelation coefficient of two descriptors (R_{int}) and the correlation coefficient (q2) is designed with leave-one-out (LOO) cross-validation method. It was familiarized into the whole subset to progress the performance. There are two main features such as controlling the exploration of several optimal subsets by q2 in the LOO cross-validation and examining speed of all top subsets by R_{int} together with q2 to differentiate this from other methods.

10.3.5 Leaps-and-Bounds Regression

In this method leaps-and-bound regression was used for the selection of descriptors (Xu and Zhang 2001), this regression can quickly find the best descriptor subset devoid of checking all the conceivable subsets based on the following inequality (Furnival and Wilson 1974).

$$RSS\,(A) \leq RSS\,(Ai)$$

where A is any set of descriptors, RSS denoted the sum of residual squares, and Ai is a subset of *A*. based on the number of the above methods of subsets are evaluated to search the best subsets. All the above methods are essentially linear and also have some disadvantages that may well not be in effect where the relationship between the activity and descriptors is nonlinear and causes coincidental correlation where several variables are screened for inclusion in the model.

10.3.6 QSAR Modeling and Development

QSAR modeling is one of the major cheminformatics approaches to determining the dependency of chemical, biological properties on the molecular features, traditionally used as lead optimization approaches (Reddy et al. 2012, 2013a, b). The QSAR based virtual screening has been used in pharmacokinetic property protection and chemical risk calculation to find the potent hit molecules (Singh et al. 2006; Dessalew and Singh 2008). The advanced approach has been allowed the enhancement procedure for QSAR model prediction (Suryanarayanan et al. 2013; Vijaya Prabhu and Singh 2018; Panwar and Singh 2020).

10.3.7 Internal Model Validation

In this technique, the developed QSAR models were endorsed within the LOO cross-validation method, where a compound is removed randomly from the dataset each phase, and the model is assembled with the other molecules. Then the final model was then used for calculating the action of the abolished molecules. This development is recurring several times till all the molecules are eliminated once. The following equation is used for the calculation of the cross-validated squared correlation coefficient.

$$Q^2 = 1 - \frac{\sum (Y_{\mathrm{Obs}} - Y_{\mathrm{Pred}})^2}{\sum (Y_{\mathrm{Obs}} - Y)^2}$$

where Y_{Obs} represent the observed activity of the training set, Y_{pred} represents the training set activity, and Y represents the mean values of the training set activity, and also, the modification (R^2($_{\mathrm{adj}}R^2$) of R^2 that corrects the quantity of explanatory terms in a model also calculated. In this method, the addition of descriptors was established which raises its value of the $_{\mathrm{adj}}R^2$ when the new term improves the model than expected by chance (Roy et al. 2012). The R^2 calculated to overcome the drawbacks by the following expression.

$$\mathrm{adj}R^2 = \frac{(n-1)R^2 - P}{n-p-1}$$

where p denotes the predictor, variables used for the model development, the overall implication of the regression coefficients and variance ratio F was calculated by the following equation.

$$F = \frac{\sum (Y_{\mathrm{cal}} - Y)^2}{P}$$

10.3.8 External Model Validation

This model was used to define the prophetic ability of the established model and set of test activity values for the calculation of the predictive R2 value by the following equation

$$R^2{}_{\text{pred}} = 1 - \frac{\sum \left(Y_{\text{pred (test)}} - Y_{\text{(test)}}\right)^2}{\sum \left(Y_{\text{(test)}} - Y_{\text{(training)}}\right)^2}$$

where the predicted and observed activity of the test compounds were represented by $Y_{\text{pred (test)}}$ and $Y_{\text{(test)}}$, respectively. The mean activity value of the training set was represented by $Y_{\text{(training)}}$ and the $R^2{}_{\text{Pred}}$ indicates the correlation coefficient of all the test compounds.

10.3.9 Randomization Test

The Y-randomization method is mainly used to check the robustness of the expected QSAR model and implemented by permuting the reaction values, including activity (Y) and the descriptors (X) matrix was unchanged. The sum of deviance of correlation coefficient of randomized model from the squared mean correlation coefficient of the non-randomized model was calculated by the following equation (Roy et al. 2015).

$$R_P^2 = R^2 X \sqrt{\left(R^2 - R_r^2\right)}$$

The above equation denotes that the value of R_P^2 and R^2 must be equivalent for the development of the QSAR model. This led Todeschini to define the correction for R^2 by the following equation.

$$cR_P^2 = R\sqrt{\left(R^2 - R_r^2\right)}$$

This model developed to penalize both randomized and non-randomized models difference among the squared correlation coefficient and the values cR_P^2 was calculated for each model.

10.4 Molecular Descriptors

In QSAR, molecular descriptors provide the important statistics of molecules in expressions of physicochemical parameters like electronic, geometrical, steric, hydrophobic, solubility, and topological descriptors (Helguera et al. 2008; Xue and Bajorath 2000). Molecular descriptors are significantly mapping and assembly of molecules into a set of mathematical and binary values that symbolize the several molecular properties that are significant for understanding the activity of the molecules. There are two broad categories of molecular descriptors used based on 3D alignment and conformation of the molecules (Kumar Singh et al. 2007).

10.4.1 2D QSAR Descriptors

This method shares common properties of the compound based on independence of 3D alignment of the molecules. The descriptors used in these approaches are mainly applicable for measurement of chemical entities with its topological and geometrical properties to compare the quantum-chemical and fragment counting method.

10.4.2 Constitutional Descriptors

This type is extremely fast and easy to capture and correlates the molecular properties that are associated with features founding its structure. Some of the constitutional descriptors comprise the total number of molecules and atoms of a different identity. Besides, some properties related with bonds including single, double, and triple and aromatic types bonds.

10.4.3 Electrostatic and Quantum-Chemical Descriptors

The Quantum-chemical descriptor provides statistics on the electronic environment of the molecules and highlights both the negative and positive charges of the molecule and the molecular polarizability and demonstrates the intermolecular bonding. The highest occupied and lowest unoccupied orbitals used to practice the quantum-chemical descriptors (Lewis 2005; Stanton et al. 1992).

10.4.4 Topological Descriptors

Topological descriptors are involved to delight the configuration of the molecules as a diagram, with atoms and covalent bonds as vertices and edges, respectively. This method is used for defining the molecular connectivity of the many indices preparatory with the Wiener index that sums all the bonds among non-hydrogen atoms. The Randic indices x is the type of topological descriptors which define the sum of geometric averages of edges within the paths of given lengths, Balabans's J index (1982) and Shultz index (1989) (Balaban 1982). Kier and Hall indices xv provide the information of valence electrons, using symmetrical means of valence connectivity beside paths and Galvez topological charges indices also provide the information like Kier and Hall indices xv to measure the topological valences and net charge transfer between atoms by a given number of bonds. Topological sub-structural molecular design (TOSS-MODE/TOPS-MODE) is another type of descriptors worked on the basis of spectral moments of bond neighboring modified matrix with bond polarizability (Estrada and Uriarte 2001). Atom type electron topological indices (Hall and Kier 2000) described the intrinsic atom state and the disconcertion with the electronic and topological organization.

10.4.5 Geometrical Descriptors

These descriptors are involved to provide the data on atomic van der Waals areas formed molecular surface depending on the three-dimensional organization of atoms creating the molecules and provide the data on molecular volume. The principal moments of the gravitation indices capture the three-dimensional organization of the molecule obtained by the projection (Labute 2000).

10.5 3D QSAR Descriptors

The 3D-QSAR approach is more composite than 2-D QSAR and is involved in numerous phases to find the numerical descriptors of the molecular structure. First, the conformation of the molecules is predicted from investigational records or molecule mechanics then it is refined by minimizing the energy (Guner 2002; Akamatsu 2002). In the second step, these conformers are consistently associated in space and finally its occupied conformers are investigated computationally for different descriptors. In this method, various descriptors including geometric, quantum-chemical, and physical characteristics were used to designate the 3D features of the lead molecules. These molecular descriptors are combined and generate a pharmacophore model to elucidate the various features of the molecules such as the number of hydrogen bond acceptors, donors that are crucial for the evaluation of desired biological activity (Chang and Swaan 2006). Then the final model of 3QSAR was obtained by the evaluation of stability and statistical significance of the pharmacophore model. Several studies have reported the various techniques of 3D-QSAR modeling and are widely used for drug design (de Groot and Ekins 2002; Van Drie 2003). Table 1 represents the classification of QSAR models based on different criteria which are widely used in drug discovery approaches.

10.6 Alignment-Dependent 3D QSAR Descriptors

Before the calculation of descriptors, several approaches are required for the alignment of molecules that strongly depend on features of the receptor for predicted ligands. Besides where the data is accessible the alignment of molecules is studied using receptor–ligand complexes. Then, the computational tools are used for overlaying the configurations in space hence these methods are highly dependent on atom–atom or substructure charting (Lemmen and Lengauer 2000).

10.6.1 Comparative Molecular Field Analysis (CoMFA)

The CoMFA generally used two crucial energy fields such as coulombic and van der Waals energy fields to study the molecules. Here the molecules are aligned and placed in a 3-D grid. This analysis allows identifying positive and negative charged regions of the structures; hence it was extensively used 3D QSAR methods to describe the shape-dependent steric and electrostatic possessions of the lead molecules with its biological activity. Based on the 3D configurations on the 3-D grid molecules are aligned by both steric and electrostatic potential energy values at each grid point. It generally adopts the calculation of the minimum energy conformer of the bioactive molecules. The known crystal structure and its matches may be used to describe the bioactive conformers (Gohda et al. 2000). Partial least square analysis (PLS) and principal component analysis (PCA) methods are generally used in CoMFA model development and then it was subjected for geometric importance and strength. The predictive stability of the CoMFA model is overly delicate to the orientation of the bioactive conformers (Yasuo et al. 2009). However, the lowest energy conformers of the bioactive molecules in the absence of receptors are flouted (MacKerell Jr. 2004; Hasegawa et al. 2000; Koehn and Carter 2005). It uses Lennard-Jones and Coulombic utility to compute the steric and electrostatic interface (Flower 2002a, b).

10.6.2 Comparative Molecular Similarity Indices Analysis (CoMSIA)

CoMSIA is like CoMFA which also immerses the molecules in the regular grid lattice and calculates the similarity between the probe atoms. Compared to the CoMFA it uses and calculates the different functions like steric, hydrophobic properties, hence the probe atoms have additional hydrophobic properties. The application of Gaussian types in CoMSIA as an alternative of Coulombic function and Lennard-Jones allows precise data in the grid situated within the molecules. In the case of CoMFA, the huge values are acquired in these points owing to the prospective functions and random cut-offs (Klebe et al. 1994). In this method it uses the hydrophobic and hydrogen bond donor and acceptor and also steric and coulomb contribution to study the similarity indices (Flower 2002a, b). Then the bell-shaped Gaussian function was used for the calculation of electrostatic and steric components of the energy (Acharya et al. 2011).

10.6.3 Weighted Holistic Invariant Molecular Descriptors (WHIM)

This type of descriptor provides the invariant statistics by analyzing the principal component analysis on the coordinates of molecules and transforms the molecule into the space that captures the variance. Several statistics were reused to calculate the proportion, variance, and symmetry of the molecules. The combination of directional and non-directional descriptors is also defined in this method. Principal component analysis of the molecules and impact of each atom can be analyzed by chemical properties and the atoms can be weighted by mass, atomic electronegativity, polarizability, and molecular electrostatic potential (Douali et al. 2003).

10.6.4 VolSurf

The VolSurf method relies on examining the grid of a molecule with specific probes like hydrophobic interaction. Then the lattice boxes are utilized to compute the descriptors on the surface of the 3D contours. The utilization of several investigations and cut-off values of various molecular properties such as surface, molecular volume, and hydrophilic regions are quantified. Besides, derivative quantities like molecular globularity that are related to the surface of hydrophilic regions of the whole molecules are also computed, and geometric-based descriptors are also available in this method (Cruciani et al. 2000; Crivori et al. 2000).

10.6.5 Grid-Independent Descriptors (GRIND)

As VolSurf, the GRIND also uses the probing of a grid with a specific probe to overwhelm the issues with interpretability in configuration free descriptors. Theregions of most favorable energies of the molecules are designated and provide the distance between the regions is large. Then the probe-based dynamisms of the molecules are calculated in a way and in the final step, the distance among the nodes in the grid in the set of bins. The distance of each bin encodes the nodes with the maximum energy and stored and then the values are used as numerical descriptors. Then the stored information of each node can also be used to track the exact region of the molecules (Pastor et al. 2000).

10.7 QSAR Based Screening

QSAR is an effective technique for constructing the accurate models used to find a statistically important relationship of chemical structure and continuous PIC50 and PEC50 value or biological property using regression analysis and also for binary/categorical properties like active, inactive, and nontoxic (Cherkasov et al. 2014). In the past years, QSAR modeling has been modified in numerous ways ranging from 1 D to nD and different methods to find the association among the chemical structure and the biological activity. The classical QSAR has some limitations to analyze the insignificant series of congeneric lead molecules and simple deterioration methods. Recently, several modifications and improvements in QSAR modeling were differentiated and progressed to modeling and screening of diverse chemical structures from large data sets with an inclusive diversity of machine learning techniques (Cherkasov et al. 2014; Ekins et al. 2015; Goh et al. 2017; Mitchell 2004).

10.8 QSAR Modeling and Validation

High-throughput screening is one of the suitable approaches for QSAR modeling to an explosion of the amount of data and the major limitation of this method was it results in errors on both chemical and experimental while constructing predictive models (Ekins et al. 2015; Young et al. 2008; Southan et al. 2009; Williams and Ekins 2011). To overcome these issues, Fourches et al. (2015) reported the guidelines for both chemical and biological statistics curation with mandatory steps for QSAR modeling. These guidelines allow the correction and identification of the removal of counterions mixtures and also the normalization of specific chemotypes, standardization of tautomeric forms, structural cleanings like detection of valence violation, and ring aromatization. Besides, the removal of replicas and curation of molecules results in the production of a single bioactivity result (Fourches et al. 2015). A set of updated procedures for the researchers was established by the Organization for Economic Cooperation and Development (OECD) to attain the narrow acceptance of QSAR models, conferring to these strategies, the QSAR model should be related with the following points such as it should define the endpoint, the domain of applicability, unambiguous algorithm, suitable measures of strength, and mechanistic interpretation.

Recently, a number of computational methods have been employed to discover the lead molecules in the early stage of drug discovery in a data-driven process, which was obtained from HTS campaigns (Nantasenamat et al. 2014). Subsequently the finding of novel lead molecules in HTS is cost-effective, QSAR plays a vital role in ordering lead molecules either synthesis or biological evaluation. QSAR models are highly utilized for both finding and optimization of lead molecules. Later the connectivity and promising stability between the pharmacokinetic and toxicological parameters and selectivity of the molecule which required for the development of novel, nontoxic, and efficient drugs could be attained via several optimization steps. This process considerably reduces the employment, time, and commercial method to find potent lead molecules with desirable biological properties. Hence QSAR has been extensively used in several applications including academic and scientific organization all over the world (Cherkasov et al. 2014). Figure 10.1 represents the general scheme of the QSAR modeling established by virtual screening method. Initially, the datasets were retrieved from data sources and are curated to remove the inconsistent data. Then the data has been utilized in QSAR models development and validated with OECD guidelines and best practice modeling and used to identify the

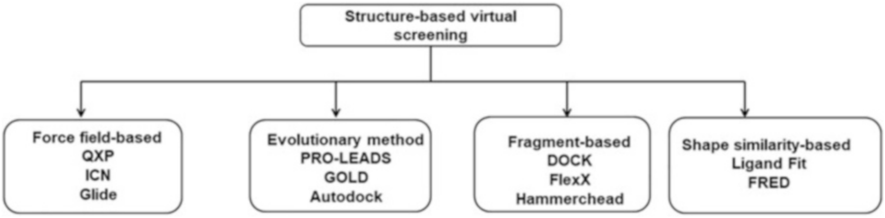


Fig. 10.1 Different methods used in structure-based virtual screening approach

lead molecules prediction from bulky chemical libraries (Cherkasov et al. 2014). The screened molecules were analyzed to reduce the number of compounds from a large chemical library and then will be subjected to experimental assay. However, it is more important to study the workflows and additional filtering steps like empirical rule setting (Cherkasov et al. 2014). Though, the investigational confirmation of computational hits the QSAR methods should be implemented as the final essential step. Then the investigational confirmation with QSAR prediction of effectiveness and pharmacokinetic constraints should be piloted. This statistics is essential for lead molecule optimization and design of the new molecule sequence to select properties like potency, pharmacokinetics, and selectivity which are related to the effect of diverse configurations to design the new lead molecules with a specific target.

Hence, the evaluation of virtual methods is crucial before screen large libraries with suitable approach and successively generates accurate outcomes on an actual project. Thus, several software and workflows are systematically calculated with benchmark datasets. Such datasets have both known active and inactive molecules called decoys (Irwin 2008; Selvaraj et al. 2014). Preferably, both active and inactive molecules are selected based on experimental assays. Receiver operating characteristics curves (ROCs) are used as shared metrics to evaluate the performance of the virtual screening methods (Triballeau et al. 2005). Besides, the enrichment factors (EF), enrichment curves (EC), and predictiveness curves are theoretically different and all share the similar objective to evaluate the capability of a process to find the active compounds and also distinguish them from decoy compounds (Empereur-mot et al. 2015). The alignment of both active and decoy molecules has been shown to slope virtual screening assessment outcomes. The space between the two chemical spaces of the molecule is determined by active molecules and the decoy molecules were used to evaluate the synthetic overestimation of improvement (Bissantz et al. 2000). In contrast, the existence of decoy molecules with active compounds may lead to an artificial underestimation of the improvement (Good and Oprea 2008). A number of studies have reported that the application of virtual screening on the structural properties of a target such as binding sites, physicochemical properties, and structural flexibility (Cummings et al. 2005). Based on the above information and increasing the protein families in databases, a decoy set of compounds were prepared publically accessible to find consistent benchmarking datasets (Ibrahim et al. 2015; Mysinger et al. 2012).

Table 10.1 The classification of 3D-QSAR methods based on different criteria

Classification		Models
Based information used	LB-3D-QSAR	CoMFA, CoMSIA, COMPASS, GERM, CoMMA, SoMFA
	SB-3D-QSAR	COMBINE, AFMoC, HIFA, CoRIA
Based on orientation Benchmark	Alignment-dependent	CoMFA, COMSIA. GERM, COMBINE, AFMoC, HIFA, CoRIA
	Alignment-independent	COMPASS, CoMMA, HQSAR, WHIM, EVA/CoSA, GRIND
Based on chemometric technique	Linear-3D-QSAR	CoMFA, CoMSIA, AFMoC, GERM, CoMMA, SoMFA
	Nonlinear 3D-QSAR	COMPASS, QPLS

The QSAR Model validation is one of the most important steps in QSAR based applications, recently, several numbers of acceptable statistics criteria have been used for the estimation of the test compounds in QSAR modeling (Zhang et al. 2006). This critical component is established as reliable and widely accepted practices for model development. Also, establishing the model's domain applicability is the major problem in QSAR analysis. The absence of applicability domain in the model can easily calculate the activity of the molecules even with structurally diverse training set compounds. Thus, the absence of these domains as a necessary element in the QSAR model evaluation would lead to the inexcusable prediction of the model in the chemistry space, which leads the inaccurate prediction (Kovatcheva et al. 2003; Shen et al. 2003, 2002). Mandel (Mandel 1982) reported that the effective forecast domain based on the various range of descriptors including the regression analysis. In another study, Afantitis et al. (2006) demonstrated the multiple regression analysis models for a dataset of apoptotic agents and reported that applicability domain highly influences the equivalent transverse component of the hat matrix. This method can predict possible leverages outliers. Netzeva et al. (2006) and Saliner et al. (2006) have reported the applicability domain with various ranges of descriptors that are occupied by representative points which have significant drawbacks due to the representative points. Tong et al. (2004) also defined the same applicability domain to construct the QSAR models for two datasets with decision forest method to study the dependence of the model vs. applicability domain for the accurate calculation. The precision was 50% for the initial applicability domain and increased when the applicability domain increased by 30%.

10.9 Decoys Selection

The evaluation of virtual screening tools with the use of a benchmarking starts in the year 2000, with the inventive work of Bissantz et al. (2000). The evaluation of ligand enrichment is the main objective of their study, i.e., the selection of hit molecules with the best dock score was found from the docking programs. There are three docking programs such as Dock (Kuntz et al. (1982), Gold (Jones et al. 1997), FlexX (Rarey et al. 1996) shared seven scoring functions assessed on two different target proteins. For each target, a minimum of 10 known compound datasets with 990 small molecules that are implicated to be inactive (decoy compounds) were generated. (1) These molecules were used to eliminate the undesired compounds and (2) then most of the molecules were randomly selected by filtered dataset and are used to estimate and match the number of docking and scoring schemes. This procedure was applied to investigate the structural similarity of the molecules against three human G-protein coupled receptors (GPCR) (Bissantz et al. 2003). Today, the high focus is on virtual screening methods, to get the new compounds from available databases along with utilizing the decoy set for comparing the actives (Kellenberger et al. 2004; Brozell et al. 2012; Neves et al. 2012; Repasky et al. 2012; Spitzer and Jain 2012). Hence, the benchmarking databases are widely used to evaluate the various virtual screening models and support the finding of potent lead molecules with both ligand and structure-based virtual screening (Allen et al. 2015; Ruggeri et al. 2011).

10.9.1 Physicochemical Filters to the Decoy Compounds Selection

Based on the dimension of the decoy sets Diller and Li (2003) incorporated the physicochemical filters for decoy selection. In addition to the kinases inhibitors (1000), they also retrieved six kinases (EGFr, VEGFr1, PDGFrβ, FGFr1, SRC, and p38), and 32, 000 compounds from the literature were randomly selected from MDDR (MDL drug data report). This filter was specially intended to find best the decoy compounds based on comparable polarization and mass. In the year 2003, McGovern and Shoichet developed a new benchmarking database with MDDR where undesired functional groups were removed, and MDDR database with at least 20 known ligands was accessible and found a target dataset (MMP3, NEP, and XO). The enduring molecules were used as decoy sets (McGovern and Shoichet 2003). The filters were used in benchmarking databases in an early stage; the potent and highly active molecules from literature while the decoy molecules were involved in the presumed inactive molecules were selected from large databases which are significantly filters to specific criteria such as molecular weight, drug-likeness property, etc. Due to the use of MDDR and filtering the decoy compounds, these benchmarking databases show remarkable drawbacks; the difference occurring between the physicochemical parameters and decoy molecules was led to perception and then good enrichment (Verdonk et al. 2004; Huang et al. 2006).

Irwin (2008) reported that the similar decoy molecules are known as active molecules which significantly reduce the bias while the dissimilar compounds are known to reduce the probability of the target protein. Based on the above consideration they constructed DUD databases used for the evaluation of virtual screening methods (Huang et al. 2006). This database contains 2950 ligands and 95,326 decoy molecules for 40 protein targets from 6 major protein families such as serine protease, folate enzyme, nuclear hormone receptors, and kinase. The structurally similar compounds retrieved from ZINC database and the decoy molecules were calculated by estimating the Tanimoto distance based on the physicochemical properties. Around, 36 molecules were sharing the most comparable properties for a single active molecule. The evaluation of the DOCK confirmed the uncorrected databases like MDDR led over-optimistic enrichment associated with the improved database such as DUD.

10.9.2 Benchmarking Database Biases

Based on the literature, several biases have been used to build the DUD database; where the analogous has limited space for the active molecules and is restricted to the series of chemical compounds that have been explored in the databases (Good and Oprea 2008). The refinement of the active molecules from the decoy compounds is simplified by larger structural variability and can induce the virtual screening performance due to the lack of structural diversity of active molecules which limits the evaluation of ligand-based virtual screening methods for predictions of potent lead molecules and are structurally different from the reference compounds while retaining the similarity and activity. The artificial enrichment bias often displays the differences in inactive compounds and decoy compounds with their respective structural complexity and optimizes the compounds retrieved from a large dataset in the patent literature (Stumpfe and Bajorath 2011). The false-negative decoys have makes its presence, even in the active molecules in the random decoy set to show the screening performance of hit compounds (Vogel et al. 2011; Bauer et al. 2013). To eradicate the least minimum of these biases, new strategies were emerged to evaluate the virtual screening methods.

10.9.3 Structure-Based Method

In a structure-based approach, the crystal structure coordinates are often used to design the novel potent molecules. Recently several structure-based methods are used in drug discovery methods, generally the basic two subclass of methods such as molecule growth method and fragment-position method. There are two distinguished programs like GRID and multiple copy simultaneous searches (MCSS) that were used in the fragment-position method to find the binding pockets that are energetically favorable for interacting fragments (Dean 2005). In the fragment method, a fragment is engaged in the binding pocket then ligand molecules are grown by binding fragments. For instance, the application of algorithms should be able to reproduce known chemotypes for different drug targets (Dean 2005). The programs such as small molecule growth (SMoG), GrowMol, GenStar, and GROW are often used in fragment methods. Docking and scoring are the two significant mechanisms in structure-based virtual screening, where molecular docking brings the two molecular types organized with prophesied positioning, while the scoring function assesses the binding affinity of these two molecular species. Since 1980

Table 10.2 Different docking protocol used in drug discovery process

S. No	Name of the tool	License terms	Scoring function
1	Auto Dock	Free	Based on force-field methods
2	DOCK	Free	Chem score, GB/SA solvation scoring function
3	Flex X	Commercial	FlexX score, screen score, drug score
4	FRED	Free	PLP. Gaussian shape score
5	Glide	Commercial	Glide score, glide comp
6	GOLD	Commercial	Gold score, chem score
7	Ligand Fit	Commercial	Lig score, PLP, PMF

several docking programs including GOLD, AutoDock, Surflex-Dock have been developed (Kitchen et al. 2004; Moustakas et al. 2006; Rester 2006; Schneider and Fechner 2005). Most of the docking platforms adopt the protein to be rigid and due to its flexibility. However, some of the docking platforms that study both ligands and proteins are more flexible to produce enhanced results than rigid docking (Dean 2005). Table 10.2 contains various information's of the docking approaches, their licensing terms and applicability domains. Different docking approaches have successfully docked the molecules into a binding site depending on the precision of the scoring function that ranks the molecules based on the mode of interaction and how they will bind into the receptor site. Generally, four key classes of scoring functions such as knowledge-based, force-field based, consensus scoring, and empirical scoring functions are used for the prediction of approximate binding free energy. Among them, calculation of scoring functions based on force-field was used as the standard molecular mechanic's energy such as electrostatic interaction and van der Waals. In the empirical scoring function, the binding free energy can be calculated by sum of ligand and receptor interaction parameters. Whereas the knowledge-based scoring functions use the sum of protein–ligand interaction and atom-pair interaction to computing the binding affinity of the promising atom in contact with each other are predicted (Dean 2005; Schneider and Fechner 2005).

10.9.4 Ligand-Based Method

The unavailability of the 3D crystal structure of the target results in ligand-based design methods. Such lack of information leads to alternative methods to use the potent lead molecules of a target protein as the source to find the potent novel structures. The ligand-based method uses the pharmacophoric, structural, and topological features of known molecules and to discover the presumed ligands. In pharmacophore-based screening, the ligand pharmacophore model was generated, and then it was used as a query to screen the novel potent molecules from the large chemical databases which are complementary to the primary target. The ligand-based virtual screening depends on the essential resemblance of the query molecules, in this approach, the test molecules are superimposed to the reference molecule and similarity of the molecules was assigned. Such similarity values are then used as scoring factors to find the best hit molecules. However, these alignments are highly required for manual interference and a substantial volume of time hence the descriptor-based approach is introduced for faster screening. Based on the number of properties that have been used descriptors are classified into 1D, 2D, and 3D. 1D descriptor represents the molecular properties such as mass and logP, whereas 2D descriptors categorized as linear molecular properties have two cases such as real value and binary descriptors (Hessler et al. 2005; Reddy et al. 2007). These methods are applied differently to break molecules into a fragment and then used for virtual screening, for example, the Topomer search from Tripos is one of the fragment-based methods which can perform the R-group search or whole molecules search (Tripos 2007).

10.10 Inverse-QSPR/QSAR

The inverse QSAR modeling is used to evaluate the values of descriptors and also generate the QSAR model with high activity and build the small molecule structure from these values (Skvortsova et al. 2001). Figure 10.2 represents the overall workflow of QSAR approach. The presence of numerical signatures of the active molecules and its activity may be considered a major challenge in this method, hence it should be re-translated into chemical structure with high activity to overcome this problem (Schneider and Baringhaus 2013; Speck-Planche and Cordeiro 2017). It uses multiple linear regression models to build a chemical graph that corresponds to the multiple linear regression equation (Schneider and Baringhaus 2013). For example, the particular descriptors have been introduced for inverse QSAR based on the multiple regression analysis equations and algorithms (Faulon et al. 2003; Churchwell et al. 2004; Weis et al. 2005). Inverse QSAR was divided into two-stage processes and is theoretically based on the significant principle adopted from conventional QSAR for the prediction of higher activity values of more necessary chemical structures. This two-stage process of inverse QSAR conjugates the challenges of the descriptor value generation that corresponds to the higher activity than the currently available training set. Besides these high activity compounds are significantly optimized by Gaussian mixture models and cluster wise multiple regression for the development due to the multi-parametric nature of training data were ordered into several clusters. With available techniques, the new onboard techniques are the two-stage inverse QSAR modeling and that was used to construct the descriptor space with autoencoder modeling for encoding a line notation of molecules of recurrent neural networks (RNNs). The optimized coordinates in latent space can directly translate into another line notation by RNNs and this method does not depend on any descriptors and has the possibility to mechanically address two-stage inverse QSAR (Gómez-Bombarelli et al. 2018). Recently the field of cheminformatics, the combination of mathematics and computational resources has emerged as a potent approach to solve the chemistry-related problems and handling the large datasets. The statistical method QSAR/QSPR models the chemical molecules with similar structure and properties and mainly

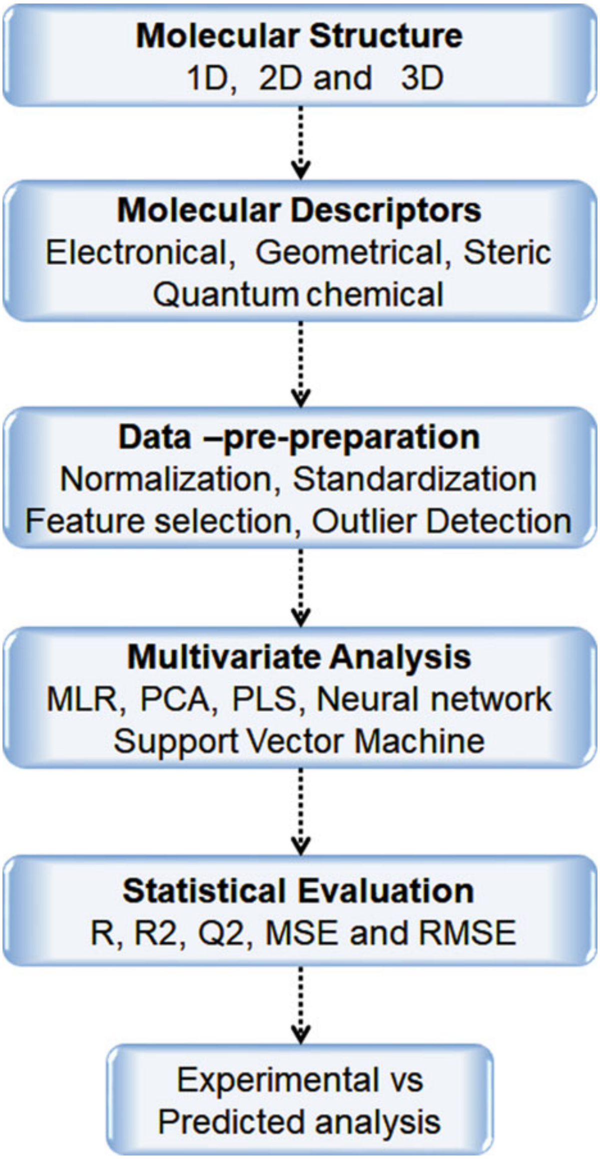


Fig. 10.2 The overall workflow of the QSAR methods in the drug development process

highlights the relationship between the structure and properties of the molecules (Nieto-Draghi et al. 2015). The identification or synthesis of novel molecules with desired properties has been examined since the support of QSPR models. The recent advance in the QSAR, i.e. tuned for i-QSPR models primarily established to calculate the property values. The stochastic method also used to monitor the structure prediction using molecular identifiers such as SMILES neural networks for the group of diversity.

10.10.1 Unguided Generation of Molecules

Computing all possible combinations such as atoms, graphs, fragments, and bonds is the easiest way for the generation of molecules; Fink et al. (2005) enumerate the chemical molecules up to 11 heavy atoms and generate the database with possible graphs and substitution atoms. This filter is useful for removal of unrealistic structures including molecules with bad valence. They also generated databases like GDB-13, GDB-17 database (Blum and Reymond 2009; Ruddigkeit et al. 2012). The software like MOLMAKER (Clark et al. 1996) and Makino et al. (1999) are recently used methods that perform reaction-based computing to produce the possible products library. These methodologies are also used in a two-step algorithm to predict the entire molecules by combinatorial association and then subjected for the screening using QSPR models. To perform i-QSPR the virtual library molecules have combined the fragments and used to screen the molecules to find the potent molecules with given application conditions. I-QSPR with GCM also has extensively used methods to find novel molecules with desired properties from known structures. The generation of a molecule is described with valence and types of groups attached at the point and the combination of these groups was controlled by about 4 roles such as aromatic group is attached as single point supplement constituent, supplement group, in this case, aromatic ring of the double point supplement does not form a double bond and also could not combine. But in some special conditions, the screening was performed by the GCM model and the improvements were proposed by bonding rules and are characterized by their binding ability (Brignole et al. 1986). These groups were considered as intermediate to their connectivity, the low and high binding reactions were set as feasibility criterion. Derringer and Markham (1985) demonstrated the contribution techniques to design the polymers with three specific properties like water absorption, density, and temperature by an unsystematic grouping of seven functional groups. This screening was implemented after the addition of each fragment. Pretel et al. (1994) generated the possible creation of aromatic groups with designed solvents and generation by scheming transitional groups having more than one free valence atom. Bolis et al. (1991) reported the thermolysin enzyme inhibitors by computerizing the group selection step and produced a classification process to identify the desired property. The use of GCM implemented with i-QSAR represents the more advantages but the combinatorial explosion is challenging to grasp. This problem can be easily elucidated by setting the proper setting rules by choosing the specific groups for the generation. In this specification, the groups must be contributing the potential bonds by their valence and the atoms can be bonded properly. Though GCM has some limitations to constrain the prediction of features to the existing fragments and the size of this fragment should cooperate among increasing the difficulty versus limiting the feature space consideration.

Recently molecular hologram QSAR (HQSAR) plays a key role in the finding of sub-structural features in molecules that are highly relevant to its biological activity. The key factor of this method is differentiating the other methods such as Free Wilson and CASE analysis of the molecular hologram generated fragments including the cyclic and overlapping fragments. Thus, each atom of these molecules occurs in multiple fragments, unlike the maximal common algorithm. HQSAR yields one of the demonstrations of QSAR modeling that was attained in a reflective analysis of the data set. Randomization of testing and redistributing the activity data which attempt the statistical model to relate the scrambled data with a molecular descriptor (Deshpande et al. 2002; Tropsha et al. 2003).

10.11 Workflow and Application of QSAR Modeling in Virtual Screening

Since 1964, an extensive variety of QSAR approaches has been designed with the concept of Free, Wilson, Hanch, and Fujita. The classical 2-D-QSAR methods use 2D molecule substituents and their physicochemical properties to analyze the quantitative prediction. Since then the evaluation of 3-D-QSAR method has been recognized as fast and development of first novel method known as comparative molecular field analysis familiarized by Crammer et al. (1988) which acts as a basis for the improvement of other advance approaches such as CoMSIA, SOMFA, CoMMA and also multidimensional nD-QSAR methods like 4D, 5D QSAR, etc., to overcome the difficulties. Hence, recent fragment-based methods show significant attraction and attention because effective prediction of molecular fragments with essential properties and potent activities is fast and robust (Zhang et al. 2007). Several lead molecules against a variety of diseases were predicted with implementation of QSAR approaches, for example, antimicrobial and antitumor compounds with strong activity and prediction of series of Xanthines derivatives as adenosine antagonists were also predicted QSAR modeling (Li et al. 2013). It also has been implemented in various studies to evaluate epothilones–tubulin depolymerization inhibitors (Lee and Briggs 2001). In addition, QSAR models significantly used structurally diverse antifolates like cycloguanil, aminopterin, pyrimethamine, and 13 pyrrolo [2, 3-d] pyrimidines (Santos-Filho and Hopfinger 2001). The implementation of topological polar surface area in 2D-QSAR has been used for the development of 14 sets of pharmacologically active compounds (Prasanna and Doerksen 2009). Hydroazones derivatives were also predicted as electron acceptors for xanthine oxidase with QSAR model and antiviral QSAR models were implemented to predict the potent lead molecules against 40 viral species with mt-QSAR model and Markov chain theory is used to compute the novel multitarget entropy of QSAR model (Prusis et al. 2004; Prado-Prado et al. 2011). In the drug designing approach, validation of the QSAR model is very important to conclude the results whether it satisfies the expectation or not. R2 and Q2 are two statistical measures that have been used for validation (Catalin 2014; Tang et al. 2016), where R2 represents the coefficient of multiple regression to measure the data and its fitness. The previous reports demonstrated that the value of R2 should be $\geq$0.6 to consider the best fit model. The Q2 represents the squared correlation which acts as an important criterion for the robustness. However, the values of R2 are not enough to calculate the model and tested for their capability to calculate the lead molecules with an external test set (Tang et al. 2016). It was proved by the good predictability of R2-Q2 values that should be in the range of below 0.3 which indicates the best probability of the models. The other applications of the QSAR study were described in the investigation of antidiabetic drugs based on sitagliptin as a potent antioxidant agent. Several descriptors such as rotatable bonds, hydrophobicity, hydrogen bond donor, acceptor atom were used based on the QSAR equation with improved pharmacological effect as DPP4 inhibitors sitagliptin as a novel potent molecule (Catalin 2014). Computational methods have been used as good predictive tools for the evaluation of inhibitory molecules, where the QSAR studies are often used with docking methods and neural networks. The implementations of important fields such as steric, electrostatic, and hydrophobic with QSAR were used for the prediction of xanthine oxidoreductase inhibitors (Veerasamy et al. 2011). Another study reported that the series of quinolones derivatives also predicted by the QSAR model with better caspase-3 inhibitors. Based on the QSAR model the new series of quinolone compounds and then calculated their caspase inhibitory activity (Sharma et al. 2008). The implementation of QSAR model with radial distribution function (RDF) also used for the prediction of potent inhibitors against HIV-1 protease. It indicates the uses of best descriptors like RDF010u, RDF010m, F04[C-N] which play an essential role in the enzyme binding (Ravichandran et al. 2010). The CoMFA studies with 3D-QASR were used to predict the series of compounds and have been proven as a valuable method for constructing the predictive model and both the electrostatic and steric fields were used as descriptors (Baraldi 1999). The molecular descriptors such as group count, logP, solvent accessible surface area, dielectric energy were considered to compare the anticancer activity of the lead molecules (Alam and Khan 2014).

10.12 Future Directions and Conclusion

QSAR is a widely used technique in the drug designing process. Though, the classical QSAR approach has a useful correlation with important congeneric series of molecules. Besides, the 3D QSAR technique has broadly been used in the general yield of a statistically robust model with limitations to describe the potent molecules. Analyzing the QSAR based virtual screening leads to the identification of promising hit molecules. Several QSAR projects are developed through it and do not show a successful model building stage which leads to poor understanding of several interdisciplinary applications and common unawareness of the pest practice in the field (Tropsha 2010; Ban et al. 2017). Also, several studies have been reported by researchers to target their determinations to the malicious statistical cycle, with the leading objective of model validation. In QSAR modeling is highly limited to the best statistical method. Though, recognizing the precise selection of the statistical method and exterior validations are essential for a crucial step in computational based drug discovery approaches to develop the new compounds with desired properties. Therefore, development of novel machine learning algorithms and other data curation techniques has emerged as alternatives to classical methods to avoid the quantity of examined molecules available in the literature. Hence, researchers must think critically and prioritize the potent lead molecules, this study is highly dependent on the overall success of QSAR based virtual screening approaches in drug discovery processes.

Acknowledgements Vivek Srivastava thankfully acknowledge to Department of Biotechnology, Faculty of Engineering and Technology, Rama University Uttar Pradesh, Kanpur, India and also Chandrabose Selvaraj and Sanjeev Kumar Singh thankfully acknowledge to RUSA-Phase 2.0 Policy (TNmulti-Gen), Dept. of Edn, Govt. of India (Grant No: F.24-51/2014-U).

References

Acharya C, Coop A, Polli JE, Mackerell AD Jr (2011) Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Curr Comput Aided Drug Des 7:10–22

Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes. Bioorg Med Chem 14:6686–6694

Akamatsu M (2002) Current state and perspectives of 3D QSAR. Curr Top Med Chem 2:1381–1394

Alam S, Khan F (2014) QSAR and docking studies on xanthone derivatives for anticancer activity targeting DNA topoisomerase II α. Drug Des Dev Ther 8:183–195

Allen BK, Mehta S, Ember SW, Schonbrunn E, Ayad N, Schürer SC (2015) Large-scale computational screening identifies first in class multitarget inhibitor of EGFR kinase and BRD4. Sci Rep 5:16924

Augen J (2002) The evolving role of information technology in the drug discovery process. Drug Discov Today 7:315–323

Bajorath J (2002) Integration of virtual and high-throughput screening. Nat Rev Drug Discov 1:882–894

Bajorath JR (ed) (2011) Chemoinformatics and computational chemical biology. Humana Press, Totowa

Balaban AT (1982) Highly discriminating distance-based topological index. Chem Phys Lett 89:399–404

Ban F, Dalal K, Li H, LeBlanc E, Rennie PS, Cherkasov A (2017) Best practices of computer-aided drug discovery: lessons learned from the development of a preclinical candidate for prostate cancer with a new mechanism of action. J Chem Inf Model 57:1018–1028

Baraldi PG (1999) Comparative molecular field analysis (CoMFA) of a series of selective adenosine receptor A2A antagonists. Drug Dev Res 46:126–133

Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM (2013) Evaluation and optimization of virtual screening workflows with DEKOIS 2.0 – a public library of challenging docking benchmark sets. J Chem Inf Model 53:1447–1462

Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43:4759–4767

Bissantz C, Bernard P, Hibert M, Rognan D (2003) Protein-based virtual screening of chemical databases. II. Are homology models of G-protein coupled receptors suitable targets? Proteins 50:5–25

Blum LC, Reymond JL (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. J Am Chem Soc 131:8732–8733

Bolis G, Di Pace L, Fabrocini FJ (1991) A machine learning approach to computer-aided molecular design. Comput Aided Mol Des 5:617–628

Brignole EA, Bottini SB, Gani R (1986) A strategy for the solvents for liquid extraction of solvents for separation processes. Fluid Phase Equilib 29:125

Brozell SR, Mukherjee S, Balius TE, Roe DR, Case DA, Rizzo RC (2012) Evaluation of DOCK 6 as a pose generation and database enrichment tool. J Comput Aided Mol Des 26:749–773

Catalin B (2014) More effective DPP4 inhibitors as antidiabetics based on sitagliptin applied QSAR and clinical methods. Curr Comput Aided Drug Des 10:237–249(13)

Chang C, Swaan PW (2006) Computational approaches to modeling drug transporters. Eur J Pharm Sci 27:411–424

Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57:4977–5010

Churchwell CJ, Rintoul MD, Martin S, Visco DP, Kotu A, Larson RS, Sillerud LO, Brown DC, Faulon J (2004) The signature molecular descriptor. 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides. J Mol Graph Model 22:263–273

Clark DE, Pickett SD (2000) Computational methods for the prediction of 'drug-likeness'. Drug Discov Today 5:49–58

Clark DE, Firth MA, Murray CW (1996) Molmaker: de novo generation of 3D databases for use in drug design. J Chem Inf Comput Sci 36:137

Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967

Crivori P, Cruciani G, Carrupt PA, Testa B (2000) Predicting blood–brain barrier permeation from three-dimensional molecular structure. J Med Chem 43:2204–2216

Cruciani G, Crivori P, Carrupt PA, Testa B (2000) Molecular interaction fields in drug discovery: recent advances and future perspectives. J Mol Struct THEOCHEM 503:17–30

Crum-Brown AFT (1868) On the connection between chemical constitution and physiological action. Pt 1. On the physiological action of the salts of the ammonium bases, derived from strychnia, Brucia. Thebia, Codeia, Morphia and Nicotia. R Soc Edin 2:151–203

Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaeger EP (2005) Comparison of automated docking programs as virtual screening tools. J Med Chem 48:962–976

de Groot MJ, Ekins S (2002) Pharmacophore modeling of cytochromes P450. Adv Drug Deliv Rev 54:367–383

Dean PM (2005) Computer-aided design of small molecules for chemical genomics. Humana Press Inc., Totowa

Derringer GC, Markham RL (1985) A computer-based methodology for matching polymer structures with required properties. J Appl Polym Sci 30:4609–4617

Deshpande M, Kuramochi M, Karypis J (2002) Frequent substructure-based approaches for classifying chemical compounds. In: Proc of the 8th international conference on knowledge discovery and data mining, Edmonton

Dessalew N, Singh SK (2008) 3D-QSAR CoMFA and CoMSIA study on benzodipyrazoles as cyclin dependent kinase 2 inhibitors. Med Chem 4:313–321

Dessalew N, Bharatam PV, Singh SK (2007) 3D-QSAR CoMFA study on aminothiazole derivatives as cyclin-dependent kinase 2 inhibitors. QSAR Comb Sci 26:85–91

Diller DJ, Li R (2003) Kinases, homology models, and high throughput docking. J Med Chem 46:4638–4647

Douali L, Villemin D, Cherqaoui D (2003) Neural networks: accurate nonlinear QSAR model for HEPT derivatives. J Chem Inf Comput Sci 43:1200–1207

Ekins S, Lage de Siqueira-Neto J, McCall L-I, Sarker M, Yadav M, Ponder EL (2015) Machine learning models and pathway genome data base for Trypanosoma cruzi drug discovery. PLoS Negl Trop Dis 9:e0003878

Empereur-mot C, Guillemain H, Latouche A, Zagury JF, Viallon V, Montes M (2015) Predictiveness curves in virtual screening. J Chem Informatics 7:52

Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375

Estrada E, Uriarte E (2001) Quantitative structure-toxicity relationships using TOPS-MODE. 1. Nitrobenzene toxicity to tetrahymena pyriformis. Environ Res 12:309–324

Faulon JL, Visco DP Jr, Pophale RS (2003) The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. J Chem Inf Comput Sci 43:707–720

Fink T, Bruggesser H, Reymond JL (2005) Virtual exploration of the small-molecule chemical universe below 160 Daltons. Angew Chem Int Ed 44:1504–1508

Flower DR (2002a) Predicting chemical toxicity and fate. CRC Press, Roca Baton

Flower DR (2002b) Drug design: cutting edge approaches. Royal Society of Chemistry, Cambridge

Fourches D, Muratov E, Tropsha A (2015) Curation of chemogenomics data. Nat Chem Biol 11:535–535

Furnival GM, Wilson RW (1974) Regressions by leaps and bounds. Technometrics 16:499–511

Gao H, Williams C, Labute P, Bajorath J (1999) Binary Quantitative structure−activity relationship (QSAR) analysis of estrogen receptor ligands. J Chem Inf Comput Sci 39:164

Gasteiger J (2003) Handbook of chemoinformatics: from data to knowledge. Wiley, New York

Goh GB, Hodas NO, Vishnu A (2017) Deep learning for computational chemistry. J Comput Chem 38:1291–1307

Gohda K, Mori I, Ohta D, Kikuchi T (2000) A CoMFA analysis with conformational propensity: an attempt to analyze the SAR of a set of molecules with different conformational flexibility using a 3D-QSAR method. J Comput Aided Mol Des 14:265–275

Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 17:241–253

Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernandez-Lobato JM, Sanchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 4:268–276

Good AC, Oprea TI (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? J Comput Aided Mol Des 22:169–178

Guner OF (2002) History and evolution of the pharmacophore concept in computer-aided drug design. Curr Top Med Chem 2:1321–1332

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389

Hall LH, Kier LBJ (2000) Chem Inf Comput Sci 30:784–791

Hall DG, Manku S, Wang F (2001) Solution- and solid-phase strategies for the design, synthesis, and screening of libraries based on natural product templates: a comprehensive survey. J Comb Chem 3:125–150

Hammet LP (1935) Some relations between reaction rates and equilibrium constants. Chem Rev 17:125–136

Hasegawa K, Arakawab M, Funatsu K (2000) Rational choice of bioactive conformations through use of conformation analysis and 3-way partial least squares modeling. Chemom Intell Lab Syst 50:253–261

Hecht P (2002) High-throughput screening: beating the odds with informatics-driven chemistry. Curr Drug Discov 10:21–24

Helguera AM, Combes RD, Gonzalez MP, Cordeiro MN (2008) Applications of 2D descriptors in drug design: a DRAGON tale. Curr Top Med Chem 8:1628–1655

Hessler G, Zimmermann M, Matter H, Evers A, Naumann T, Lengauer T, Rarey M (2005) Multiple-ligand-based virtual screening: methods and applications of the MTree approach. J Med Chem 48:6575–6584

Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801

Ibrahim TM, Bauer MR, Boeckler FM (2015) Applying DEKOIS 2.0 in structure-based virtual screening to probe the impact of preparation procedures and score normalization. Aust J Chem 7:21

Irwin JJ (2008) Community benchmarks for virtual screening. J Comput Aided Mol Des 22:193–199

Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748

Kapetanovic IM (2008) Computer-aided drug discovery and development (CADDD): in-silico-chemico-biological approach. Chem Biol Interact 171:165–176

Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 57:225–242

Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3:935

Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146

Koehn FE, Carter GT (2005) The evolving role of natural products in drug discovery. Nat Rev Drug Discov 2005:206–220

Kovatcheva A, Buchbauer G, Golbraikh A, Wolschann P (2003) QSAR modeling of alpha-campholenic derivatives with sandalwood odor. J Chem Inf Comput Sci 43:259–266

Kumar Singh S, Dessalew N, Bharatam PV (2007) 3D-QSAR CoMFA study on oxindole derivatives as cyclin dependent kinase 1 (CDK1) and cyclin dependent kinase 2 (CDK2) inhibitors. Med Chem 3:75–84

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288

Labute PA (2000) Widely applicable set of descriptors. J Mol Graph Model 18:464–477

Lee KW, Briggs JM (2001) Comparative molecular field analysis (CoMFA) study of epothilones-tubulin depolymerization inhibitors: phramacophore developemt using 3D QSAR methods. J Comput Aided Mol Des 15:41–55

Lemmen C, Lengauer TJ (2000) Computational methods for the structural alignment of molecules. Comput Aided Mol Des 14:215–232

Lewis RA (2005) A general method for exploiting QSAR models in lead optimization. J Med Chem 48:1638–1648

Li P, Tian Y, Zhai H, Deng F, Xie M, Zhang X (2013) Study on the activity of non-purine xanthine oxidase inhibitor by 3D-QSAR modeling and molecular docking. J Mol Struct 5:56–65

Liu SS, Liu HL, Yin CS, Wang LSJ (2003) VSMP: a novel variable selection and modeling method based on the prediction. Chem Inf Comput Sci 43:964–969

MacKerell AD Jr (2004) Empirical force fields for biological macromolecules: overview and issues. J Comput Chem 25:1584–1604

Makino S, Ewing TJA, Kuntz ID (1999) DREAM++: flexible docking program for virtual combinatorial libraries. J Comput Aided Mol Des 13:513–532

Mandel J (1982) Use of the singular value decomposition in regression-analysis. Am Stat 36:15–24

Matter H, Baringhaus KH, Naumann T, Klabunde T, Pirard B (2001) Computational approaches towards the rational design of drug-like compound libraries. Comb Chem High Scr 4:453–475

McGovern SL, Shoichet BK (2003) Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 46:2895–2907

Mitchell JBO (2004) Machine learning methods in chemoinformatics. Wiley Interdiscip Rev Comput Mol Sci 4:468–481

Moustakas DT, Lang PT, Pegg S, Pettersen E, Kuntz ID, Brooijmans N, Rizzo RC (2006) Development and validation of a modular, extensible docking program: DOCK 5. J Comput Aided Mol Des 20:601–619

Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 55 (2012):6582–6594

Nantasenamat C, Monnor T, Worachartcheewan A, Mandi P, Isarankura-Na-Ayudhya C, Prachayasittikul V (2014) Predictive QSAR modeling of aldose reductase inhibitors using Monte Carlo feature selection. Eur J Med Chem 76:352–359

Netzeva TI, Gallegos SA, Worth AP (2006) Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory. Environ Toxicol Chem 25:1223–1230

Neves MA, Totrov M, Abagyan R (2012) Docking and scoring with ICM: the benchmarking results and strategies for improvement. J Comput Aided Mol Des 26:675–686

Nieto-Draghi C, Fayet G, Creton B, Rozanska X, Rotureau P, de Hemptinne JC, Ungerer P, Rousseau B, Adamo C (2015) A general guidebook for the theoretical prediction of physicochemical properties of chemicals for regulatory purposes. Chem Rev 115:13093–13164

Oprea TI, Davis AM, Teague SJ, Leeson PD (2001) Is there a difference between leads and drugs? A historical perspective. J Chem Inf Comput Sci 41:1308–1315

Panwar U, Singh SK (2020) Atom-based 3D-QSAR, molecular docking, DFT, and simulation studies of acylhydrazone, hydrazine, and diazene derivatives as IN-LEDGF/p75 inhibitors. Struct Chem 2020:1–16

Pastor M, Cruciani G, McLay I, Pickett S, Clementi S (2000) GRid-INdependent Descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. J Med Chem 43:3233–3243

Prado-Prado FJ, García I, García-Mera X, González-Díaz H (2011) Entropy multi-target QSAR model for prediction of antiviral drug complex networks. Chemom Intell Lab Syst 107:227–233

Prasanna S, Doerksen RJ (2009) Topological polar surface area: a useful descriptor in 2D-QSAR. Curr Med Chem 16:21–41

Pretel EJ, López PA, Bottini SB, Brignole EA (1994) Computer-aided molecular design of solvents for separation processes. AICHE J 40:1349–1360

Proudfoot JR (2002) Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. Bioorg Med Chem Lett 12:1647–1650

Prusis P, Dambrova M, Andrianov V, Rozhkov E, Semenikhina V, Piskunova I, Ongwae E, Lundstedt T, Kalvinsh I, Wikberg JES (2004) Synthesis and quantitative structure−activity relationship of hydrazones of N-amino-N‘-hydroxyguanidine as electron acceptors for xanthine oxidase. J Med Chem 47:3105–3110

Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489

Ravichandran V, Shalini S, Sundram KM, Dhanaraj SA (2010) QSAR study of substituted 1, 3, 4-oxadiazole naphthyridines as HIV-1 integrase inhibitors. Eur J Med Chem 45:2791–2797

Reddy AS, Pati SP, Kumar PP, Pradeep HN, Sastry GN (2007) Virtual screening in drug discovery - a computational perspective. Curr Protein Pept Sci 8:329–351

Reddy KK, Singh SK, Dessalew N, Tripathi SK, Selvaraj C (2012) Pharmacophore modelling and atom-based 3D-QSAR studies on N-methyl pyrimidones as HIV-1 integrase inhibitors. J Enzyme Inhib Med Chem 27:339–347

Reddy KK, Singh SK, Tripathi SK, Selvaraj C (2013a) Identification of potential HIV-1 integrase strand transfer inhibitors: in silico virtual screening and QM/MM docking studies. SAR QSAR Environ Res 24:581–595

Reddy KK, Singh SK, Tripathi SK, Selvaraj C, Suryanarayanan V (2013b) Shape and pharmacophore-based virtual screening to identify potential cytochrome P450 sterol 14-α-demethylase inhibitors. J Recept Signal Transduction 33:234–243

Repasky MP, Murphy RB, Banks JL, Greenwood JR, Tubert-Brohman I, Bhat S, Friesner RA (2012) Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. J Comput Aided Mol Des 26:787–799

Rester U (2006) Dock around the clock - current status of small molecule docking and scoring. QSAR Comb Sci 25:605–615

Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSAR model. J Chem Inf Model 52:396–408

Roy K, Kar S, Das RN (2015) Background of QSAR and historical developments. In: Das KRKN (ed) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic, Boston, pp 1–46

Ruddigkeit L, van Deursen R, Blum LC, Reymond JL (2012) Visualization and virtual screening of the chemical universe database GDB-17. J Chem Inf Model 52:2864

Ruggeri C, Drinkwater N, Sivaraman KK, Bamert RS, McGowan S, Paiardini A (2011) Identification and validation of a potent dual inhibitor of the P. falciparum M1 and M17 aminopeptidases using virtual screening. PLoS ONE 10:e0138957

Rusinko A III, Young SS, Drewry DH, Gerritz SW (2002) Optimization of focused chemical libraries using recursive partitioning. Comb Chem High Scr 5:125–133

Saliner AG, Netzeva TI, Worth AP (2006) Prediction of estrogenicity: validation of a classification model. Environ Res 17:195–223

Santos-Filho OA, Hopfinger AJ (2001) A search for sources of drug resistance by the 4D-QSAR analysis of a set of antimalarial dihydrofolate reductase inhibitors. J Comput Aided Mol Des 15:1–12

Schneider G, Baringhaus KH (2013) De novo design: from models to molecules. In: De novo molecular design. Wiley, Weinheim, pp 1–55

Schneider G, Fechner U (2005) Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov 4:649

Selvaraj C, Singh P, Singh SK (2014) Molecular insights on analogs of HIV PR inhibitors toward HTLV-1 PR through QM/MM interactions and molecular dynamics studies: comparative structure analysis of wild and mutant HTLV-1 PR. J Mol Recognit 27:696–706

Sharma S, Ravichandran V, Jain PK, Mourya VK, Agrawal RK (2008) Prediction of caspase-3 inhibitory activity of 1,3-dioxo-4-methyl-2,3- dihydro-1h-pyrrolo[3,4-c] quinolines: QSAR study. J Enzyme Inhib Med Chem 23:424–431

Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A (2002) Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. J Med Chem 45:2811–2823

Shen M, Xiao Y, Golbraikh A, Gombar VK, Tropsha A (2003) Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. J Med Chem 46:3013–3020

Singh SK, Dessalew N, Bharatam PV (2006) 3D-QSAR CoMFA study on indenopyrazole derivatives as cyclin dependent kinase 4 (CDK4) and cyclin dependent kinase 2 (CDK2) inhibitors. Eur J Med Chem 41:1310–1319

Skvortsova MI, Fedyaev KS, Palyulin VA, Zefirov N (2001) Inverse structure-property relationship problem for the case of a correlation equation containing the Hosoya index. Dokl Chem 379:191–195

Southan C, Várkonyi P, Muresan S (2009) Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. J Chem 1:10

Speck-Planche A, Cordeiro MNDS (2017) Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. Mol Divers 21:511–523

Spitzer R, Jain AN (2012) Surflex-dock: docking benchmarks and real-world application. J Comput Aided Mol Des 26:687–699

Stanton DT, Egolf LM, Jurs PC, Hicks MG (1992) Computer assisted prediction of normal boiling points of pyrans and pyrroles. J Chem Inf Comput Sci 32:306–316

Stumpfe D, Bajorath J (2011) Applied virtual screening: strategies, recommendations, and caveats. In: Sotriffer C (ed) Virtual screening: principles, challenges, and practical guidelines. Wiley, Weinheim, pp 291–318

Suryanarayanan V, Kumar Singh S, Kumar Tripathi S, Selvaraj C, Konda Reddy K, Karthiga A (2013) A three-dimensional chemical phase pharmacophore mapping, QSAR modelling and electronic feature analysis of benzofuran salicylic acid derivatives as LYP inhibitors. Environ Res 24:1025–1040

Tang H, Yang L, Li J, Chen J (2016) Molecular modelling studies of 3,5- dipyridyl-1,2,4-triazole derivatives as xanthine oxidoreductase inhibitors using 3D-QSAR, TopomerCoMFA, molecular docking and molecular dynamic simulations. J Taiwan Inst Chem Eng 68:64–73

Tong W, Lowis DR, Perkins R, Chen Y, Welsh WJ, Goddette DW, Heritage TW, Sleehan DM (1998) Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. J Chem Inf Comput Sci 38:669

Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R (2004) Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. Environ Health Perspect 112:1249–1254

Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. J Med Chem 48:2534–2547

Tripos (2007) SYBYL8.0. In: Discovery software for computational chemistry and molecular modeling. St. Louis, Missouri, USA

Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29:476–488

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. Quant Struct Act Relat Comb Sci 22:69–77

Van Drie JH (2003) Pharmacophore discovery: lessons learned. Curr Pharm Des 9:1649–1664

Varnek A, Baskin II (2011a) Chemoinformatics as a theoretical chemistry discipline. Mol Inf 30:20–32

Varnek A, Baskin II (2011b) Chemoinformatics as a theoretical chemistry discipline. Mol Informatics 30:20–32

Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK (2011) Validation of QSAR models - strategies and importance. Int J Drug Des Discov 2:511–519

Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD, Watson P (2004) Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 44:793–806

Verma J, Khedkar VM, Coutinho EC (2010) 3D-QSAR in drug design- a review. Curr Top Med Chem 10:95–115

Vijaya Prabhu S, Singh SK (2018) Atom-based 3D-QSAR, induced fit docking, and molecular dynamics simulations study of thieno [2, 3-b] pyridines negative allosteric modulators of mGluR5. J Recept Signal Transduction 38:225–239

Vogel SM, Bauer MR, Boeckler FM (2011) DEKOIS: demanding evaluation kits for objective in silico screening—a versatile tool for benchmarking docking programs and scoring functions. J Chem Inf Model 51:2650–2665

Wedebye EB, Dybdahl M, Nikolov NG, Jonsdottir SO, Niemela JR (2015) QSAR screening of 70, 983 REACH substance for genotoxic carcinogenicity, mutagenicity and development toxicity in the Chem Screen project. Reprod Toxicol 55:64–72

Wegner JK, Frö hlich H, Zell AJ (2004) Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). Chem Inf Comput Sci 44:921

Weis DC, Faulon JL, LeBorne RC, Visco DP (2005) The signature molecular descriptor. 5. The design of hydrofluoroether foam blowing agents using inverse-QSAR. Ind Eng Chem Res 44:8883–8891

Williams AJ, Ekins S (2011) A quality alert and call for improved curation of public chemistry databases. Drug Discov Today 16(2011):747–750

Xu J, Hagler A (2002) Chemoinformatics and drug discovery. Molecules 7:566–600

Xu J, Stevenson J (2000) Drug-like index: a new approach to measure drug-like compounds and their diversity. J Chem Inf Comput Sci 40:1177–1187

Xu L, Zhang WJ (2001) Comparison of different methods for variable selection. Anal Chim Acta 446:475–481

Xue L, Bajorath J (2000) Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Comb Chem 3:363–372

Yang SP, Song ST, Tang ZM, Song HF (2003) Optimization of antisense drug against conservative local motif in simulant secondary structures of HER-2 mRNA and QSAR analysis. Acta Pharmacol Sin 24:897–902

Yasuo K, Yamaotsu N, Gouda H, Tsujishita H, Hirono S (2009) Structure-based CoMFA as a predictive model - CYP2C9 inhibitors as a test case. J Chem Inf Model 49:853–864

Young D, Martin T, Venkatapathy R, Harten P (2008) Are the chemical structures in your QSAR correct? QSAR Comb Sci 27:1337–1345

Zhang S, Golbraikh A, Tropsha A (2006) Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. J Med Chem 49:2713–2724

Zhang S, Wei L, Bastow K, Zheng W, Brossi A, Lee KH, Tropsha A (2007) Antitumor Agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. J Comput Aided Mol Des 21:97–112

Computational Genomics 11

Fabrício Almeida Araújo, Ana Lidia Queiroz Cavalcante,
Marcus de Barros Braga, Rodrigo Bentes Kato,
Rommel Thiago Jucá Ramos, and
Edian Franklin Franco De Los Santos

Abstract

The use of computational methodologies for analysis of biological data is not recent; however, with the reduction of the cost of DNA sequencing associated with the increase in the volume of genomic data produced by the sequencing platforms, it has become essential to use computational approaches to handle and extract more information from the data of complete genomes and/or transcriptomes using bioinformatics tools. The challenge for this starts with simple sequence alignments, until the assembly of the whole genomes with the challenge to process the high volume of data, which requires high computational capacity and/or improvement of the algorithms in order to optimize the use of computers. This chapter will show how DNA sequences are decoded, how sequences are compared through alignment, what are the main approaches to assembly genomes, and how to evaluate their quality followed by gene prediction techniques, and finally, how interaction networks can be implemented from genomic data after processed by the steps presented here.

F. A. Araújo · A. L. Q. Cavalcante · R. T. J. Ramos (✉)
Federal University of Pará, Belém, PA, Brazil

M. de Barros Braga
Federal Rural University of the Amazon, Paragominas, Pará, Brazil

R. B. Kato
Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

E. F. F. D. L. Santos
Federal University of Pará, Belém, PA, Brazil

Instituto de Innovación en Biotecnología e Industria (IIBI), Instituto Tecnologico de Santo Domingo (INTEC), Santo Domingo, Dominican Republic

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_11

Keywords

Computational genomics · Transcriptomes · Genomes · Optimization · Sequences

11.1 DNA Decoding

11.1.1 Sequencing Platforms

Decoding DNA in biological samples has become an essential step in a variety of research applications. With the advancement of technologies, DNA was sequenced more quickly and identified with greater precision in terms of genetic composition and organization, which is fundamental information for understanding biological processes, in addition to directing post-genomic studies such as transcriptomics (Toledo-Arana and Solano 2010).

The DNA structure, described by James Watson and Francis Crick in 1953, made it possible for other researchers to apply sequencing methodologies to determine the nucleotide sequence of nucleic acids, called first generation sequencing technologies (Holley et al. 1965). In this first generation, the initial efforts were focused on RNA sequencing, generating, in 1965 by Robert Holley et al., the first complete nucleic acid sequence, the tRNA alanine from *Saccharomyces cerevisiae (*Heather and Chain 2016*)*. At that time, several researchers started adapting their methods to sequence DNA.

From the mid-1970s, DNA sequencing was leveraged through the Maxam–Gilbert Method (Maxam and Gilbert 1977) and the "Plus and minus" Method (Sanger and Coulson 1996). However, the breakthrough that changed the process of how DNA was sequenced only came in 1977, when Sanger's dideoxy "chain termination" technique was published (Sanger et al. 1977).

The Sanger method has undergone numerous changes, such as the development of semi-automatic sequencers with electrophoresis in capillaries filled with gel, and a detection system using confocal fluorescence excited by laser, which brought advantages such as: reducing the handling of toxic chemicals, application of samples by electroinjection and simultaneous electrophoresis with up to 384 independent capillaries, for the generation of fragments of approximately 750 base-pairs (Ansorge et al. 1987; Ansorge et al. 1986; Kambara et al. 1988; Luckey et al. 1990; Prober et al. 1987; Smith et al. 1985; Swerdlow and Gesteland 1990). In 1986, Leroy Hood's laboratory in Caltech (California—USA), together with Applied Biosystems, launched the first semi-automatic sequencer, based on the Sanger method (Smith et al. 1986).

Simultaneously, with Sanger's sequencing efforts, a luminescent method was developed to measure pyrophosphate synthesis. Pyrosequencing, as this technique was called, was subsequently licensed to 454 Life Sciences, a biotechnology company founded by Jonathan Rothberg. This company evolved into the first major commercial "next-generation sequencing" (NGS) technology (Holley et al. 1965). In 2004, the first high-throughput (HTS) sequencing machine that was massively available to the public was the 454 GS/20 Roche machine, which offered an increase in the number of reads (up to 100 bp), as well as improved data quality (Voelkerding et al. 2009). The greater number of reactions, generated in parallel sequencing on a micrometer scale, often made it possible as a result of improvements in microfabrication and high-resolution images, was the point that defined the second generation of DNA sequencing (Shendure and Ji 2008).

After the 454 machine, other sequencing platforms were developed, such as Solexa, which was later acquired by Illumina in 2005 (Voelkerding et al. 2009) and SOLiD of the Applied Biosystems system (Thermo Fisher Scientific) in 2006, all based on fluorescence detection and characterized by low cost and high sequencing coverage (McKernan et al. 2009). In addition, another platform developed in 2010, but not based on fluorescence or luminescence, was Ion Torrent. This technology incorporates nucleotides due to the pH difference that is generated by the release of H+ ions during polymerization (Rothberg et al. 2011). Most of these new technologies allow the construction of paired genomic libraries, being useful in the resolution of repetitive regions in the genome during the new assembly process. All of these factors contributed to the increase in projects for sequencing complete genomes (Mardis 2011; Scholz et al. 2012).

In 2009, a new way of sequencing single DNA molecules in real time was established by Pacific Biosciences, culminating in the launch of a sequencer known as SMRT (single-molecule real time). This process takes place through a single DNA polymerase molecule fixed at the bottom of a ZMW (zero-mode waveguide detector) detector with the size of a few nanometers, made of a metal film and deposited on a glass substrate (Van Dijk et al. 2014), it can generate long readings of 10 kb (Schadt et al. 2010). However, there is a great perspective regarding the technology of single molecules related to the sequencing in nanopores. This approach was established for the first time even before the second-generation sequencing appeared (Holley et al. 1965).

In early 2012, the first nanopores sequencing platform was announced by Oxford Nanopore, introducing two main versions of sequencers: GridION and MinION, capable of generating large amounts of data, with a simple sample preparation resulting in long reads to a low cost. MinION is a small, portable device, capable of sequencing 30 Gb of DNA, while GridION can generate up to 150 Gb of data transmitted in real time for immediate analysis. Another technology launched by the same company was PromethION, which can generate up to 8 Tb of data (van Dijk et al. 2018).

11.1.2 Whole-Genome Sequencing and Whole-Transcriptome Sequencing

The improvement of NGS technologies made it possible to carry out genome sequencing and complete transcriptome projects on a large scale. The analysis of genomes and complete transcriptome enabled the identification of gene function within the biological context. For example, with the sequencing of a genome, it was possible the identification of genes that generally do not function independently, and their functions are not controlled directly by the promoter, but by many other regulatory elements, such as intensifiers, response elements, and silencers (Heather and Chain 2016; Holley et al. 1965). While the transcriptome sequencing allows quantifying the heterogeneity of gene expression from cells to tissues and organs, this method is important as it offers the initial steps for the functional annotation and characterization of genes and genomes that were previously revealed by DNA sequencing (Altman and Raychaudhuri 2001); assembly projects for the rebuilding of genetic interaction networks to comprehend cellular functions, growth and development, and biological systems (Hsiao et al. 2000); produces molecular fingerprints of disease development and prognoses to identify potential targets for diagnosis and drug development (Celis et al. 2000), and it also makes it possible to study the interaction between the host and the pathogen by the development of new strategies that can be used for therapeutic and prophylactic intervention (Manger and Relman 2000). Thus, the analysis of the complete transcriptome provides a basis for exploring the regulatory pathways and genetic networks that both qualitatively and quantitatively control the phenotypes important for agriculture and human medicine (Jiang et al. 2015).

11.2 Sequence Alignment

The sequence of bases in DNA has a huge importance, as it contains the code for the formation of several proteins and, therefore, contains the complexity and diversity of life itself (Martorell-Marugán et al. 2019). The unique order of these bases in DNA creates the basic hereditary units, which are the genes. The human genome project initially estimated that there would be 20,000 genes in the human genome (Lander et al. 2001; Venter et al. 2001), and these estimates were later revised to 25,000–30,000 genes (Pennisi 2003). Based on the DNA sequence, enzymes such as RNA polymerase create single-stranded messenger RNA (mRNA) that later translates into proteins. This entire process of decoding the DNA sequence in a protein is referred to as the "central dogma of life" (Crick 1970). Depending on the organism, genes may not encode proteins that, being composed of amino acids, are much more complex than nucleic acids. There are 20 main amino acids that form proteins, and each protein can group them in different numbers and order. This amino acid sequence of proteins is also crucial, as it not only determines the physical-chemical properties of proteins, but also determines the different conformations that they can create in a three-dimensional space (Anfinsen 1973). These changes result in complex protein structures that, in turn, perform unique biological functions, such as transport, functional regulation, and homeostasis. Therefore, it is of great importance to identify the correct sequence of nucleotides in DNA/RNA and of amino acids in proteins.

The comparison of biological sequences allows us to confront the differences between organisms and species at the gene level. Comparative genomics, a branch of science that exhaustively uses bioinformatics techniques to track genes in various species and study their similarities and differences, uses these studies to infer the functional and structural characteristics of newly discovered or existing proteins. The analysis of biological sequences does not differ much from the techniques used to compare strings and texts and, therefore, the concept of alignment becomes very important. Sequences that evolve in species and clades through mutations include insertions, deletions (indels), and incompatibilities. When comparing two biological sequences, an alignment is generated to visualize the differences between the sequences at each position (Martorell-Marugán et al. 2019).

Sequence alignment is one of the main tasks of bioinformatics. It consists of aligning a query sequence with a reference sequence, which is usually in a public database of sequences, with the aim of determining whether they have correspondences with each other that are statistically significant (Gusfield 1997). It differs from the classic computational problem of exact string matching (Cormen et al. 2001), where there is an interest in finding exact matches. String alignment is an approximate string match or string match problem that allows for errors (Navarro 2001). The problem, in its most general form, is to find into a text (or sequence of characters) the position where a certain pattern occurs, allowing a limited number of errors in the correspondences. The distance between the two sequences is defined as the minimum sequence of operations necessary to transform one into the other. With respect to probability, a cost is assigned to operations, so that the most likely operations cost less. The objective is to minimize the total cost (Li and Homer 2010). Ultimately, the final goal of sequence alignment is to determine the similarity between parts of the genomic code. Among the known applications of this type of task, we can mention the discovery of genes, prediction of function, and assembly of the genome sequence.

11.2.1 Biological Sequence Alignment

An alignment between two strings is simply the matching of pairs between the letters in each string. The alignment of nucleotide or amino acid sequences is able to reflect the evolutionary relationship between two or more homologous sequences that share a common ancestor. If the same letter is present in both sequences, the position was preserved in the evolution. If the letters are different, then it is possible to infer that the two strings are derived from an ancestral letter (which may be one of the two or none) (Koonin and Galperin 2013). However, sequences that are homologous can have different lengths, which can be partly explained by insertions or deletions in the sequences. In this way, a letter or a section of letters can be paired with dashes in the other sequence to signify this insertion or exclusion (Fig. 11.1).

```
A T T G C A T C   A A G C T A T A   A A T T G C A A
A T G - A C - C   A A - G - A T A   A A - - C C A A
```

Fig. 11.1 Possible short sequence alignments

11.2.1.1 Pairwise Sequence Alignment and Dynamic Programming

Pairwise alignment consists of comparing two sequences with one another to find the best possible alignment between them. The process involves a scoring system for each position where there is a match, mismatch, and indels. Since matches are preferred over deletions, they normally receive the highest scores and the lowest scores are assigned to insertions. The similarity between two sequences is inversely proportional to the number of mismatches and indels in the alignment. Different scoring models were developed based on the statistically relevant frequency of one amino acid becoming another.

There are two types of alignments for sequence analysis in pairs based on the dynamic programming method: Global and Local Alignment.

Global Alignment

Also called end-to-end alignment. The idea behind the method is to try to align all the residues in each sequence. This approach is useful when the sequences being compared are similar and of approximate size. Needleman and Wunsch were the first to present an algorithm capable of finding the global alignment between two amino acid sequences. The algorithm is based on dynamic programming and achieves the global alignment of two sequences (Needleman and Wunsch 1970). The algorithm covers three main steps: initialization, calculation, and trace back. A matrix of dimensions *i, j* is initialized, where *i* and *j* are the length of the two strings in comparison. Next, the highest score *F*(*i,j*) for each comparison in each position is calculated,

$$F(i,j) = \max \{F(i-1, j-1) + s(Xi, Yi), F(i-1,j) - d, F(i, j-1) - d\}, \tag{11.1}$$

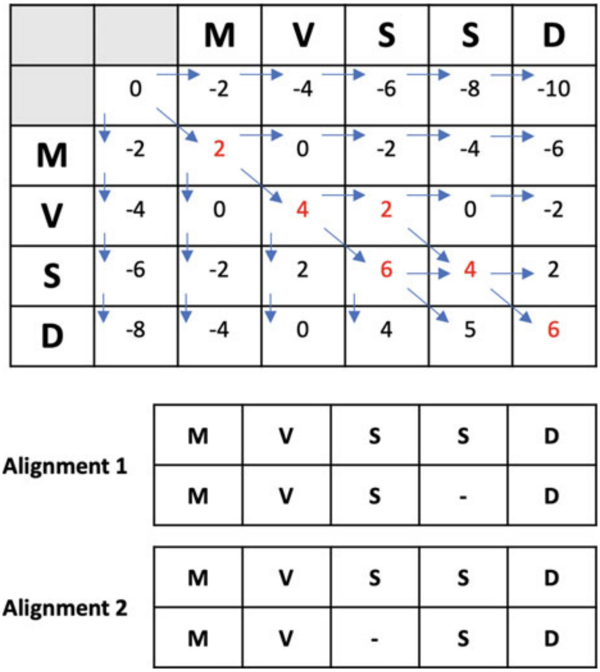
where *s*(*Xi, Yi*) is the match/mismatch score and *d* is the penalty for deletion.

After calculating the maximum score for each position in the matrix (Fig. 11.2), the trace back starts from the last cell (bottom right) in the matrix. At each step, it moves from the current cell to the one from which the current cell value was derived. A match or mismatch is assigned if the maximum score was derived from a diagonal cell. An insertion/deletion is assigned if the score was derived from the top or left cell. After the trace back is complete, there are two sequences aligned end to end with an optimal alignment score (Durbin et al. 1998).

Local Alignment

This type of arrangement is most useful for different sequences that probably contain regions of similarity in the larger context of the sequence. Smith and Waterman (Smith et al. 1981) introduced a different algorithm for scoring similarities in order to find optimal local alignment subsequences, even at the cost of the global score. The algorithm achieves local alignment of strings and is quite similar to the Needleman–Wunsch's method. Local alignment can be used in situations where you want to align smaller substrings from two sequences. In the biological context,

Fig. 11.2 Needleman–Wunsch matrix. The calculation uses scores: +2 for match, −1 for mismatch, and −2 for gap. The arrows show the matrix cell from where the value is generated. Cells with values in red show the trace back that creates the alignment



		M	V	S	S	D
	0	-2	-4	-6	-8	-10
M	-2	2	0	-2	-4	-6
V	-4	0	4	2	0	-2
S	-6	-2	2	6	4	2
D	-8	-4	0	4	5	6

Alignment 1

M	V	S	S	D
M	V	S	-	D

Alignment 2

M	V	S	S	D
M	V	-	S	D

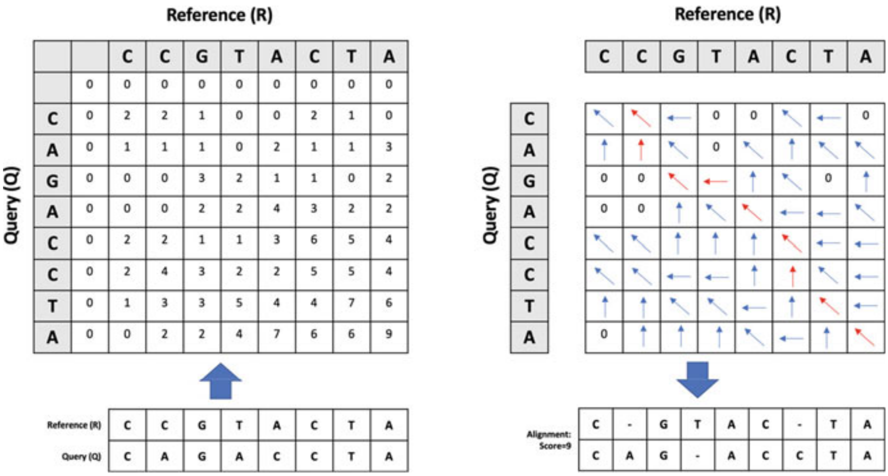
such a situation can arise during the search for a domain or motif within larger sequences. The algorithm comprises the same steps as Needleman–Wunsch, however, with two main differences. The calculation of the maximum score also includes an option of *0*:

$$F(i,j) = \max\{0, F(i-1, j-1) + s(Xi, Yi), F(i-1,j) - d, F(i, j-1) - d\}. \tag{11.2}$$

The assignment of *0* as the maximum score corresponds to the beginning of a new alignment. This allows the alignments to end anywhere in the matrix. The trace back, therefore, starts from the highest value of $F(i, j)$ in the matrix and ends where it finds *0* (Fig. 11.3).

11.2.1.2 Multiple Sequence Alignment (MSA)

When it comes to biological sequence analysis, one of the biggest challenges is to decode the large number and length of the sequences. Biological databases store a vast amount of proteins and DNA sequences and gather more than 100 million sequences, of the most distinct species of nature. Although alignment methods based on dynamic programming are quite accurate and can achieve good alignments based on scores, they are slow and impractical for these databases with millions of sequences. The time complexity of dynamic programming algorithms is O(mn), that is, the product of the sequence lengths. As an initial way of trying to improve the



Reference (R)

Query (Q)

		C	C	G	T	A	C	T	A
	0	0	0	0	0	0	0	0	0
C	0	2	2	1	0	0	2	1	0
A	0	1	1	1	0	2	1	1	3
G	0	0	0	3	2	1	1	0	2
A	0	0	0	2	2	4	3	2	2
C	0	2	2	1	1	3	6	5	4
C	0	2	4	3	2	2	5	5	4
T	0	1	3	3	5	4	4	7	6
A	0	0	2	2	4	7	6	6	9

Reference (R)	C	C	G	T	A	C	T	A
Query (Q)	C	A	G	A	C	C	T	A

Reference (R)

Query (Q)

	C	C	G	T	A	C	T	A
C	↖	↖	←	0	0	↖	←	0
A	↑	↑	↖	0	↖	↑	↖	↖
G	0	0	↖	←	↑	↖	0	↑
A	0	0	↑	↖	↖	←	←	↖
C	↖	↖	↑	↑	↑	↖	←	←
C	↖	↖	←	←	↑	↑	↖	←
T	↑	↑	↖	↖	←	↑	↖	←
A	0	↑	↑	↑	↖	←	↑	↖

Alignment: Score=9	C	-	G	T	A	C	-	T	A
	C	A	G	-	A	C	C	T	A

Fig. 11.3 Smith–Waterman matrix with linear gap penalty. The calculation uses scores: +2 for, −1 for mismatch, and 1 for gap. The left matrix represents input sequences and the right matrix represents sequences are alignments. The left matrix is the corresponding (n + 1) by (m + 1) score matrix. The right matrix is the trace back matrix, with red arrows indicating the optimal alignment path. The null pointer is represented as 0

speed of comparison between sequences, heuristic algorithms such as BLAST (Altschul et al. 1990), BLAT (Kent 2002), and FASTA (Lipman and Pearson 1985; Pearson and Lipman 1988) were created. In the same direction, algorithms such as LSCluster (Husi et al. 2013), Usearch (Edgar 2010), Vsearch (Rognes et al. 2016), Diamond (Buchfink et al. 2015), and Ghostx (Suzuki et al. 2014) have been proposed to try to improve the search efficiency by similarity. In general, these algorithms look for exact matches and extend the alignment of those matches, trying to estimate the ideal score alignment. Thus, heuristic algorithms with approximate correspondence approaches try to solve the multiple sequence alignment by finding similarities between them, as is the case of the CLUSTAL software family (Higgins et al. 1992; Higgins and Sharp 1988; Thompson et al. 1994), which uses the progressive algorithm of Feng and Doolittle (Feng and Doolittle 1987).

BLAST (Basic Local Alignment Search Tool) is a software based on the idea that the best scoring sequence alignment should contain the largest number of identical matches or high-scoring sub-alignments. The algorithm works by performing the following steps:

1. Reducing the query sequence into small subsequences called seeds;
2. Searching for these seeds across the entire database looking for exact matches;
3. Extending the size of the exact matches into an un-gapped alignment until a maximum scoring extension is reached.

The use of seeds to first search for exact matches greatly increases the entire search process and alignment without gaps loses only a small set of significant matches. BLAST's accuracy and sensitivity have made it one of the most widely used search algorithms in the biological world (Martorell-Marugán et al. 2019). A variant of BLAST called Position-Specific-Iterative BLAST (PSI-BLAST) extends the basic BLAST algorithm. PSI-BLAST (Altschul et al. 1997) performs several BLAST iterations and uses the hits found in one iteration as a query for the next iteration. Although PSI-BLAST responds slower to the large amount of calculations required, it is considered a reliable tool to finding distant homology relationships (Martorell-Marugán et al. 2019).

Although BLAST and PSI-BLAST are still widely used, some lately developed methods offer results with greater precision and sensitivity. Hidden Markov Models (HMM) have been used efficiently in numerous applications to understand and explore biological data. An example is HMM-HMM (HHblits) fast sequence search (Remmert et al. 2012). The tool can be used as an alternative to BLAST and PSI-BLAST and is 50–100 times more sensitive. This high sensitivity of the tool can be attributed to the algorithm that is based on the comparison of the HMM representations of the sequences. Although profile–profile or HMM–HMM alignments are very slow due to calculations, the HHblits prefilter reduces the required alignment scaling from millions to thousands, increasing its speed considerably. HHblits represents each sequence in the database as an HMM profile. This pre-processing reduces the number of HMM comparisons to search for similarity, selecting only those target sequences where the highest alignment without gap exists. At the end, a Smith–Waterman alignment shows a significant E-value.

There is another set of methods used to perform Multiple Sequence Alignment (MSA), while reducing errors inherent to progressive methods, they are called iteratives. These categories work in a similar way to progressive methods, but they realign the initial sequences repeatedly as well as they add new sequences to the growing MSA. A very used iteration-based algorithm is called MUSCLE (Multiple Sequence Alignment by Log-Expectation) and improves the performance of progressive methods through a more accurate distance measurement to assess the relationship between two sequences (Edgar 2004).

Both pairwise and MSA algorithms use substitution matrices to assign points to the sequence alignments. These matrices evaluate potential substitutions for protein and nucleic acid sequences. Each possible residue substitution receives a score that reflects the probability of change. Two protein substitution matrix models are the best known: Percent Accepted Mutation (PAM) (Dayhoff et al. 1978) and Blocks Substitution Matrix (BLOSUM) (Henikoff and Henikoff 1992).

These above-mentioned methods use traditional approaches to solve the MSA problem (Edgar and Batzoglou 2006). In practice, as the MSA is an NP-hard problem, these methods often fail to do multiple alignments. To overcome this problem, Metaheuristic approaches to the MSA problem were developed. The Metaheuristics' basic concept allows the description at an abstract level and can take advantage of domain-specific knowledge in the heuristics' form, which is, in its turn, controlled by the strategy in a higher-level. Metaheuristics are generally used as a guide to overcome heuristic problems. These methods exploit the search space in an account to find an almost optimal result. Usually, Metaheuristics are categorized in: (Dey et al. 2017):

1. Single-Solution Based Methods: this category includes a local search algorithm that is primarily concerned with modifying and improving individual possible solutions. The unique solution-based method includes Tabu Search (TS) (Naama et al. 2013),
   Simulated Annealing (SA) (Lindgreen et al. 2007), Variable Neighborhood Search (VNS) (Mladenović and Hansen 1997), Iterated local Search (ILS) (Lourenço et al. 2003), but not restricted to these methods only.
2. Population-Based Methods: population-based metaheuristics begin its process with the starting population and keep iterating until any stop criteria are met. These methods are imitations or inspired by natural phenomena. They use bio-inspired operators such as selection, crossing, mutation, to generate the pool of descendants from the previous population. This is the main difference between Metaheuristic single-solution based and population-based methods. Population-based Metaheuristic methods include Evolutionary Algorithms (EAs), Genetic Algorithm (GA) (Ortuño et al. 2013), Differential Evolution (DE) (Maulik and Saha 2009), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Group Search Optimizer (GSO), Artificial Immune System (AIS) (Dasgupta et al. 2011), etc.

As mentioned before, the sequence alignment problem can be categorized into two main types, pairwise alignment and multiple sequence alignment. Each category was designed for different purposes. The pairwise alignment involves just aligning two sequences. However, in Multiple Sequence Alignment, the main objective is to find similarity between more than just two sequences. Aligning two strings is a relatively simple task and does not take a hard computational time. The problem starts to become more complex when the number of strings increases. Dey et al. (2017) proposed a taxonomy (Fig. 11.4) for the different existing approaches to sequence alignment.

11.3 Genome Assembly and Annotation

From the data generated by NGS technologies, several new applications have emerged, such as the study of microbial communities, the discovery of structural variants in genomes, and the analysis of gene structure and expression (Chen et al. 2017). Due to the small length of sequences generated by the most common NGS platforms, many of these analyses begin with the computational process of sequence assembly, which consists of grouping the generated fragments based on their base identity (Nagarajan and Pop 2013).

There are two general approaches to assembling NGS fragments: reference-based and de novo approaches. In the reference-based assembly, a reference genome of the

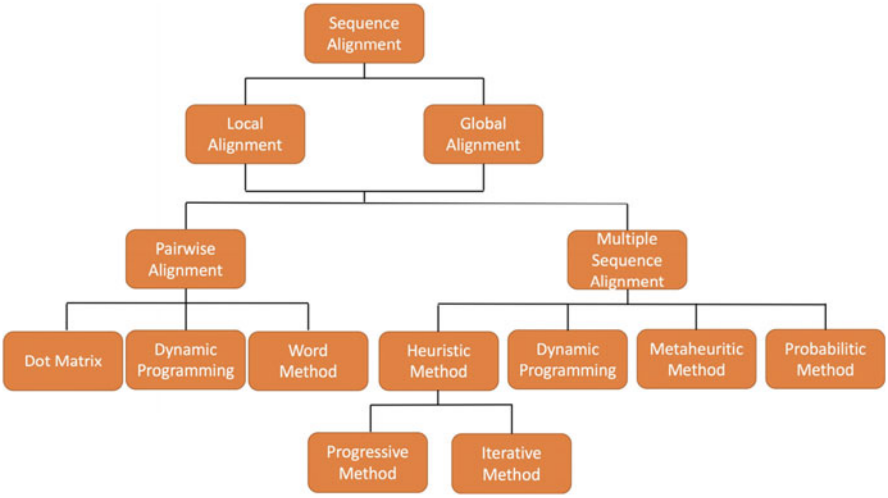


Fig. 11.4 Different methods for sequence alignment

same organism or related species is used as a guide to align the reads; this is, in many cases, the analysis of resequenced data (Pop et al. 2004).

11.3.1 Reference-Based Assembly

The reference-based assembly requires less computational cost when compared to the de novo approach. It is a technique to identify the differences between the reads obtained in a sequencing compared to a previously available reference genome, traditionally used in resequencing, but not limited to this (Hacia 1999).

In the reference-based assembly each read is compared to the reference sequence, base by base, trying to map those bases. In this scenario, there are four possibilities: matches, mismatches, and insertions and deletions (indels). A match happens when one base is mapped against the reference sequence; a mismatch happens when one base is not mapped against the reference genome. An insertion happens when one base is present in the read sequence but not in the reference sequence, while a deletion happens when the base is present in the reference sequence but not in the read sequence. The combination of insertions and deletions is called indels (Hoffmann et al. 2009). In order to evaluate how well the alignment is, the number of matches is counted and then divided by the size of the sequence. This division represents the percentage of identity (Raghava and Barton 2006). These concepts are shown in Fig. 11.5.

Due to the high coverage of sequencing provided by Next-Generation Sequencing platforms, analysis of variants in the genome such as SNPs (Morris and Zeggini 2010) and SNVs (Schnepp et al. 2019) is based on mapping reads against a reference genome, where the alignment achieves a minimum score, usually represented by the

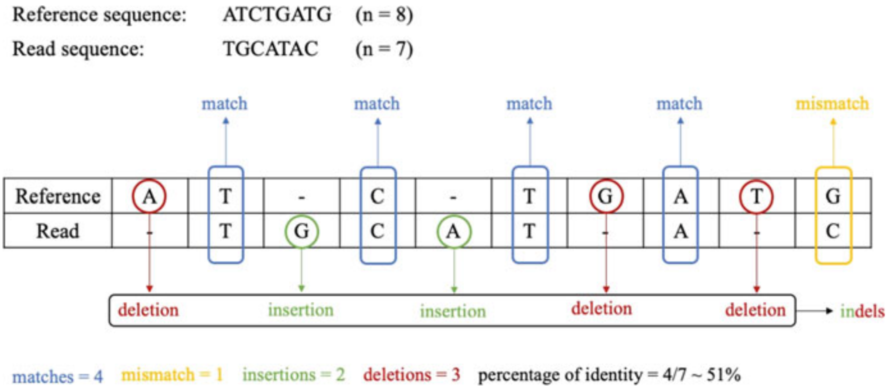


Fig. 11.5 Main alignment concepts: matches, mismatches, insertions, deletions, and percentage of identity

amount of matches, mismatches, indels, and/or percentage of identity (Hoffmann et al. 2009).

11.3.1.1 Mapping Algorithms and Tools

The alignment presented in Fig. 11.5 shows two very simple and small sequences to demonstrate simple concepts. In reality, considering the size of reference genome sequences and the amount of data generated by NGS technologies, some robust methodology is required. Mapping efficiency of reads against the reference sequence, highly accurate, is determinant for the quality of downstream analysis (Keel and Snelling 2018).

Over 50 different mapping algorithms exist (Fonseca et al. 2012). Most of them require special data structures, indices, constructed for reads of sequences and the reference sequence. Based on how these algorithms use their indices, it is possible to group them in two categories: hash tables-based algorithms and Burrows–Wheeler transform (BWT)-based algorithms (Li and Homer 2010).

Hash tables-based algorithms are grouped in types: those that hash the genome and those that hash the reads. The main concept for both types is to construct a hash table for subsequences of reads and genomes. The hash key for each entry in the hash table is a subsequence, and the value for that key is a list of the coordinates where this subsequence can be located (Hatem et al. 2013). Examples of hash tables-based algorithms: GSNAP (Wu and Nacu 2010), FANGS (Misra et al. 2010), and MAQ (Li et al. 2008).

BWT-based algorithms are very efficient in indexing data and maintaining small memory usage when a search is performed. Current BWT-based tools use a modified version of BWT algorithm that uses a different type of data structure, called FM-index, Created by Ferragina and Manzini (2000). The transformation of the genomes into a FM-index improves the search performance, improving the algorithm as a whole. Because of its efficiency, BWT-based algorithms became the most used in mapping applications (Zhang et al. 2013). In this context, two software stand out: BWA (Li et al. 2008) and Bowtie (Langmead et al. 2009).

Bowtie begins by constructing the FM-index for the reference sequence, then it uses a modified version of the Ferragina and Manzini mapping algorithm to locate the position of the alignment. Currently, two versions of bowtie can be found, Bowtie and Bowtie 2 (Langmead and Salzberg 2012). Bowtie 2 was developed mainly to handle reads longer than 50 base-pairs, while the first version of bowtie handles only sequences up to 35 base-pairs.

BWA is very similar to Bowtie; it also uses a modified version of the Ferragina and Manzini (2000) mapping algorithm to find exact matches. To handle inexact matches, BWA searches for matches among subsequences of the reference sequence minding a certain distance defined. In general, Bowtie is best suited for most analyses, while BWA performs better for longer reads (Hatem et al. 2013).

11.3.1.2 Advantages and Disadvantages

New genomes assemblies that can be used as references are available every day. This makes the probability of a related species genome have already been assembled very high. It means that a great part of reads of new sequenced species can be mapped to those already assembled, thus assisting in the process of assembling new species. Also, the computational cost of assembling using a reference is much lower when compared to the de novo approach (Lischer and Shimizu 2017).

However, the reference-based assembly has some advantages. Those advantages rely on biases that can be found in new assemblies towards the reference sequence. Also, diverged regions might not be correctly reconstructed or maybe be missing, leading to the reduction of the diversity of the assembly targeted (Schneeberger et al. 2011). Additionally, chromosomal rearrangements between species and errors in the reference sequence can lead to mistaken assembly (Ekblom and Wolf 2014). The accumulation of all these problems can lead to the increasing divergence between target and reference species (Card et al. 2014).

11.3.2 De Novo Assembly

The de novo assembly approach is based on the overlapping of the reads or part of them with another (Martin and Wang 2011). This strategy is useful to unknown genomes: new strains or species, and is able to represent regions which cannot be identified by reference assembly due to its absence in the reference genome.

To improve the accuracy of the de novo assembly is highly recommended to remove the low quality bases (Phred metric) of the ends of the reads, and that reads with low quality scores to avoid missassemblies. After the launch of next generation sequencing platforms, the challenge became to group short readings (<30 bp) based on its identity to produce long sequences (contigs), and in turn contigs can be ordered and oriented to generate scaffolds (Baker 2012; El-Metwally et al. 2013; Martin and Wang 2011).

One of main difficulties is assembly of repeated regions, greater than reads length, of the genome (El-Metwally et al. 2013) and some bases or regions cannot be represented in the assembled genome, these regions are called gaps and are usually represented by N (Baker 2012).

Actually, the reads length increases and the use of paired libraries is useful to address some repeated regions for the prokaryote, but it remains a real problem during the assembly of eukaryotes due to the larger repeated regions (Nowak et al. 2019).

The main strategies used to assembly genomes are Eulerian De Bruijn Graph, Hamiltonian De Bruijn Graph, String graph, and Overlap-Layout-Consensus (OLC) (Sohn and Nam 2018) which are implemented with some differences by many assemblers, such as ALLPATHS (Butler et al. 2008), Velvet (Zerbino and Birney 2008), ABySS (Simpson et al. 2009), SOAPdenovo (Luo et al. 2012), and SPAdes (Bankevich et al. 2012).

Before executing any approach the user can adopt some strategies to correct errors on the reads, most of them based on the frequency of reads or k-mers to define the confident and erroneous sequences (Sohn and Nam 2018).

The quality of the results obtained from the genome assembly process can be evaluated based on the contig length, amount of bases and contig generated, and how large the sequences produced are, to explore the results the common metrics to be evaluated are: N50—The N50 value means that 50% of the bases generated by the assembly process are part of contigs with length greater than or equal to N50 value; NG50—the same for N50 but the percent of bases in the reference genome selected; L50—the number of contigs used to reach the N50 value (Earl et al. 2011). Other metrics should be evaluated too, such as number of base produced, number of contigs and missassemblies when a reference genome is available to check with software such as Quast (Gurevich et al. 2013).

11.3.3 Hybrid Assembly

New hybrid strategies have been developed to take advantage of each type of assembly. Among them, it is possible to highlight techniques that combine reads and assemblies from different sequencing technologies and different assembly algorithms that can be applied in several tasks, such as de novo assemblies, sequencing error correction, and sequence quality improvement (Hatakeyama et al. 2018). This type of hybrid assembly makes use of reads from different sequencers to reconstruct the genome, mostly using overlap-layout-consensus based methods. Another hybrid assembly approach occurs when different assemblers are used. Rather than performing assembly from reads, this kind of hybrid strategy, also known as meta-assembly, uses assemblies generated by different assemblers, combining the results (contigs and/or scaffolds) produced by those tools to produce a new sequence. However, the concepts of hybrid assembler and hybrid assembly should not be confused. When it comes to assemblers, “hybrid” refers to the ability of an assembler to work with short and long reads, while in relation to the assembly process, “hybrid” refers to the use of more than one type of assembly strategy (DBG/OLC), sequencer (regardless of read type), or input dataset (read/mount) (Miller et al. 2017).

11.3.4 Gene Prediction and Annotation

After finishing the genome assembly process, having a whole genome or a draft, the identification of the Open Read Frames (ORFs), sequence between the start and stop codon, is the next step, followed by gene annotation: the process to get metadata regards the genic product for each ORF identified. The gene annotation often is based on biological database that shows the function, products, and processes that gene can be involved beyond other information. Some methods adopted to do the gene prediction are based on a training dataset, so when this set of genes chosen for training is not good, it can lead to bad results. For the training task, most programs today use Markov models (HMM—Hidden Markov Models or IMM—Interpolated Markov Models) (for example, SNAP (Korf 2004); GlimmerHMM (Majoros et al. 2004); GeneMark (Lukashin and Borodovsky 1998); GlimmerIMM (Salzberg et al. 1998)) for this training, where the genes are modeled with the Markov models that use a series of states to represent a generic structure of the genes. Data training for gene prediction and annotation programs is often chosen at random from a subset of high-quality genes that ideally represent the variation found in a genome. When programs for gene prediction and annotation are trained on a grass genes subset with random GC content, they are effectively being trained on two classes of genes at the same time, and this may result in poor output when genes are predicted in new sequences of genome.

Actually, the sequence of the human genome can be done for less than a thousand dollars. Due to this reduction in the sequencing price, there was an advance in the assembly and alignment algorithms. As a result, obtaining a high-quality assembly draft became an achievable goal for most genome projects. This caused the bottleneck in genomic studies to change focus, genome annotation has become a challenging task due to the difficulty of collecting or predicting proteins, mainly for large genomes, requiring other data sources, such as RNA-Seq and databases to train, optimize, and configure gene annotation tools. (Yandell and Ence 2012).

The manual curation can be used to improve the quality of gene annotation to describe the Gene Ontology (GO) Terms (Consortium 2015) or the gene products based on biological annotation database, such as Blast2GO (Conesa et al. 2005) and GoFeat (Araujo et al. 2018), which use annotated genes and its structural similarity to take new information and insights, ever based on computational approaches representing most of the annotations found on the biological databases.

The accuracy of gene annotation is essential to next analysis to evaluate the genes found and their relationship in the organism, which will drive to discoveries about functions and phenotypes which can be associated to the organism to many applications, such as pathogen–host interactions and antibiotic resistance.

One of the main limitations in the genomic annotation task comes from need of database with annotations already made. There are some areas of biology that are more studied and therefore have more data (complete data, better described, and sometimes curated) for known processes, beyond the amount of databases for specific analysis that are not integrated on the big databases such as Genbank, DDBJ, and EBI. Nowadays with the evolution of annotation programs, most of them are now automated, for example, RAST (Aziz et al. 2008; Seemann 2014), PATRIC (Wattam et al. 2014). These pipelines basically have two tasks: searching for patterns that identify the species gene (e.g., ESTs—Expressed Sequence Tag, proteins, RNA-Seq) and characterizing these patterns into a database (e.g., Interpro (Mitchell et al. 2019), Uniprot (Apweiler et al. 2004), Pfam (El-Gebali et al. 2019)) using Blast (Altschul et al. 1990) or Diamond (Buchfink et al. 2015) (Table 11.1).

11.4 Biological Interaction Network

Biological networks are used in different biological sciences, such as the study of the interactome, cancer study, drug prediction, metagenome analysis, proteomic analysis, molecular interactions, and cell interactions, among other areas.

A biological network can be defined as a collection of units (biomolecules), potentially interacting as a system. In other words, a biological interaction network can be represented an abstraction of the interactions obtained through mathematical or computational models, where a uniform set of nodes connected by a uniform set of edges that can be directed or undirected are represented. In this type of network, the nodes can represent biomolecules (genes, protein, neuron, organisms, cells, among others), and the borders usually represent relationships and interactions (biochemical, transcriptional, energy flow, regulation, co-expression, metabolic, among other) (Beretta et al. 2019a; Proulx et al. 2005).

The study and analysis of networks is part of network biology. This paradigm allows us to understand the complex interactions of biomolecules within cells by representing and analyzing biological systems through tools and methods derived from graph theory, mathematics, physics, statistics, machine learning, and other, applies to and omics and biological data (Pellegrini 2019; Zhang et al. 2014).

The inferences of biological networks using NGS data allow obtaining relevant information about expression and regulation processes inside the organisms. Biological interaction networks can be built using different methods of reverse engineering that use high and low throughput data, as well as statistical, mathematical, and computational techniques that allow reconstructing how the elements of biological networks integrate as a system (Chasman et al. 2016; Tieri et al. 2019).

The power of biological networks lies in the possibility of being able to abstract from complex biological systems in the form of a graph, which allows analyzes and descriptions of these systems, as well as detecting interactions and processes that could not be discovered by studying the elements individually (Marbach et al. 2012; Pellegrini 2019).

Table 11.1 List of the main software tools developed for gene prediction and annotation since 2007

Software	Organism	Year	Type	Method
GLIMMER (Delcher et al. 2007)	Bacteria, archea and viruses	2007	Ab initio	IMM (Interpolated Markov Model)
RAST (Aziz et al. 2008)	Bacteria and archea	2008	Pipeline	Glimmer + PubSEED
Mgene (Schweikert et al. 2009)	Eukaryote	2009	Ab initio	Structural HMM (Hidden Markov Model) combined with discrimination training techniques similiar to SVMs (support vector machine)
Prodigal (Hyatt et al. 2010)	Prokaryote	2010	Ab initio	Dynamic programming + HMM
MAKER2 (Holt and Yandell 2011)	Smaller eukaryotic and prokaryotic	2011	Pipeline combiner	Evidence or ab initio or ab initio evidence driven
MOCAT (Kultima et al. 2012)	Prokaryote and eukaryote	2012	Pipeline	Use prodigal or MetaGeneMark
MetaGUN (Liu et al. 2013)	Smaller eukaryotic	2013	Ab initio	SVM (support vector machine) and prokaryotic
GeneMark-ET (Lomsadze et al. 2014)	Prokaryote	2014	Ab initio	HMM (Hidden Markov Model)
Prokka (Seemann 2014)	Prokaryote	2014	Pipeline	Abinitio + evidence-based for functional annotation
GASS (Wang et al. 2015)	Eukaryote	2015	Comparative	Shortest path model and Dynamic Programming
AugustusCGP (König et al. 2016)	Eukaryote	2016	Comparative	Logistic regression
PGAP (Tatusova et al. 2016)	Prokaryote	2016	Pipeline	GenemarkS + Glimmer + extrinsec data
Funannotate (Palmer and Stajich 2017)	Specifically for fungi, higher eukaryotes	2017	Pipeline	Evidence Modeler + Augustus + GeneMark-ES/ET + evidence + PASA
FunGap (Min et al. 2017)	Fungi	2018	Pipeline	Augustus + Maker + Braker1
Vgas (Zhang et al. 2019)	Vírus	2019		Ab initio + ZCURVEV + BLASTp similarity-based

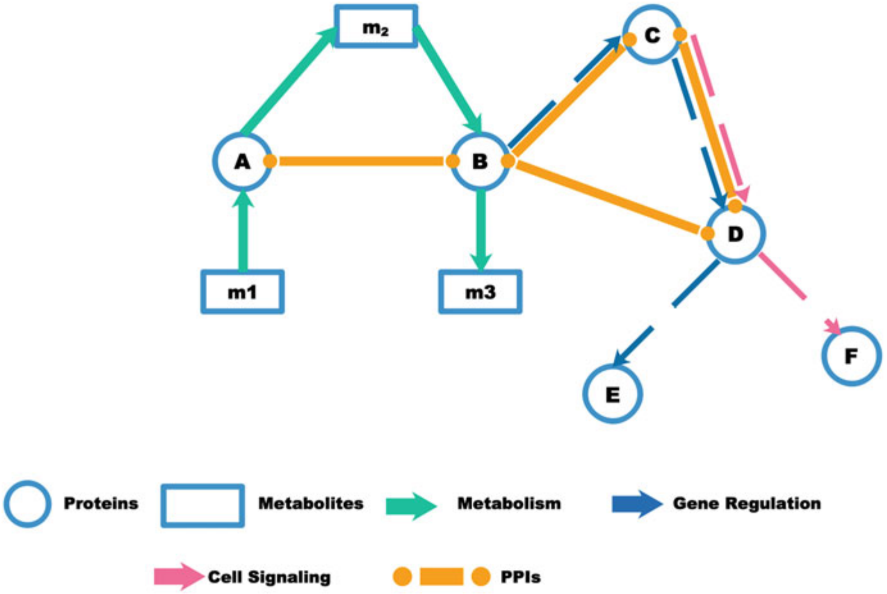


Fig. 11.6 Types of biological interactions that can be represented by networks. Adapted from (Koh et al. 2012)

Biological networks have been used to study transcription-regulation processes in Escherichia Coli, where it has been demonstrated that essential molecular elements contribute to the specialization of the dynamics of global responses, which allows the bacteria to have a more robust and quicker response to processes and environmental signs. They are also used to study networks of metabolic interactions; study host-pathogen interactions, discover new measurements, identify biomarkers; identification of genes involved in specific cell cycle processes; identification of disease-related genes critical biological processes (Pitkänen et al. 2010). Networks have been used to generate models of the relationship between elements of biological data sets, as well as the analysis of chromatin formation within cells (Tordini et al. 2016); the identification of metabolic pathways related to genetic regulation (Karlebach and Shamir 2008); as well as to model protein-protein interactions that take place within organisms (Beretta et al. 2019a; Pizzuti and Rombo 2014).

The networks of biological inferences can be classified into four types: protein-protein, Gene regulation networks, metabolic networks, signaling networks, and co-expression networks (Fig. 11.6).

11.4.1 Biological Network Properties

The biological interaction networks have specific architectures and properties that enable the analysis and interpretation of the complexity of the interactions present within the different domains and elements present in them (Aoki et al. 2007; Beretta et al. 2019b). Some of the elements and property that are part of the networks are as follows:

- Node: it is an individual element within the network;
- Edge: represent the interactions and interconnections between nodes;
- Components: are a group or groups of nodes that are mutually connected;
- Degree/connectivity: the number of edges connected between a focal node and the other nodes;
- Network density: describes the portion of potential connections on a network that are real connections. A potential connection is a connection that can exist between two nodes, regardless of whether they exist or not;
- Betweenness: is the metric that measures how a node is in the path between the other nodes. Nodes with a high centrality may have a strong influence due to their control over the passage of information within the network;
- Closeness: is the measure of the shortest path between one node and the other nodes within the network;
- Clustering coefficient: it is a measure of the proportion of neighbors reached through a node to the other neighbors. This metric shows the degree to which the nodes in a network tend to group. This metric allows measuring the cohesion of the network;
- Degree of distribution: it is the distribution of the frequencies of the degrees of the nodes individually for an entire network;
- Modules or clustering: a set of densely interconnected nodes within the network;
- Motif: they are small subnets or patterns that are statistically overrepresented within the network;
- Clique: consists of a fully connected subnet within a given network;
- Directed graph: nodes in a directed graph are connected by an asymmetric relationship, such as predation;
- Undirected graph: nodes in an undirected graph are connected by a symmetric relationship, such as physical interactions.

11.4.2 Types of Biological Networks

11.4.2.1 Metabolic Networks

This network type allows the annotation of genes and metabolic ones by determining elements of relationships, structure, and dynamics of metabolic networks. This network infers the enzymatic function of a specific protein or reconstructs the metabolic pathways in which it participates (Pavlopoulos et al. 2011; Tieri et al. 2019).

11.4.2.2 Signaling Network

It allows representing abstractions of molecular interactions and chemical modifications that act in a chain to transport stimuli (hormones, pathogens, nutrients) detected by the cell membrane receptors to the cell nucleus, to coordinate the beginning of the appropriate metabolic and genetic responses. For the reconstruction of this type of network, techniques such as genetic knockout have been used, which allow studying the different responses of organisms to this stimulus (Tieri et al. 2019).

11.4.2.3 Gene Regulation Network

Also known as the transcription regulation network, networks represent the casual interactions between transcription factors and genes. They are usually represented as a directed graph, whose direction is defined by the genes' expression. The interconnections between genes can represent biochemical processes such as reaction, transformation, interaction, activation, or inhibition (De Smet and Marchal 2010; Oates and Mukherjee 2012; Tieri et al. 2019).

The gene regulation networks present the set of activation and inhibition gene interactions within cells. Several transcription factors and gene products participate in the transcription process, regulating, directly or indirectly, other biomolecules within the genome, through regulatory chains. On the other hand, feedback loops can also be generated within this process, regulating negatively (downregulated) or positively (upregulated) gene product production (Fionda 2019).

11.4.2.4 Protein-Protein Interaction Network

Protein-protein interaction (PPIs) networks consist of proteins and their interactions. In this network, the nodes represent the proteins, and the edges correspond to the interactions between the proteins. Proteins are organized into different putative complexes, each performing a specific task or process within cells. A protein-protein interaction occurs when two or more proteins come together temporarily to modify each other, trigger signal transduction, or perform specific biological functions for a prolonged time (Pizzuti and Rombo 2014; Zhang et al. 2014).

The construction and analysis of PPI networks enable identifying protein complexes, which permits the study and understanding of the mechanisms that regulate life, explaining the evolutionary orthology signal, the prediction of biological functions of uncharacterized proteins, and drug target detections for specific diseases. One of the most used techniques for detecting protein complexes is clustering techniques, which allow groups of proteins that share similarity or common domains to be identified (Jancura et al. 2012; Pizzuti and Rombo 2014; Zhang et al. 2014).

The PPI networks allow to represent the relationships that occur between proteins within cells, that is, the interactome of the organisms that are under study. The use of protein-protein interaction detection techniques, such as high-throughput affinity purification combined with mass spectrometry and the yeast two-hybrid assay and PPI prediction algorithms, have made it possible to construct and study more complex and complete interactomes. Despite the advances, it is essential to note that current knowledge about the interactome is incomplete and noisy. The techniques used have limitations in terms of the number of genuinely physiological interactions and present some false positives and false negatives (Pellegrini 2019; Pizzuti and Rombo 2014; Zhang et al. 2014).

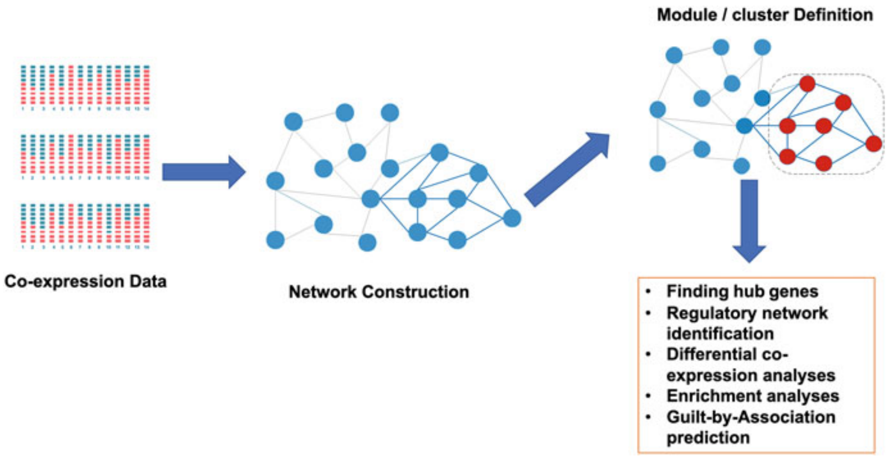

Co-expression Data
Network Construction
Module / cluster Definition
• Finding hub genes
• Regulatory network identification
• Differential co-expression analyses
• Enrichment analyses
• Guilt-by-Association prediction

Fig. 11.7 Workflow for generating and analyzing a co-expression network. The figure shows the different kinds of analyzes that can be performed with this type of network. Adapted from (van Dam et al. 2018)

11.4.2.5 Biological Co-Expression Network

These networks allow exploring the existing transcript-transcript associations and interactions, where genes are generally interconnected since there is an association of co-expression between them. This network is built from gene expression data and is represented as an undirected network (Fig. 11.7) (De Smet and Marchal 2010; Tieri et al. 2019).

Co-expression networks (CEN) can be used for various purposes, such as identifying genes with the most significant influences within networks, prioritizing disease candidate genes, functional annotation of genes, and identifying regulatory genes within networks (van Dam et al. 2018).

Co-expression networks are built by analyzing gene expression profiles' similarity, using techniques such as the correlation coefficient, distance metrics, and developed algorithms based on statistical metrics. The interconnections between the genes are determined using a cut-off that allows structuring the interconnections within the networks. The use of these types of control through cut-off allows the network to represent complex processes and patterns within the organisms (van Dam et al. 2018; Tieri et al. 2019).

The CENs have properties such as transitivity, allow the identification of dense communities of genes within the networks, which indicate that the member genes of these communities are functionally related; inside the cluster, there are nodes (genes) that present a high degree of interconnection, these are called central genes (hubs), this type of genes are generally more important to define the functionality of the network, in addition to being able to explain the functioning of each module better. Another property of CENs is the free-scale property. This kind of network has a degree of distribution that follows the power-law distribution, in which most nodes have low degrees, in contrast to the existence of a few nodes with high degrees, which indicates the probability that one node can connect to another is directly proportional to its degree. Due to this property, it is possible to identify a small set of central genes (hubs) and a broad set of genes with few interactions. This characteristic allows networks to be more robust (van Dam et al. 2018; Tieri et al. 2019).

Co-expression networks have been used to understand the relationships between genes' expression and the study of different phenomena and interactions between genes. CENs were used to identify genes related to the synthesis and metabolization of fenbendazole and flunixin meglumine in pig livers; this research found eight gene modules that showed a high relation to the level of transcripts relating the metabolism of these medications (Howard et al. 2017).

Exciting research that used co-expression networks were carried out by Shaik and Ramakrishna (Shaik and Ramakrishna 2013). This study presents the common genes for responses to water and bacterial stress present in rice and Arabidopsis. The team was able to identify several common gene modules that showed high co-expression and specific hubs related to these stresses.

CENs were used to study the genes involved in developing the skeletons and muscle mass of mice for myostatin. In the study developed by Yang (Yang et al. 2015), the researchers built co-expression networks using microarray data, which allowed them to study biological processes and metabolic pathways related to the development of muscles and skeletons in wild mice. This study allowed confirmed and identified new transcriptional regulators.

CEN can be used for the detection of biomarkers. The research developed by the team of Zhao and Li (2019), studied gestational diabetes mellitus and managed to identify ten potential biomarkers that help in diagnosing and therapy of this disease through co-expression networks.

In the research of Yuan et al. (2018), biomarkers for the diagnosis of adrenocortical carcinoma were analyzed. Within this study, they analyzed 12 central genes (hubs) within the networks that showed a correlation associated with the prognosis and progress of the disease; another team (Kommadath et al. 2014) used co-expression networks to detect candidate regulatory genes that present differential expression and that contribute to the spread of Salmonella enterica in pigs.

References

Altman RB, Raychaudhuri S (2001) Whole-genome expression analysis: challenges beyond clustering. Curr Opin Struct Biol 11(3):340–347

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25 (17):3389–3402

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181 (4096):223–230

Ansorge W, Sproat BS, Stegemann J, Schwager C (1986) A non-radioactive automated method for DNA sequence determination. J Biochem Biophys Methods 13(6):315–323

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M (1987) Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res 15 (11):4593–4602

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol 48(3):381–390

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M et al (2004) Uniprot: the universal protein knowledgebase. Nucleic Acids Res 32(suppl_1):D115–D119

Araujo FA, Barh D, Silva A, Guimarães L, Ramos RTJ (2018) Go feat: a rapid web-based functional annotation tool for genomic and transcriptomic data. Sci Rep 8(1):1–4

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M et al (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9(1):1–15

Baker M (2012) De novo genome assembly: what every biologist should know. Nat Methods 9 (4):333–337

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD et al (2012) Spades: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19(5):455–477

Beretta S, Denti L, Previtali M (2019a) Graph theory and definitions. Academic Press, Cambridge, MA

Beretta S, Denti L, Previtali M (2019b) Network properties. Academic Press, Cambridge, MA

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using diamond. Nat Methods 12(1):59–60

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB (2008) Allpaths: de novo assembly of whole-genome shotgun microreads. Genome Res 18 (5):810–820

Card DC, Schield DR, Reyes-Velasco J, Fujita MK, Andrew AL, Oyler-McCance SJ, Fike JA, Tomback DF, Ruggiero RP, Castoe TA (2014) Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. PLoS One 9(9):e106649

Celis JE, Kruhøffer M, Gromova I, Frederiksen C, Østergaard M, Thykjaer T, Gromov P, Yu J, Pálsdóttir H, Magnusson N et al (2000) Gene expression profiling: monitoring transcription and translation products using dna microarrays and proteomics. FEBS Lett 480(1):2–16

Chasman D, Siahpirani AF, Roy S (2016) Network-based approaches for analysis of complex biological systems. Curr Opin Biotechnol 39:157–166

Chen Q, Lan C, Zhao L, Wang J, Chen B, Chen YPP (2017) Recent advances in sequence assembly: principles and applications. Brief Funct Genomics 16(6):361–378

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21 (18):3674–3676

Consortium GO (2015) Gene ontology consortium: going forward. Nucleic Acids Res 43(D1): D1049–D1056

Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) The Knuth-Morris-Pratt algorithm. In: Introduction to algorithms, 2nd edn. MIT Press, Cambridge, MA

Crick F (1970) Central dogma of molecular biology. Nature 227(5258):561–563

Dasgupta D, Yu S, Nino F (2011) Recent advances in artificial immune systems: models and applications. Appl Soft Comput 11(2):1574–1587

Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. In: Atlas of protein sequence and structure, vol 5. The National Biomedical Research Foundation, Silver Spring, MD, pp 345–352

De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nat Rev Microbiol 8(10):717–729

Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont dna with glimmer. Bioinformatics 23(6):673–679

Dey A, Saha I, Maulik U (2017) A survey on multiple sequence alignment using metaheuristics. In: 2017 7th international conference on communication systems and network technologies (CSNT). IEEE, pp 279–284

Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge

Earl D, Bradnam K, John JS, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M et al (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res 21(12):2224–2241

Edgar RC (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797

Edgar RC (2010) Search and clustering orders of magnitude faster than blast. Bioinformatics 26 (19):2460–2461

Edgar RC, Batzoglou S (2006) Multiple sequence alignment. Curr Opin Struct Biol 16(3):368–373

Ekblom R, Wolf JB (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7(9):1026–1042

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A et al (2019) The pfam protein families database in 2019. Nucleic Acids Res 47(D1):D427–D432

El-Metwally S, Hamza T, Zakaria M, Helmy M (2013) Next-generation sequence assembly: four stages of data processing and computational challenges. PLoS Comput Biol 9(12):e1003345

Feng DF, Doolittle RF (1987) Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. J Mol Evol 25(4):351–360

Ferragina P, Manzini G (2000) Opportunistic data structures with applications. In: Proceedings 41st annual symposium on foundations of computer science. IEEE, pp 390–398

Fionda V (2019) Networks in biology. In: Ranganathan S, Gribskov M, Nakai K, Schön-bach C (eds) Encyclopedia of bioinformatics and computational biology. Academic Press, Oxford, pp 915–921. https://doi.org/10.1016/B978-0-12-809633-8.20420-2

Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28(24):3169–3177

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) Quast: quality assessment tool for genome assemblies. Bioinformatics 29(8):1072–1075

Gusfield D (1997) Algorithms on stings, trees, and sequences: computer science and computational biology. ACM Sigact News 28(4):41–60

Hacia JG (1999) Resequencing and mutational analysis using oligonucleotide microarrays. Nat Genet 21(1):42–47

Hatakeyama M, Aluri S, Balachadran MT, Sivarajan SR, Patrignani A, Grüter S, Poveda L, Shimizu-Inatsugi R, Baeten J, Francoijs KJ et al (2018) Multiple hybrid de novo genome assembly of finger millet, an orphan allotetraploid crop. DNA Res 25(1):39–47

Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV (2013) Benchmarking short sequence mapping tools. BMC Bioinformatics 14(1):184

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107(1):1–8

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89(22):10915–10919

Higgins DG, Sharp PM (1988) Clustal: a package for performing multiple sequence alignment on a microcomputer. Gene 73(1):237–244

Higgins DG, Bleasby AJ, Fuchs R (1992) Clustal v: improved software for multiple sequence alignment. Bioinformatics 8(2):189–191

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5(9):e1000502

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A (1965) Structure of a ribonucleic acid. Science 147:1462–1465

Holt C, Yandell M (2011) Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12(1):491

Howard JT, Ashwell MS, Baynes RE, Brooks JD, Yeatts JL, Maltecca C (2017) Gene co-expression network analysis identifies porcine genes associated with variation in metabolizing fenbendazole and flunixin meglumine in the liver. Sci Rep 7(1):1–12

Hsiao LL, Stears RL, Hong RL, Gullans SR (2000) Prospective use of dna microarrays for evaluating renal function and disease. Curr Opin Nephrol Hypertens 9(3):253–258

Husi H, Skipworth RJ, Fearon KC, Ross JA (2013) Lscluster, a large-scale sequence clustering and aligning software for use in partial identity mapping and splice-variant analysis. J Proteome 84:185–189

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11(1):119

Jancura P, Mavridou E, Carrillo-de Santa Pau E, Marchiori E (2012) A methodology for detecting the orthology signal in a PPI network at a functional complex level. BMC Bioinformatics 13: S18

Jiang Z, Zhou X, Li R, Michal JJ, Zhang S, Dodson MV, Zhang Z, Harland RM (2015) Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. Cell Mol Life Sci 72(18):3425–3439

Kambara H, Nishikawa T, Katayama Y, Yamaguchi T (1988) Optimization of parameters in a dna sequenator using fluorescence detection. Bio/Technology 6(7):816–821

Karlebach G, Shamir R (2008) Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol 9(10):770–780

Keel BN, Snelling WM (2018) Comparison of burrows-wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: application to illumina data for livestock genomes1. Front Genet 9:35

Kent WJ (2002) Blat—the blast-like alignment tool. Genome Res 12(4):656–664

Koh GC, Porras P, Aranda B, Hermjakob H, Orchard SE (2012) Analyzing protein–protein interaction networks. J Proteome Res 11(4):2014–2031

Kommadath A, Bao H, Arantes AS, Plastow GS, Tuggle CK, Bearson SM, Stothard P et al (2014) Gene co-expression network analysis identifies porcine genes associated with variation in salmonella shedding. BMC Genomics 15(1):1–15

König S, Romoth LW, Gerischer L, Stanke M (2016) Simultaneous gene finding in multiple genomes. Bioinformatics 32(22):3388–3395

Koonin EV, Galperin M (2013) Sequence—evolution—function: computational approaches in comparative genomics. Springer, Dordrecht

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5(1):59

Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, Arumugam M, Pan Q, Liu B, Qin J et al (2012) Mocat: a metagenomics assembly and gene prediction toolkit. PLoS One 7(10): e47656

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al (2001) Initial sequencing and analysis of the human genome. Nature 409 (6822):860–921

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9(4):357

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol 10(3):R25

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11(5):473–483

Li H, Ruan J, Durbin R (2008) Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome Res 18(11):1851–1858

Lindgreen S, Gardner PP, Krogh A (2007) Mastr: multiple alignment and structure prediction of non-coding rnas using simulated annealing. Bioinformatics 23(24):3304–3311

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227 (4693):1435–1441

Lischer HE, Shimizu KK (2017) Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 18(1):1–12

Liu Y, Guo J, Hu G, Zhu H (2013) Gene prediction in metagenomic fragments based on the SVM algorithm. BMC Bioinformatics 14:S12

Lomsadze A, Burns PD, Borodovsky M (2014) Integration of mapped rna-seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res 42(15):e119–e119

Lourenço HR, Martin OC, Stützle T (2003) Iterated local search. In: Handbook of metaheuristics. Springer, Boston, pp 320–353

Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, Smith LM (1990) High speed dna sequencing by capillary electrophoresis. Nucleic Acids Res 18(15):4417–4421

Lukashin AV, Borodovsky M (1998) Genemark.hmm: new solutions for gene finding. Nucleic Acids Res 26(4):1107–1115

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y et al (2012) Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1(1):2047-217X

Majoros WH, Pertea M, Salzberg SL (2004) Tigrscan and glimmerhmm: two open source ab initio eukaryotic gene-finders. Bioinformatics 20(16):2878–2879

Manger ID, Relman DA (2000) How the host 'sees' pathogens: global gene expression responses to infection. Curr Opin Immunol 12(2):215–218

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G (2012) Wisdom of crowds for robust gene network inference. Nat Methods 9(8):796–804

Mardis ER (2011) A decade's perspective on dna sequencing technology. Nature 470 (7333):198–203

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12 (10):671–682

Martorell-Marugán J, Tabik S, Benhammou Y, del Val C, Zwir I, Herrera F, Carmona-Sáez P (2019) Deep learning in omics data analysis and precision medicine. Codon Publications, Brisbane, pp 37–53

Maulik U, Saha I (2009) Modified differential evolution based fuzzy clustering for pixel classification in remote sensing imagery. Pattern Recogn 42(9):2135–2149

Maxam AM, Gilbert W (1977) A new method for sequencing dna. Proc Natl Acad Sci 74 (2):560–564

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC et al (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19(9):1527–1541

Miller JR, Zhou P, Mudge J, Gurtowski J, Lee H, Ramaraj T, Walenz BP, Liu J, Stupar RM, Denny R et al (2017) Hybrid assembly with long and short reads improves discovery of gene family expansions. BMC Genomics 18(1):541

Min B, Grigoriev IV, Choi IG (2017) Fungap: Fungal genome annotation pipeline using evidence-based gene model evaluation. Bioinformatics 33(18):2936–2937

Misra S, Narayanan R, Lin S, Choudhary A (2010) Fangs: high speed sequence mapping for next generation sequencers. In: Proceedings of the 2010 ACM symposium on applied computing. ACM, New York, pp 1539–1546

Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI et al (2019) Interpro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 47(D1):D351–D360

Mladenović N, Hansen P (1997) Variable neighborhood search. Comput Oper Res 24 (11):1097–1100

Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34(2):188–193

Naama B, Bouzeboudja H, Allali A (2013) Application of Tabu search and genetic algorithm in minimize losses in power system. Using the b-coefficient method. Energy Procedia 36:687–693

Nagarajan N, Pop M (2013) Sequence assembly demystified. Nat Rev Genet 14(3):157–167

Navarro G (2001) A guided tour to approximate string matching. ACM Comput Surv (CSUR) 33 (1):31–88

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453

Nowak RM, Jastrzębski JP, Kuśmirek W, Sałamatin R, Rydzanicz M, Sobczyk-Kopcioł A, Sulima-Celińska A, Paukszto Ł, Makowczenko KG, Płoski R et al (2019) Hybrid de novo whole-genome assembly and annotation of the model tapeworm hymenolepis diminuta. Sci Data 6 (1):1–14

Oates CJ, Mukherjee S (2012) Network inference and biological dynamics. Ann Appl Stat 6 (3):1209

Ortuño FM, Valenzuela O, Rojas F, Pomares H, Florido JP, Urquiza JM, Rojas I (2013) Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: structural information, non-gaps percentage and totally conserved columns. Bioinformatics 29 (17):2112–2121

Palmer J, Stajich J (2017) Funannotate: eukaryotic genome annotation pipeline

Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG (2011) Using graph theory to analyze biological networks. BioData Min 4(1):10

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci 85(8):2444–2448

Pellegrini M (2019) Community detection in biological networks. In: Encyclopedia of bioinformatics and computational biology. Elsevier, Amsterdam

Pennisi E (2003) A low number wins the GeneSweep pool. Science 300:1484

Pitkänen E, Rousu J, Ukkonen E (2010) Computational methods for metabolic reconstruction. Curr Opin Biotechnol 21(1):70–77

Pizzuti C, Rombo SE (2014) Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. Bioinformatics 30 (10):1343–1352

Pop M, Phillippy A, Delcher AL, Salzberg SL (2004) Comparative genome assembly. Briefings in. Bioinformatics 5(3):237–248

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K (1987) A system for rapid dna sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238(4825):336–341

Proulx SR, Promislow DE, Phillips PC (2005) Network thinking in ecology and evolution. Trends Ecol Evol 20(6):345–353

Raghava GP, Barton GJ (2006) Quantification of the variation in percentage identity for protein sequence alignments. BMC Bioinformatics 7(1):1–4

Remmert M, Biegert A, Hauser A, Söding J (2012) Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nat Methods 9(2):173–175

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) Vsearch: a versatile open source tool for metagenomics. PeerJ 4:e2584

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348–352

Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated markov models. Nucleic Acids Res 26(2):544–548

Sanger F, Coulson A (1996) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Sel Pap Frederick Sanger Comment 94:382

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74(12):5463–5467

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19(R2):R227–R240

Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N et al (2011) Reference-guided assembly of four diverse arabidopsis thaliana genomes. Proc Natl Acad Sci 108(25):10249–10254

Schnepp PM, Chen M, Keller ET, Zhou X (2019) Snv identification from single-cell rna sequencing data. Hum Mol Genet 28(21):3569–3583

Scholz MB, Lo CC, Chain PS (2012) Next generation sequencing and bioinformatic bottle-necks: the current state of metagenomic data analysis. Curr Opin Biotechnol 23(1):9–15

Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong CS, Philips P, De Bona F, Hartmann L, Bohlen A et al (2009) mgene: accurate svm-based gene finding with an application to nematode genomes. Genome Res 19(11):2133–2143

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14):2068–2069

Shaik R, Ramakrishna W (2013) Genes and co-expression modules common to drought and bacterial stress responses in arabidopsis and rice. PLoS One 8(10):e77261

Shendure J, Ji H (2008) Next-generation dna sequencing. Nat Biotechnol 26(10):1135–1145

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) Abyss: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123

Smith TF, Waterman MS et al (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197

Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE (1985) The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent dna primers for use in dna sequence analysis. Nucleic Acids Res 13(7):2399–2412

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE (1986) Fluorescence detection in automated dna sequence analysis. Nature 321(6071):674–679

Sohn Ji, Nam JW (2018) The present and future of de novo whole-genome assembly. Brief Bioinform 19(1):23–40

Suzuki S, Kakuta M, Ishida T, Akiyama Y (2014) Ghostx: an improved sequence homology search algorithm using a query suffix array and a database suffix array. PLoS One 9(8):e103833

Swerdlow H, Gesteland R (1990) Capillary gel electrophoresis for rapid, high resolution dna sequencing. Nucleic Acids Res 18(6):1415–1419

Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J (2016) Ncbi prokaryotic genome annotation pipeline. Nucleic Acids Res 44(14):6614–6624

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680

Tieri P, Farina L, Petti M, Astolfi L, Paci P, Castiglione F (2019) Network inference and reconstruction in bioinformatics. Encycl Bioinform Comput Biol 2:805–813

Toledo-Arana A, Solano C (2010) Deciphering the physiological blueprint of a bacterial cell: revelations of unanticipated complexity in transcriptome and proteome. BioEssays 32 (6):461–467

Tordini F, Aldinucci M, Milanesi L, Liò P, Merelli I (2016) The genome conformation as an integrator of multi-omic data: the example of damage spreading in cancer. Front Genet 7:194

van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP (2018) Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform 19 (4):575–592

Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30(9):418–426 van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C (2018) The third revolution in sequencing technology. Trends Genet 34(9):666–681

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) The sequence of the human genome. Science 291(5507):1304–1351

Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55(4):641–658

Wang Y, Chen L, Song N, Lei X (2015) Gass: genome structural annotation for eukaryotes based on species similarity. BMC Genomics 16(1):150

Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R et al (2014) Patric, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 42(D1):D581–D591

Wu TD, Nacu S (2010) Fast and snp-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7):873–881

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13 (5):329–342

Yang X, Koltes JE, Park CA, Chen D, Reecy JM (2015) Gene co-expression network analysis provides novel insights into myostatin regulation at three different mouse developmental timepoints. PLoS One 10(2):e0117607

Yuan L, Qian G, Chen L, Wu CL, Dan HC, Xiao Y, Wang X (2018) Co-expression network analysis of biomarkers for adrenocortical carcinoma. Front Genet 9:328

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome Res 18(5):821–829

Zhang J, Lin H, Balaji P, Feng WC (2013) Optimizing burrows-wheeler transform-based sequence alignment on multicore architectures. In: 2013 13th IEEE/ACM international symposium on cluster, cloud, and grid computing. IEEE, pp 377–384

Zhang B, Tian Y, Zhang Z (2014) Network biology in medicine and beyond. Circulation: cardiovascular. Genetics 7(4):536–547

Zhang KY, Gao YZ, Du MZ, Liu S, Dong C, Guo FB (2019) Vgas: a viral genome annotation system. Front Microbiol 10:184

Zhao X, Li W (2019) Gene coexpression network analysis identified potential biomarkers in gestational diabetes mellitus progression. Mol Gen Genom Med 7(1):e00515

A Guide to RNAseq Data Analysis Using Bioinformatics Approaches

12

Preeti Sharma, B. Sharan Sharma, and Ramtej J. Verma

Abstract

The emergence of Next Generation Sequencing (NGS), such as DNA, RNA and other small RNA sequencing technologies, gave rise to a huge amount of raw data on a massive scale. To analyse that data and to obtain the biological interpretation as a challenging act, advancements in computational biology and bioinformatics applications emerged as the need of the hour. RNAseq accounts for exploration of comprehensive expression profile of genes and quantifies the presence of RNA content in the biological sample. In addition to this, RNAseq also provides information for alternative splice variants, novel gene identification, differentially expressing genes, etc. The workflow for RNAseq data analysis requires quality check of the data, mapping onto a reference genome/transcriptome, read quantification, differential expression analysis and functional annotation. Various tools and softwares with different algorithms have been developed to provide biological understanding of the data and to meet the demands of the analyst. An overview of the tools and softwares has been provided in the chapter that can be exploited to analyse the data for different investigations. Also, a glimpse of

P. Sharma (✉)
Department of Zoology, Biomedical Technology and Human Genetics, University School of Sciences, Gujarat University, Ahmedabad, Gujarat, India

PanGenomics International Pvt Ltd, Sterling Accuris Diagnostics, Ahmedabad, Gujarat, India

B. S. Sharma
Genexplore Diagnostics and Research Centre, Ahmedabad, Gujarat, India

Rivaara Labs Pvt Ltd, KD Hospital, Ahmedabad, Gujarat, India

R. J. Verma
Department of Zoology, Biomedical Technology and Human Genetics, University School of Sciences, Gujarat University, Ahmedabad, Gujarat, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_12

other RNAseq techniques such as single cell RNAseq and small RNA sequencing has been discussed as an introduction to newer forms of RNA sequencing.

Keywords

Next generation sequencing · Transcriptome · Pre-processing · Quantification · Normalization

12.1 Introduction

With the advent of NGS technologies, RNA sequencing (RNAseq) occurred as a pivotal approach to evaluate the expression of a whole genomic profile. Sooner, the technique was exploited tremendously for certain advantages over others, such as identification of novel genes, unlike microarrays, detection of alternative splice variants, differentially expressing transcripts, etc.

The vast and varied applicability of RNAseq by offering results in multiple forms led to the generation of huge loads of data, also referred to as 'Big Data'. Resultantly, the technological expansion in the era of NGS also directed the evolution in the field of computational biology. Different tools and softwares were developed to analyse and interpret the results from the data generated on different platforms, such as SoLiD sequencing, Ion Torrent Platform, Illumina sequencing, etc. The procedure for RNAseq data analysis takes place in a number of steps which involves cDNA preparation, fragmentation followed by adapter ligation, cDNA library preparation and amplification (Han et al. 2015), etc. The fragments are read and sequenced to obtain the raw sequence data in the prescribed formats. These raw data sequences are then analysed to extract meaningful results from the sequences using various tools and pipelines.

The workflow for data analysis involves quality check and pre-processing of the raw reads, assembly to a reference genome, quantification of transcripts and identification of differentially expressed transcripts. The transcripts of interest are then annotated to different databases for functional enrichment, gene ontology analysis and pathway enrichment, etc. (Garber et al. 2011). A schematic workflow of the steps involved in data analysis has been shown in Fig. 12.1.

To explore deep into the genome or transcriptome (Sharma et al. 2020), other RNAseq technologies such as single cell RNAseq, small RNA sequencing etc. were developed. The development of these modified versions of RNAseq technologies also led to certain variabilities during sample processing, technical noise, normalization processes, etc. The challenges in data analyses for these processes accounted for advancements in development of computational tools and bioinformatics applications with certain modifications.

The present chapter provides an overview of workflow for analysis of RNAseq data on different sequencing platforms using bioinformatics approaches. Also, a brief outlook of different tools and softwares, based on different algorithms, can provide an understanding of using them in multiple dimensions depending upon the type of analysis to be performed (Table 12.1).

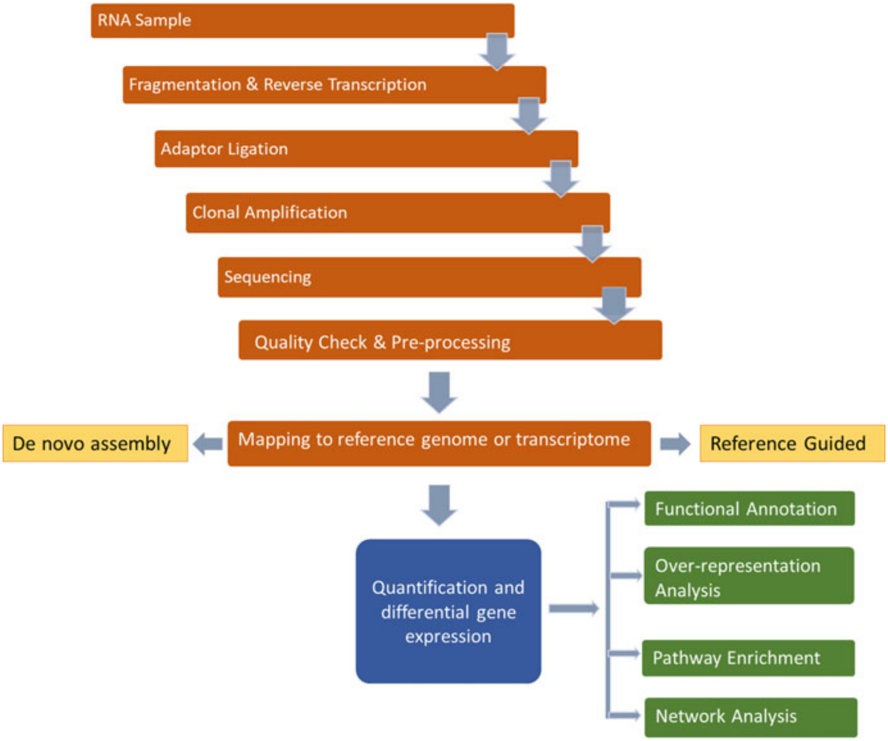

RNA Sample
Fragmentation & Reverse Transcription
Adaptor Ligation
Clonal Amplification
Sequencing
Quality Check & Pre-processing
De novo assembly
Mapping to reference genome or transcriptome
Reference Guided
Quantification and differential gene expression
Functional Annotation
Over-representation Analysis
Pathway Enrichment
Network Analysis

Fig. 12.1 Schematic Workflow showing steps in RNAseq data analysis

12.2 Platforms Available for Sequencing

Since the commencement of sequencing technologies various platforms have been developed which are based on different chemistries. The differences in the sequencing platforms also lie in the data output, performance and data quality. Some of the sequencing platforms and chemistries are discussed below:

12.2.1 SOLiD

SOLiD stands for Sequencing by Oligo Ligation and Detection and the technique was developed in 2005 (Hedges et al. 2011). It is based on oligonucleotide ligation to ligate dsDNA strands with the help of enzyme DNA ligase. A primer-binding adapter is bound to the target sequence on a bead, which is then amplified using emulsion PCR. A universal primer is hybridized to the adapter, followed by exposure of beads to a library of 8-nucleotide probes tagged with four different fluorescent dyes at 5’end and a hydroxyl group at 3’end. Based on the complementarity of

Table 12.1 List of tools available for different analytical processes of RNAseq data analysis

S. no.	Process	Tool	Link
1.	Quality check	FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
		Kraken	https://github.com/DerrickWood/kraken2
		HTSeq	https://htseq.readthedocs.io/en/master/
		NGS QC Toolkit	http://www.nipgr.res.in/ngsqctoolkit.html
		RNASeQC	https://github.com/getzlab/rnaseqc
2.	Pre-processing	BBDuk	https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk.sh
		Cutadapt	https://bioinformaticshome.com/tools/rna-seq/descriptions/cutadapt.html
		FASTX Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
		SortMeRNA	https://bioinfo.lifl.fr/RNA/sortmerna/
		Trimmomatic	https://github.com/timflutre/trimmomatic
3.	Alignment of reads		
	Reference guided	Bowtie	http://bowtie-bio.sourceforge.net/index.shtml
		Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
		Burrows-Wheeler Aligner (BWA)	http://bio-bwa.sourceforge.net/
		Bayesembler	https://github.com/bioinformatics-centre/bayesembler
		Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/
		IsoLasso	http://alumni.cs.ucr.edu/~liw/isolasso.html
	De novo assemblers	CLC Genomics Workbench	https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-clc-genomics-workbench/
		Oases	https://github.com/dzerbino/oases
		rnaSPAdes	https://cab.spbu.ru/software/rnaspades/
		Rnnotator	https://www.osti.gov/biblio/1231732-rnnotator
		SOAPdenovo-trans	http://sourceforge.net/projects/soapdenovotrans/
		Trans-ABySS	https://github.com/bcgsc/transabyss
		Trinity	https://github.com/trinityrnaseq/trinityrnaseq/wiki
		Velvet	https://www.ebi.ac.uk/~zerbino/velvet/
4.	Assembly evaluation tools	Busco	https://busco.ezlab.org/
		Detonate	http://deweylab.biostat.wisc.edu/detonate/
		rnaQUAST	https://github.com/ablab/rnaquast
		TransRate	https://hibberdlab.com/transrate/

(continued)

Table 12.1 (continued)

S. no.	Process	Tool	Link
	Co-expression networks		http://gnw.sourceforge.net/
		WGCNA	http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA.
5.	Functional, network and pathway analysis tools	BioCyc	https://biocyc.org/
		FunRich	http://www.funrich.org/
		GeneSCF	http://genescf.kandurilab.org/
		GOexpress	http://bioconductor.org/packages/release/bioc/html/GOexpress.html
		PathwaySeq	https://rna-seqblog.com/pathwayseq-pathway-analysis-for-rna-seq-data/
		ToPASeq	https://www.bioconductor.org/packages/release/bioc/html/ToPASeq.html
		RNA-Enrich	http://lrpath.ncibi.org
6.	miRNA prediction and analysis	miRDeep2	https://www.mdc-berlin.de/content/mirdeep2-documentation
		miRExpress	http://mirexpress.mbc.nctu.edu.tw/
		miR-PREFeR	https://github.com/hangelwen/miR-PREFeR
		miRDeep-P	http://faculty.virginia.edu/lilab/miRDP/
		miRPlant	http://www.australianprostatecentre.org/research/software/mirplant
		ShortStack	https://github.com/MikeAxtell/ShortStack
		mireap	https://github.com/liqb/mireap

first two bases, the probes get attached to the target sequence with the help of the enzyme DNA ligase. The fluorescent tag is then cleaved from the fragment at 5th and 6th base of the probe which is joined by phosphorothioate linkage. The fluorescence of the dyes generated due to cleavage is measured at different spectra. After the completion of first round of sequencing, the second-round sequencing starts with primer of length N-1, and so on. The sequencing of the target is ensured by measuring the fluorescence signals at each round of sequencing. However, the technique was low-cost and provided results with high accuracy due to the two-base sequencing, the main disadvantages were the time-consumption and shorter read lengths (Wyrzykiewicz and Cole 1994).

12.2.2 Ion Torrent Semiconductor Sequencing

The Ion Torrent sequencing is well-versed as ‘semiconductor sequencing’, where the target is sequenced by measuring changes in the pH variation due to release of hydrogen ion after incorporation of a specific nucleotide (Quail et al. 2012). A cDNA library is prepared here by fragmenting the RNA using enzymatic degradation. The fragmented libraries are then ligated with complementary probes embedded on beads and mixed with PCR reagents and oil to perform emulsion PCR. Here, each microsphere of emulsion, specifically known as Ion Sphere Particles (ISPs), is covered with multiple copies of same DNA fragment for clonal amplification. After amplification the ISPs with template fragment are enriched from the mixture using biotin labelled magnetic beads and the rest are melted off. The positive templates are then prepared for sequencing and loaded onto Ion chips which contain millions of microwells with many copies of single-stranded DNA template and other sequencing reagents such as DNA polymerase, dNTPs in each well. The incorporation of the complementary nucleotide results into the change in pH level and is converted to digital signals to obtain the sequence of the target sequence. The technology is not based on fluorescence signals and does not require optical reading for detection so the sequencing is rapid and number of bases gets incorporated in less time. The technology limits in reading of homopolymer sequences in the template, such as 'TTTTTT', and becomes challenging to distinguish between the multiple oligomers, resulting into an increase in the error rate (Merriman et al. 2012).

12.2.3 Illumina Sequencing Technology

Illumina sequencing also known as 'sequencing by synthesis' approach (Ansorge 2009). Here, the target sequence is cleaved into smaller fragments of 100–150 bp to form a library and is ligated to customized adapters followed by generation of multiple copies of the same read using PCR. The adapter ligated templates are then washed onto a flow cell where millions of clusters are formed by the process of 'bridge amplification' PCR. The amplification process is carried out with DNA polymerase and modified dNTPs with a terminator tagged with a fluorescent label corresponding to each base. This terminator blocks the addition of another nucleotide and only one base is added by the polymerase at a time. The fluorescence is detected by imaging the signals, indicating a base that has been added to the sequence. With the addition of four nucleotides, the terminators are removed preparing the slide for next cycle of sequencing. The signals are then converted to construct the entire sequence. As the sequencing takes place in fixed cycles and of uniform read length, the sequences generated are also of uniform length (Meyer and Kircher 2010).

12.3 Quality Check and Pre-Processing of Reads

12.3.1 Formats Available for Storage of Raw Data

The sequences, can be referred to as raw reads, generated by sequencing on different platforms are stored in multiple files of short reads. After sequencing, the raw data is generated and can be stored in different file formats such as FASTQ, FASTA, SAM/BAM, etc.

- FASTQ is the most commonly used file format. It allows storing of data with corresponding quality values known as Phred scores. The files in fastq format are with extension '.fq' or '.fastq'. A FASTQ file contains four lines of textual information. The first line starts with a sign '@', generally known as a sequence identifier. The second line consists of a sequence of nucleotides, i.e. A, T, G, C. The third line consists of a '+' sign which is usually a separator and indicates the end of the sequence. The fourth line provides a quality score corresponding to the sequence in the second line (Deorowicz and Grabowski 2011).
- FASTA format is also one of the data storing formats and is available with extension '.fa' and '.fasta'. The sequences are recognized by a '>' sign in the beginning followed by a descriptive information about the sequence. This format is generally used while alignment or reference genome mapping by different tools and softwares. The sequence consists of nucleotides A, T, G, C and N (for undetermined base) (Gilbert 2003). The sequence can be viewed using text editor tools or LINUX/UNIX environment.
- BAM/SAM—The raw sequence data generated from the sequencer have no genomic information and are need to be aligned to a reference genome. After mapping or aligning to a reference genome, the output is generated in SAM/BAM format. SAM is Sequence Alignment/Map format which stores the sequences in an aligned format against the reference genome. A SAM file is a tab-delimited file, recognized by a '.sam' extension and can be viewed using text editor tools (Li et al. 2009). A BAM file is binary version of SAM file and is often found with '.bam' extension (Niemenmaa et al. 2012).

12.3.2 Quality Check Using Available Softwares and Tools

The data generated after sequencing often contains contaminants such as poor-quality reads, PCR artefacts, adapter sequences, over-represented sequences, etc. which interferes in downstream analytical operations of the data. Hence, the data needs to be quality checked to obtain clean and filtered high quality reads. For this, many softwares are available to assess the quality of the reads. These softwares perform a quality check (QC) on the data and provide a QC report depicting low-quality sequencing reads impeding the quality of the data. FASTQC is a commonly used tool for assessing the quality of the data. It measures scores associated with data such as read length, quality score, GC percentage, k-mers, etc. and produces results in different modules (Andrews 2010).

The *per base sequence quality* module assesses the overall quality of the bases at each position of the read which is represented by a box whisker plot. A higher score determines better quality of the base call. Likewise, *per sequence quality score* report presents a subset of overall sequences having low-quality scores. This constitutes a small fraction of the total sequences; however, a large subset possessing bad quality scores indicates some systematic errors.

The *per base GC content* shows the GC content of each base in the sequence. A shift in the graph of GC content with the underlying genome indicates presence of over-represented sequences creating a sequence bias. Further in this, *per sequence GC content* marks for GC content across whole length of sequences comparable to normal distribution plot of GC content. A shift of the plot from the normal distribution on the graph indicates some systematic bias which is independent of base position. Some other modules such as *per base N content*, *sequence length distribution, duplicate sequences, over-represented sequences* and *over-represented k-mers,* etc. also provide report for the quality of the data.

12.3.3 Pre-Processing of Data

Before using the data for functional annotation and differential expression, etc. it is required to be pre-processed for removal of contaminated reads. For this, various tools are available such as Fastx-toolkit (Gordon and Hannon 2010), NGStoolkit (Mulcare 2004), Trimmomatic (Bolger et al. 2014), etc. Fastx-toolkit is most commonly used tool to filter out the good data from the bad quality data. During the course of filtration, the data is processed for removal of low-quality bases, adapter sequences, and other such reads interfering with the quality of the data.

The sequencing data is often contaminated with adapter sequences which are synthetically designed fragments of DNA added to the target sequences. These sequences are generally removed by the sequencers after the completion of sequencing process. But less often they remain attached to the sequenced read and are responsible for background noise in the data. Various tools such as Cutadapt (Martin 2011), Trimmomatic (Bolger et al. 2014), etc. are most frequently used tools for removal of adapter sequences.

Other contaminants are bases with low-quality, i.e. those with high error rate of being incorrect. The quality of base is assigned by a phred score (Q score) value, which is commonly used to measure the accuracy of the base call while sequencing the read by the sequencer. A quality score of <20 is generally considered of poor quality with high chances of inaccuracy. Fastx-toolkit is the most commonly used tool to trim off the reads with phred score <20.

Few other sequences such as rRNA sequences also act as contaminants in case of whole transcriptome sequencing. To remove the rRNA reads, rRNAFilter (Wang et al. 2017), SortMeRNA (Kopylova et al. 2012) and RiboPicker (Schmieder and Edwards 2011) are commonly used tools for the process.

12.4 Assembling Reads to Reference Genome/Transcriptome

12.4.1 Alignment of Reads

The raw reads generated after sequencing are then mapped onto a reference genome or transcriptome of the same species or the nearest relative, whichever available. (Roberts et al. 2011; Trapnell et al. 2010). The mapping of reads is affected by complexities of the genome, polymorphisms, gene isoforms, alternative splicing, etc.

leading to reduced percentage of mapped reads. The percentage of reads assembled indicates the accuracy of the results and presence of contaminated sequences (Conesa et al. 2016). The mapping can be done uniquely to one position or can also be mapped to multiple reads due to presence of repetitive sequences. In case of reference transcriptome multiple reads are found more often because of the presence of all isoforms of genes in the transcriptome.

12.4.2 Reference Guided/de Novo Assembly

In reference guided assembly, the reads are mapped onto a reference genome or transcriptome, whichever available, to assemble them into transcripts. The reads to be mapped are split into parts where one part maps to the exonic part and the other one to the intronic region. Reads mapping on the reference genome minimizes the complexities in the assemblies as they are mapped specifically to their genomic locations (Voshall and Moriyama 2018). Several assemblers are available for reference guided assemblies, such as Bayesembler (Maretty et al. 2014), Cufflinks (Ghosh and Chan 2016), Stringtie (Pertea et al. 2015), etc. Different assemblers use different strategies to assemble reads with highest percentage of read coverage, such as Cufflinks uses few numbers of transcripts to assemble large number reads to the genome or transcriptome, whereas Bayesembler uses Bayesian likelihood to estimate the most likely combination of transcripts constructed for each splice junction. Other assemblers such as IsoLasso (Li et al. 2011) and iReckon (Mezlini et al. 2013) use L-1 norm and specific sparse constraints, respectively, to obtain possible transcripts combinations.

The reference guided assemblers use reference genomes to align the reads and assemble them into transcripts, where graphs are prepared and isoforms are considered as paths of graphs (Li and Xuejun 2016). The accuracy of the assembly depends on the availability of complete and good quality reference genome which are usually available for the model organisms such as human, mouse, rat, Arabidopsis, Oryza, etc., but not for non-model species.

Therefore, for species with no reference genome de novo or reference-independent method is used to construct the transcripts. The de novo assembly is based on generation of short fragments of reads known as k-mers which overlaps to form a de Bruijn graph structure (Martin and Wang 2011). The assemblage of contigs using different algorithms depends on the varying lengths of the k-mers. Shorter k-mers generally cover the reference sequences completely but also provides ambiguity because of the presence of multiple reads from different transcripts. In case of longer k-mers, ambiguity is resolved but also does not cover the entire region of the reference genome/transcriptome.

Various assemblers are available based on optimization of k-mer lengths for assemblage of contigs using different algorithms. SOAPdenovo-Trans (Xie et al. 2014) and Trinity (Freedman 2016) use the preferred k-mer lengths for producing the de Bruijn graph. Trinity is a package of three independent softwares: Inchworm, Chrysalis and Butterfly, where Inchworm assembles the transcripts, Chrysalis forms the de Bruijn graph by clustering those transcripts and finally Butterfly evaluates the graphs and produces the full-length assembly (Grabherr et al. 2011). rnaSPAdes (Bushmanova et al. 2019) identifies the k-mer lengths based on the read data. rnaSPAdes is the optimized version of SPAdes (Bankevich et al. 2012), where three assemblies are produced and one can choose any of them depending upon the downstream analyses. The three assemblies contain, one assembled with all transcripts, assembly with long and highly expressing transcripts, and assembly with short and lowly expressing transcripts (Geniza and Jaiswal 2017). Another assembler Velvet/Oases assembles the contigs based on de Bruijn graph using short reads. Velvet assembles the contigs using the short reads which are then clustered into loci using Oases program (Schulz et al. 2012).

12.4.3 Quality Check (QC) of Assembled Reads

Before processing the data for further downstream analysis the assembled reads are checked for their quality. The quality metrics of the assembled reads can be evaluated using two different criteria, either by calculating number and length of contigs or by mapping the assembled reads to coded proteins for similarity search. Softwares such as rnaQUAST (Bushmanova et al. 2016), CD-HIT (Li and Godzik 2006), TransRate (Smith-Unna et al. 2016) and Bowtie (Langmead 2010), etc. can be used to measure the quality of the assembly by measuring the lengths of the contigs and N50 value of the assemblies (T O'Neil and Emrich 2013). N50 value is defined as the minimum contig length required to cover fifty percent of the genome. While N50 value is more suitable quality of a genome assembly, transcriptome assembly is checked by measuring their ExN50 value which is dynamic and real time estimation of the assembled reads (Geniza and Jaiswal 2017).

ExN50 calculates the highly expressing transcripts which accounts for half of the overall transcriptome data. Another criterion based on mapping of the assembled reads to the coded proteins provides more probable notion of completeness of the assembled transcripts. The similarity searches are generally done by aligning the assembled reads against well-annotated databases containing non-protein sequences, conserved domains of proteins with functional annotation or lineage dependent protein databases (Nakasugi et al. 2014). These include BLAST (Altschul et al. 1990), Pfam (Finn et al. 2014), UniProt/Swiss-Prot (Apweiler et al. 2004), BUSCO (Waterhouse et al. 2018), etc. However, the protein-coded similarity search is a more plausible metric of QC of an assembly, the performance is limited by the relatedness of the biological entity in question to the sequences present in the databases. The more the divergence of the organism, more will be the possibility of lower percentage of assembled reads and gaps in the assembly.

12.5 Expression Quantification and Differential Expression

The first approach for transcriptome quantification is done by quantifying the expression of number of reads of specific transcripts. The most likely used method is maximizing likelihood (Glebova et al. 2016), based on different variants of expectation maximization (EM) (Li and Dewey 2011; Li and Jiang 2012), min-cost flow (Tomescu et al. 2013) and regression (Li et al. 2011), etc. RNAseq by Expectation Maximization (RSEM) quantitates the expression at isoform level and produces the output with 95% confidence interval. Moreover, all approaches use sequence specific transcripts to assess the expression level of each transcript. RSEM processing requires transcript sequences produced by the assembler as reference transcript sequences for RNAseq analysis for species with only transcript sequences available (Li and Dewey 2011). The mapped reads on multiple isoforms can be used to quantitate the expression in terms of prospective measures such as counting Fragments Per Kilobase of transcript per Million (FPKM) (Trapnell et al. 2010).

Another most widely used tool Cufflinks-Cuffdiff (Trapnell 2013) upgraded to Cuffdiff2 provides more determined method for differential expression analysis at transcript level. The newer version Cuffdiff2 uses negative binomial model and provides FPKM reads after normalization using relative log expression and inter-sample normalization method Q (Trapnell 2013).

Normalization of read counts is one of the critical steps in differential analysis of RNAseq data. The primary step in this process is to equate the total read counts from different libraries, as the variation caused by sequencing depths and size of the library are not comparable directly. In association to the number of expressing reads and gene length, the expression analysis also depends on the sample RNA that is being processed. For instance, genes with high expression shares a large percentage of the total reads of the sample compared to the left-over reads. This could be compared to the samples where reads are distributed evenly, in which case these lowly expressed genes show false positive result of differential expression for those genes (Zyprych-Walczak et al. 2015).

12.6 Annotation

12.6.1 Functional Annotation

The output of differential gene analysis provides information for the altered expression level of particular set of genes, now the next step is to explore the biological function of the genes. This is done by analysing the functional aspects, interaction network, pathway analysis and gene ontology, etc. of the genes involved in different processes of the biological system.

For functional annotation of the genes, various databases such as PANTHER classification system (Mi et al. 2016), DAVID Gene Functional Classification Tool (Sherman et al. 2007), etc. are available which assign particular function to genes and categorize them into different protein classes and biological pathways based on their over-representation analysis (ORA) in the data (Khatri et al. 2012). Based on similar biological functions, cellular localization and pathway annotation these genes are classified into different functional categories. The genes are analysed for their over-representation in the particular category by calculating their occurrence in the specific category compared to the proportion of genes accommodated in the same category. The results can further be evaluated for significant results by applying statistical tools such as Fisher's exact test, Hypergeometric correction, etc.

12.6.2 Pathway Analysis

Annotation of differentially expressed genes to different pathways ensues to offer biological insights of genes based on their functional and structural similarities. Few methods of pathway annotation involve categorization of genes into different pathways irrespective of the mechanistic model of the pathway (Zhao et al. 2016). Another method involves analysis of certain genes enriched more than the expected count. This is known as pathway enrichment analysis which provides more functional understanding to the gene sets obtained from sequencing data. Here, the over-represented pathways are identified with strong statistical significance, such as FDR (False Discovery Rate) and p-value, relative to the expected chance of occurrence, using ranking score, overlapping genes over the size of the pathway and pathway topology, etc.

Some databases identify the enriched genes by assigning a scoring system based on their position and interaction amongst other genes in the network. Resultantly, interacting genes obtain higher weightage compared to the non-interacting ones, showing the functional relatedness of few sets of genes (Zhao et al. 2016). The analysis involves identification of set of genes from the sequencing data, selection of statistically significant enriched pathways and visualization and graphical representation of the results.

12.6.3 Gene Ontology (GO) Analysis

Gene Ontology analysis is a method to distribute genes into hierarchical classification and their representation in graphical structure. GO classification is distributed into different terms in which the genes or gene products get distributed into Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). These GO terms can be defined as:

- *Biological Process*—defines the role of the genes in the biological processes of an organism, such as, transcription, translation, signalling, apoptosis, etc.
- *Molecular Function*—provides the information related to functional activity of the gene in molecular terms. These activities include protein binding, nuclease activity, protease activity, etc.

- *Cellular Component*—provides information for cellular localization of the gene product. This includes components such as nucleus, lysosome, plasma membrane, etc.

The GO terms are said to be loosely hierarchical based on the available information regarding their biological functions and localizations. Based on this information they can be arranged in terms of 'parent terms' or more specific 'child terms'.

GO analysis also provides information for genes that are over- or under-regulated under specific conditions. This is done by calculating the enrichment analysis for the over-representation of certain set of genes in a particular condition (Gene Ontology Consortium@2015). The results are statistically evaluated based on their p-values. Various tools such as WebGeStalt (Wang et al. 2013), Clusterprofiler (Yu et al. 2012), Gorilla (Eden et al. 2009), WEGO (Ye et al. 2006), etc. are widely used.

12.7 Other RNAseq Applications

12.7.1 Single Cell RNAseq

RNAseq provides information for expression profile for a population of millions of cells. But different population of cells behave distinctly in different tissues. Single cell RNAseq is a recently developed technique designed to explore the distinct expression profile of single gene entity. Several tools have been designed to improve the procedural factures in employing this technique, such as dividing and disintegrating the cells to obtain single cell molecule (Zappia et al. 2018).

Since transcriptomic profiles of bulk samples provide a comprehensive outlook of bulk population of cells, single cell RNA sequencing meant to decipher the distinctiveness of cells at individual level. This approach is an addition to identify distinguishing variations in gene expression which are more complex and understanding of biological diversities in cellular context. Different approaches are being used to achieve unbiased, high throughput single cell RNAseq with exhaustive quantitative information at individual scale (Avital et al. 2014). One such approach is droplet based single cell RNAseq, developed independently by Klein et al. (2015) and Macosko et al. (2015). This technology is based on identification of single cells by barcoding individual cells from bulk of cells and analysing them using high throughput sequencing.

Another approach developed recently for single cell RNAseq is based on differential analysis of discrete expression pattern in different biological conditions. The approach developed by Korthauer and his team uses simulated data to detect the variations in the differential patterns under given set of biological conditions using a modelling framework (Korthauer et al. 2015).

12.7.2 Small RNA Sequencing

Small RNAs, such as siRNA (small interfering RNA), miRNA (microRNA), etc. belong to class of non-coding RNAs that plays crucial roles in regulation of gene expression at transcriptional level. The developing technologies in high throughput sequencing opened new prospects to explore the world of the miRNAs (Sharma@2020). Despite their pivotal roles, miRNAs share very less percentage in the genome. In order to obtain a comprehensive profile of miRNAs, deep sequencing is performed which is a modified version of next generation sequencing, sequencing a genomic region hundred or thousand times and allowing to detect molecules present in rare volumes (Motameny et al. 2010).

Currently, only a small number of tools and pipelines are available for analysis of miRNA data which is also a major challenge faced by many researchers. The analysis of miRNA data involves:

(a) Pre-processing of the raw data to filter out low-quality reads and other non-coding RNAs such as rRNA, tRNA, snRNA, snoRNA, etc.
(b) Mapping of reads to miRbase (largest repository of published miRNA sequences and annotations of various organisms) (Griffiths-Jones et al. 2007) to obtain known or conserved miRNAs in an organism.
(c) Prediction of novel miRNAs in an organism based on generation of hairpin loop structure using an RNA folding algorithm.
(d) Quantification of miRNAs for detection of differentially expressing miRNAs.

Further, these miRNAs regulate expression of various genes by binding to the 3’UTR (untranslated region) of their target mRNAs with near specific complementarity. Based on the complementarity between miRNA and target mRNAs various tools have been developed to detect the potential targets of candidate miRNAs using different algorithms. Tools such as microrna.org (Betel et al. 2008) and TargetScan (Lewis et al. 2005) account for detection of target mRNAs by searching for the binding sites for specific miRNAs. Few other tools such as Pictar (Lall et al. 2006), RNAhybrid (Rehmsmeier et al. 2004), miTarget (Kim et al. 2006), miRDB (Wong and Wang 2015), DIANAmicroT (Maragkakis et al. 2009) also predict putative binding mRNAs for given miRNAs using different algorithms in the background.

Identification of target mRNAs also accounts for involvement of these target mRNAs in different molecular processes and significant pathways, which is done by functional annotation, gene ontology and pathway analysis, etc. This could provide information for miRNA-mRNA regulatory network and can further be exploited for disease aetiology and therapeutic interventions.

12.8 Concluding Remarks

The rapid increase in technological expansion in the current times resulted in a tremendous upsurge of NGS technologies such as DNA sequencing, RNA sequencing and other targeted sequencing projects (Sharma et al. 2016). But to translate the data generated from sequencing, the prime requisite is development of appropriate, specialized and reliable tools and bioinformatics applications. RNA sequencing is an advanced technique of NGS which favours the quantification and presence of RNA content in the biological sample. It also infers the presence of post-transcriptional modifications, SNPs, mutations, alternative spliced transcripts and their association with disease pathogenesis (Conesa et al. 2016). The use of RNAseq technology for various applications on a massive scale also demands for development of computational tools and softwares, with significant and reliable results, to match the pace by analysis and interpretation of data parallelly.

However, RNAseq is a gold standard technique to generate a comprehensive profile of whole transcriptome and other small non-coding RNAs in the sample. It is also highly prone to biasness and discrepancies in the data due to RNA extraction process, fragmentation of RNA, cDNA synthesis, amplification and sequencing, etc. Hence, to avoid these inconsistencies various tools and pipelines have been developed, based on different algorithms, to avoid the artefacts generated at various steps during the process. Data normalization is one such step which is crucial to reduce the biasness in the data. Several researchers deliver different thoughts on using different tools for data normalization and to minimize the noise and obtain best possible results.

Furthermore, different analysis tools offer varied results depending on the algorithms and backend procedures they are based on, hence relying on single tool cannot be recommended to provide substantial results. Therefore, it is always advisable to go through different school of thoughts and use multiple tools to attain comprehensive and comparative values for conclusive considerations.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data

Ansorge WJ (2009) Next-generation DNA sequencing techniques. New Biotechnol 25:195–203

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Res 32:D115–D119

Avital G, Hashimshony T, Yanai I (2014) Seeing is believing: new methods for in situsingle-cell transcriptomics. Genome Biol 15:110

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477

Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008) The microRNA.org resource: targets and expression. Nucleic Acids Res 36:D149–D153

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD (2016) rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. Bioinformatics 32:2210–2212

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD (2019) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 8:giz100

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17:13

Deorowicz S, Grabowski S (2011) Compression of DNA sequence reads in FASTQ format. Bioinformatics 27:860–862

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10:1–7

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Freedman A (2016) Best practices for de novo transcriptome assembly with trinity

Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8:469–477

Geniza M, Jaiswal P (2017) Tools for building de novo transcriptome assembly. Curr Plant Biol 11:41–45

Ghosh S, Chan C-KK (2016) Analysis of RNA-Seq data using TopHat and Cufflinks. In: Plant bioinformatics. Springer, New York, pp 339–361

Gilbert D (2003) Sequence file format conversion with command-line Readseq. Curr Protoc Bioinformatics 00(1):A-1E.1–A-1E.4

Glebova O, Temate-Tiagueu Y, Caciula A, Al Seesi S, Artyomenko A, Mangul S, Lindsay J, Măndoiu II, Zelikovsky A (2016) Transcriptome quantification and differential expression from NGS data. In: Computational methods for next generation sequencing data analysis. Wiley, Hoboken, NJ, pp 301–327

Gordon A, Hannon G (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpublished. http://hannonlab.cshl.edu/fastx_toolkit

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol 29:644

Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ (2007) miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154–D158

Han Y, Gao S, Muegge K, Zhang W, Zhou B (2015) Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights 9:BBI-S28991

Hedges DJ, Guettouche T, Yang S, Bademci G, Diaz A, Andersen A, Hulme WF, Linker S, Mehta A, Edwards YJ (2011) Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. PLoS One 6:e18595

Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 8:e1002375

Kim S-K, Nam J-W, Rhee J-K, Lee W-J, Zhang B-T (2006) miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics 7:1–12

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201

Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28:3211–3217

Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C (2015) scDD: a statistical approach for identifying differential distributions in single-cell RNA-seq experiments. bioRxiv 035501

Lall S, Grün D, Krek A, Chen K, Wang Y-L, Dewey CN, Sood P, Colombo T, Bray N, MacMenamin P (2006) A genome-wide map of conserved microRNA targets in C. elegans. Curr Biol 16:460–471

Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics 32:11–17

Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120:15–20

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

Li W, Jiang T (2012) Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. Bioinformatics 28:2914–2921

Li Z, Xuejun L (2016) A comprehensive review on RNA-seq data analysis. Trans Nanjing Univ Aeronaut Astronaut 33(3):339–361

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li W, Feng J, Jiang T (2011) IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. J Comput Biol 18:1693–1707

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214

Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 37:W273–W276

Maretty L, Sibbesen JA, Krogh A (2014) Bayesian transcriptome assembly. Genome Biol 15:501

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12:671–682

Merriman B, Ion Torrent R&D Team, Rothberg JM (2012) Progress in ion torrent semiconductor chip based sequencing. Electrophoresis 33:3397–3417

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc 2010:pdb-prot5448

Mezlini AM, Smith EJ, Fiume M, Buske O, Savich GL, Shah S, Aparicio S, Chiang DY, Goldenberg A, Brudno M (2013) iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Res 23:519–529

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD (2016) PANTHER version 11: expanded annotation data from gene ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res 45:D183–D189

Motameny S, Wolters S, Nürnberg P, Schumacher B (2010) Next generation sequencing of miRNAs–strategies, resources and methods. Genes 1:70–84

Mulcare D (2004) NGS toolkit. Part 8: the National Geodetic Survey. NADCON tool. Prof Surv Mag 24(2):120–125

Nakasugi K, Crowhurst R, Bally J, Waterhouse P (2014) Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant Nicotiana benthamiana. PLoS One 9: e91776

Niemenmaa M, Kallio A, Schumacher A, Klemelä P, Korpelainen E, Heljanko K (2012) Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. Bioinformatics 28:876–877

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13:1–13

Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10:1507–1517

Roberts A, Pimentel H, Trapnell C, Pachter L (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27:2325–2329

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092

Sharma P, Bhunia S, Poojary SS, Tekcham DS, Barbhuiya MA, Gupta S, Shrivastav BR, Tiwari PK (2016) Global methylation profiling to identify epigenetic signature of gallbladder cancer and gallstone disease. Tumor Biol 37:14687–14699

Sharma P, Kumar S, Beriwal S, Sharma P, Bhairappanavar SB, Verma RJ, Das J (2020) Comparative transcriptome profiling and co-expression network analysis reveals functionally coordinated genes associated with metabolic processes of Andrographis paniculata. Plant Gene 23:100234

Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA (2007) The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8:R183

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 26:1134–1144

T O'Neil S, Emrich SJ (2013) Assessing De Novo transcriptome assembly metrics for consistency and utility. BMC Genomics 14:465

Tomescu AI, Kuosmanen A, Rizzi R, Mäkinen V (2013) A novel min-cost flow method for estimating transcript expression with RNA-Seq. BMC Bioinformatics 14(Suppl 5):S15

Trapnell C (2013) Cufflinks. cuffdiff (v6). Open module on GenePattern public server. GenePattern. https://software.broadinstitute.org/cancer/software/genepattern/modules/docs/Cufflinks.cuffdiff/6

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515

Voshall A, Moriyama EN (2018) Next-generation transcriptome assembly: strategies and performance analysis. In: Bioinformatics in the era of post genomics and big data. IntechOpen, London, pp 15–36

Wang J, Duncan D, Shi Z, Zhang B (2013) WEB-based gene set analysis toolkit (WebGestalt): update 2013. Nucleic Acids Res 41:W77–W83

Wang Y, Hu H, Li X (2017) rRNAFilter: a fast approach for ribosomal RNA read removal without a reference database. J Comput Biol 24:368–375

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543–548

Wong N, Wang X (2015) miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res 43:D146–D152

Wyrzykiewicz T, Cole D (1994) Sequencing of oligonucleotide phosphorothioates based on solid-supported desulfurization. Nucleic Acids Res 22:2667–2669

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34:W293–W297

Yu G, Wang L-G, Han Y, He Q-Y (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 16:284–287

Zappia L, Phipson B, Oshlack A (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol 14:e1006245

Zhao S, Zhang B, Zhang Y, Gordon W, Du S, Paradis T, Vincent M, von Schack D (2016) Bioinformatics for RNA-seq data analysis. Bioinformatics—updated features and applications. InTechOpen, London, pp 125–149

Zyprych-Walczak J, Szabelska A, Handschuh L, Górczak K, Klamecka K, Figlerowicz M, Siatkowski I (2015) The impact of normalization methods on RNA-Seq data analysis. Biomed Res Int 2015:621690

Computational Metabolomics

13

Priya Ranjan Kumar, Santosh Kumar Mishra, and Sarika Srivastava

Abstract

Metabolomics is a comprehensive and systematic determination of different metabolite levels, their interactions and dynamics from the complete set of small molecule called metabolome in any biological system. Metabolome is the complete set of metabolite present in any organism. Thousands of metabolites may be present in any animal or plant metabolome. Extracting biological information from a large metabolomics dataset is a big challenge in the field of metabolomics. Sample preparation and its classification, identification, and estimation of the quantity of individual metabolite, etc. are some of the important challenges. Rapid improvements in Nuclear Magnetic Resonance (NMR)-based methods, Mass Spectroscopy (MS), computer software and hardware which can handle large dataset leads to the development of high-throughput metabolomics methods. The pipeline for metabolomics data processing is discussed in this chapter.

Keywords

Metabolomics · Metabolites · Metabolome · High-throughput · Data processing

P. R. Kumar (✉) · S. K. Mishra
Department of Biotechnology, IMS Engineering College, Ghaziabad, Uttar Pradesh, India

S. Srivastava
School of Biosciences, IMS Ghaziabad (University Courses Campus), Ghaziabad, Uttar Pradesh, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_13

13.1 Introduction

Small molecules which are produced during the metabolic reactions due to enzymatic activities are called metabolite. These metabolites also take part in various other catabolic and anabolic reactions and many a time required for normal growth and development of cells. Entire metabolites present in an organism are called metabolome (Oliver et al. 1998). Likewise, metabolomics refers to the use of scientific methods to find and quantitatively estimate all metabolites in any organism or bio-system, as well as the monitoring of changes in the metabolome of total plant or organism. In a metabolomics experiment, following steps are performed (Goodacre et al. 2007):

1. Design of the experiment.
2. Data and associated metadata storage.
3. Data preprocessing and processing.
4. Data analysis and interpretation.

The rapid improvement in methods based on MS, NMR, and computational tools which are capable in big data processing leads to significant improvement in high-throughput metabolomics methods (Wen and Zhu 2015).

In 1990s, GC/LC-MS (gas-chromatography mass spectrometry) was the choice of technology for the analysis of metabolites in the wide range of plant species. This leads to the development of many metabolomics spectral libraries which are currently used by many tools and software for the identification of metabolites from the spectra of new bio-samples. One such metabolomics spectral library is “GOLM Metabolome Database” (Kopka et al. 2004). Now a day, GC/LC-MS, NMR and Electron microscopy techniques are used for the identification of metabolites from bio-samples. Among all these techniques, NMR is a widely used technique in metabolomics and is becoming increasingly popular in this field. All metabolomics experiments generate a complex and very big dataset. Handling and processing of these datasets for the identification of metabolites is very complex and a big challenge in this area (Boccard et al. 2010). These datasets are generated from various experimental methods like NMR or GC/LC-MS, etc. These instrumental datasets need to be initially preprocessed to get clean dataset as they contain many biases and noises. This clean data need to be further processed using different tools and metabolite libraries for the identification of metabolites (Barnes et al. 2016). Quantitative and chemometric approaches are the two major means of metabolomics data analysis. The first approach facilitates the quantitative estimation of entire metabolites present in bio-sample using spectral libraries prior to statistical analysis of data. Whereas, in chemometric approaches the intensity of spectra and its patterns are initially recorded and then it is statistically compared to find spectral features (Xia et al. 2009). The processing pipeline of NMR data for metabolite identification from a given sample is explained in this chapter.

13.2 Factors Influencing Variations and Redundancy in Metabolomics Data

There are different factors which influence the variation or redundancy in the metabolomics data. For example, differences in orders of magnitude among the concentration of measured metabolites. Many a time, it has been observed that the metabolites which are present in low concentration are more important one than the metabolites present in higher concentration. Other factors include the differences in the fold changes among metabolite concentration due to the induced variation resulting into the larger differences in metabolite concentration depending on the environmental conditions. Under the same experimental conditions, large fluctuations in concentration can be seen for some metabolites. These types of biological variations are called uninduced variation. Variations arising due to the sampling errors, analytical errors, etc. are called technical variations. For data analysis and accurate results, total uninduced and technical variation should be zero. This is, however, not always possible and hence data filtering and preprocessing is required prior to data processing and metabolite identification (van den Berg et al. 2006).

13.3 NMR Spectroscopy for Metabolite Identification

Various compounds available in a complex mixture can be identified and quantified using NMR using (De Meyer et al. 2010). Since, most of the signals heavily overlap in 1D NMR spectra, makes it is very complex to interpret. These spectra are simplified with additional spectral dimensions, which also facilitate in obtaining extra information. In metabolomics experiments we generally go till two dimensions. This may be of two types: Homonuclear experiments and Heteronuclear experiments.

13.3.1 Homonuclear Experiments

Homonuclear NMR experiment is the one where recorded dimensions span chemical shifts of the same type of nucleus. Examples are COSY (COrrelated SpectroscopY), TOCSY (TOtal Correlated SpectroscopY), NOESY (Nuclear Overhauser Effect SpectroscopY), and ROESY (Rotational nuclear Overhauser Effect SpectroscopY) (Keeler 2010; Gheysen et al. 2008).

13.3.2 Heteronuclear Experiences

The Heteronuclear NMR experiment is the one where recorded dimensions span chemical shifts of different types of nucleus. It is used to assign the spectrum of another nucleus once the spectrum of one nucleus is known. Examples are HSQC

(Heteronuclear Single Quantum Coherence) and HMQC (**H**eteronuclear **M**ultiple **Q**uantum **C**oherence) (Bodenhausen and Ruben 1980).

NMR spectra obtained from samples of biological extracts used to contain overlapping signals in the magnitude of thousand from large numbers of molecules. Analysis of these overlapping spectra for the qualitative and quantitative estimation of individual metabolite is the major challenge of metabolomics studies (Lewis et al. 2009). Good result can only be obtained from any metabolomics studies using NMR, by proper data collection and handling, data filtering/preprocessing, and data analysis (Wang et al. 2009). The complex nature of NMR data requires data preprocessing as an initial step for further analysis. It transforms and clear data from bias or noise obtained from any uninduced and technical variations to facilitate a more accurate and robust data analysis (Izquierdo-García et al. 2009). It mainly includes various steps like Fast Fourier Transformation, Phasing of spectra, Baseline correction, Normalization, Binning/Bucketing, etc.

There are plenty of software available for NMR data processing and analysis. Some of them are Chenomx NMR Suite, NMRPipe, AMIX, Hires, Automics, KnowItALL, etc. (Wang et al. 2009). Along with these software tools, few R-based NMR spectra processing packages are also available like NMRs, ChemoSpec, rNMR, etc. These R-packages provide both GUI and command line interface for the spectra processing and analysis.

After the data processing, the next step is to identify the metabolite present in the sample. This is done by generating the peak list from the processed dataset. With this peak list, one needs to search through the metabolite library to identify the metabolite. There are various metabolite libraries available for example: BMR Data Bank, MMC Database, NMRShiftdb, WebSpectra, SDBS, etc. (Ellinger et al. 2013).

13.4 Organization of 1D and 2D Dataset and their Format

Bruker Corporation is the largest vendor for the NMR instrument. Hence the “Bruker fid” is a standard data file format for the NMR data. Other than this, JCAMP-DX, ASCII, CSV, Varian VNMR, Joel, Simplot, etc. are few other widely used file format. The Bruker fid dataset contains many different files with it. It stores one scanning session in the directory with name as per the subject and its session. It belongs to its own directory. Within a single session directory of an experiment, there are:

1. A text file with name “subject,” contains information about the experiment.
2. 1, 2, 3, etc. numbered subdirectory containing acquisitioned data for each saved run of that session.

Each directory contains following files:

1. A text file with name “imnd.” It contains various parameters which are used in data acquisition from the scanner.

Table 13.1 1D and 2D datasets in pdata subdirectory of Bruker's data format

1D dataset		2D dataset	
Acqus	Acquisition parameters	acqus	Acquisition parameters F2
procs	Process parameters	procs	Process parameters F2
fid	Raw fid data points	acqu2s	Acquisition parameters F1
1r	Spectrum real part data points	proc2s	Process parameters F1
1i	Spectrum imaginary part data points	ser	Raw ser data points
		2rr	2D spectrum real part data points
		2ii	2D spectrum imaginary part data points

2. Another text file with name "acqs," which is also an Acquisition Parameter file containing details about that run.
3. Raw Free Induction Decay (FID) data used to be present in a very big binary. A single file having name "fid" for 1D NMR and multiple "fid" files along with a serial "ser" file in 2D NMR experiments use to be present in the directory.
4. Various other files, including the "pulseprogram," "log," and "grdprog.r" gradient programs.
5. A subdirectory "pdata" (Processed Data) that contains any reconstructions of the data.

Subdirectories numbered 1, 2, 3, etc. may be found within each of the "pdata" subdirectories for each new reconstruction of the raw data. Files of these subdirectories are listed in Table 13.1 for 1D and 2D datasets.

We can convert Bruker file into ASCII and further in to simple CSV file for the ease of understanding and processing using various tools and R-packages.

13.4.1 GSim: Bruker FID File into ASCII File Converter

Open an FID data file after downloading and installation of GSim software. To do this one need to select Data option available in Edit Table A spreadsheet is built by GSim tool to select and use real and imaginary FID data parts. This file can be saved by pressing ctrl+s as .ascii format.

13.4.2 Converting ASCII File into CSV File

Open the .ascii file in MS-Excel by choosing all file type (*.*). Click next button and select space check box. Chose dot as decimal separator by clicking on advance button and then click OK. Finally, click on "Finish" button to save opened file in CSV format.

13.5 Data Preprocessing Methods

Preprocessing of NMR data usually aims to reduce variances and influences as phase corrections of each spectrum, baseline corrections, etc. For NMR data, preprocessing procedures include Fourier transformation of the raw FID, phasing, noise filtering, baseline correction, normalization, and conversion to magnitude spectra (Goodacre et al. 2007). For chemical shift variability, NMR peak alignment, whether global or local is also regarded as preprocessing.

13.5.1 Fourier Transformation of FID

In NMR, oscillating signal decays exponentially as a function of time as the phase coherence between the magnetic dipoles. This oscillating signal is called the FID, which represents the signal in the time domain. In FID data, signal amplitudes are represented as a function of time which need to get converted into a spectrum where one axis is frequency instead of time (Duer 2004). For this conversion, a mathematical approach is used called "Fourier Transformation." It converts FID data into frequency domain (the spectrum of amplitude versus frequency) as shown in Fig. 13.1. The frequency domain spectrum contains two parts., which are, real and imaginary. An "absorption mode line" is obtained from real part of data and "dispersion mode line" is obtained by the imaginary part of data.

13.5.2 Phase Correction or Phasing

Absorption line does not appear sometime in the real part of spectrum which is very undesirable as it is required for best resolution spectrum. The real part (S_x) of the FID used to be a damped cosine wave and the imaginary part (S_y) a damped sine wave. The spectrum obtained by Fourier transformation contains the real part having the absorption mode line shape and the imaginary part the dispersion mode. But due to the effect of phase shift of around 45°, both real and imaginary sections of spectrum get mixture of absorption and dispersion lines. Sometime due to more phase shift of

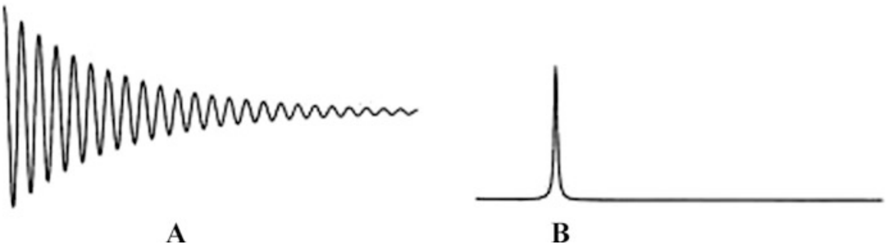


Fig. 13.1 (**a**) example of a free induction decay with only one frequency component where frequency decay exponential with time; (**b**) Fourier-transformed frequency spectrum of FID

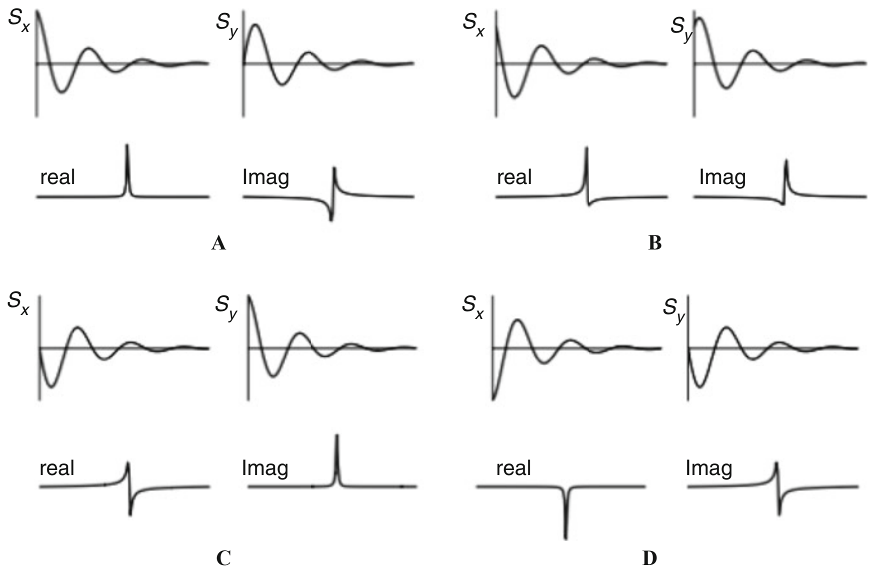


Fig. 13.2 Spectrum showing the phase shift effect. (**a**) Normal phase; (**b**) Phase shift of 45°; (**c**) Phase shift of 90°; (**d**) Phase shift of 180° (Keeler 2004)

90° real part takes the form of a damped sine wave, whereas imaginary part takes the form of damped cosine wave. Negative absorption line may appear in the real part due to the phase shift of 180°. (Keeler 2004). The effect of different phase shifts on time domain signal is shown in Fig. 13.2.

In practice, the real part of the spectrum is displayed after Fourier transformation of FID spectrum. Further, the phase is adjusted to make it correct until the spectrum appears to be in the absorption mode. The whole process is called phasing the spectrum.

13.5.3 Noise Filtering

Many a time, noises are also get recorded while recording FID. Major contributors of these noises are the amplifiers, some of the electrical components of spectrometer and thermal noise of the signal detector coil. Due to this, while FID decays with respect to time, noises continue to get recorded. Hence, if we record spectrometer data for a long time, there may be only noise in the later part of data instead of actual signal. A weak SNR (signal-to-noise ratio) may be expected in the resulting spectrum. Since, actual metabolite signals used to be present in early part of FID, hence by reducing the spectrometer data recording time, SNR may be improved (Keeler 2004). But, one need to be careful while shortening the data acquisition time so that they do not miss the actual FID data as shown in Fig. 13.3.

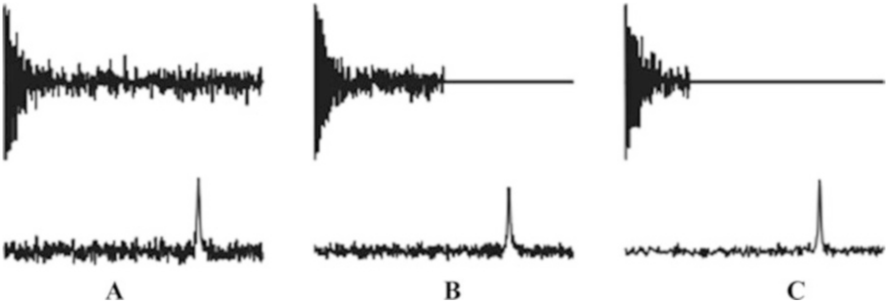


Fig. 13.3 Time dependency of SNR in spectrum while recording FID. (**a**) A long FID data recording time contributes more noise. (**b**) Low level of noise when data recording time is reduced to half. (**c**) taking the first quarter of the data (Keeler 2004)

Looking at the FID spectrum as shown in Fig. 13.3, it can be concluded that, only the starting parts of FID contain actual signal. Hence, a mathematical function can be used which can cut off the remaining part of data without affecting the starting part in order to improve SNR in the spectrum.

13.5.3.1 Baseline Correction

In 1D NMR spectra sometime, few initial data points in FID are corrupted which leads to the distortion in baseline. The main reason for this is the low frequency modulation due to the corrupted data points in the Fourier-transformed spectra. In NMR spectra, any distortion must be corrected as they reduce the values of intensity. This also creates problems in peak alignment and quantitative estimation of metabolites. Sometimes, many small peaks may also be very significant. These peaks may be sensitive to any distortion in baseline (Xi and Rocke 2008). There are two methods for baseline correction. These are time domain correction and frequency domain correction. The low frequency modulation is reduced by reconstructing the damaged data points in FID using time domain correction method. Whereas, the baseline curves are corrected directly using frequency domain correction method and it is subtracted to remove distortion.

13.5.4 Peak Alignment

NMR analysis of biofluid samples is often associated with the variations in the position of peak and its shape not correlated with the sample. The main reason for this is the instability of the instrument and the variations in the sample background matrix. The NMR data analysis and interpretation gets complicated due to these variations. Hence, peak alignment step is the must to follow preprocessing method to remove these complications before any further analysis step. The peak alignment is performed by shifting the spectrum sidewise and comparing it with the reference spectrum until the best correlation is not found (Forshed et al. 2003).

13.5.5 Binning/Bucketing

The chemical shift variability across spectra is corrected by this method. The entire spectra are segmented into small bins and then spectrum under each bins are taken for further processing. The appropriate bin size is used so that the spectral peak of one compound remains in a single bin despite small spectral shifts across the spectra. For this, the size of bins needs to be specified either in ppm or must fix total number of bins (Cobas 2011). Most preferably binning is performed from 0.04 to 10 ppm with a bin size of 0.04 ppm. Signals that are unrelated to the experiment, called the dark regions can be excluded from binning. An example of binned spectra is shown in Fig. 13.4.

13.5.6 Normalization

Ideally, the metabolite concentration used to be directly proportional to the intensities of 1H-NMR peak. Hence, it is useful in biomarker discovery and metabolite class prediction. However, peak intensities can be affected by many uninduced variables such as instrumental, experimental, etc. The influence of such variations can be minimized by using normalization methods. These are used in order to produce robust and reproducible analysis. Normalization of binned spectra can be done by constant sum method (Torgrip et al. 2008). But, this method may have some limitations such as, the metabolites present in abundance in the sample may influence the scaling of remaining metabolites. Hence, other approaches for data normalization may be used, such as, a reference sample may be selected to normalize entire data. This may reduce scaling artifacts which get generated using constant sum method. Data variation in the sample can be estimated and reduced by calculating the quotient median among all data points between target and reference spectrum.

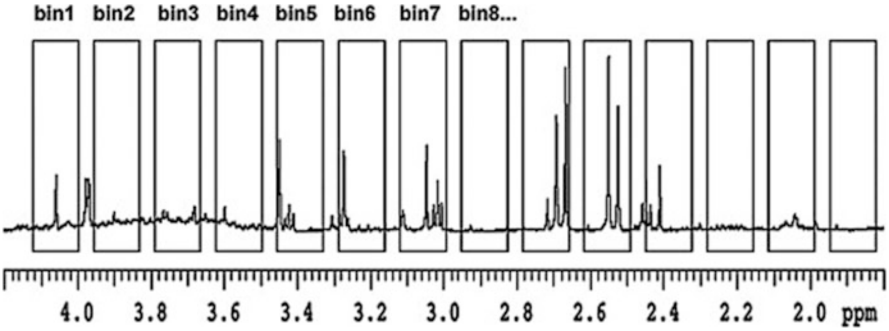


Fig. 13.4 Binned NMR spectra

13.6 Data Processing Tools/R-Packages

There are a few freely available tools/R-packages which can help in 1D NMR data processing and analysis. They are:

- NMRS,
- ChemoSpec,
- speaq,
- batman.

For 2D NMR data processing and analysis following tools can be used:

- NMRPipe (An UNIX based software for data processing and analysis),
- rNMR (An R-based tool for data analysis).

13.6.1 NMRS

Bruker FID format spectral data can be loaded directly into the NMRS tool for its analysis. The reference of spectrum can also be displayed using this tool. It can also be used to perform many basic operations like phase correction, chemical shift adjustment of certain compound to zero ppm, baseline correction and selection of spectral area. The NMRS package has been designed as an interactive process. By typing NMRS the user can have access to complete preprocessing of the data. NMRS package is dependent upon few other R-package and requires the combination of Tool Command Language (Tcl) and Tk GUI toolkit which is referred to as Tcl/Tk. Hence before installing the NMRS package, one need to configure their machine with Tcl development package (tcl-dev) and Tk development package (tk-dev). In order to perform Time–Frequency analysis, one also needs to install another R-package "R-wave." The "tkrplot" package is required for placing R graphics in a Tk widget; and "FTICRMS," to handle large matrices and data visualization (Izquierdo 2013).

13.6.2 ChemoSpec

Spectroscopic data analysis can be performed using ChemoSpec tool kit. The spectra can be plotted using specific functions of this package. Many exploratory data analysis can be performed using this tool such as PCA (Principal Component Analysis), Model based clustering, and HCA (Hierarchical Cluster Analysis). The comparison between control and treatment group samples can also be performed using this tool (Hanson et al. 2020). The package has only command line interface so it runs on R console. This tool depends upon several other R-packages such as R. utils, plyr, amap, baseline, pls, etc. Since this package is also depending upon "rgl" package for 3D visualization so it requires OpenGL support. Function "getManyCsv" helps in moving raw data sets into ChemoSpec, and it supports only . csv file format containing data in two columns containing frequency and intensity values. ChemoSpec package has a large set of functions which can help in data preprocessing and analysis.

13.6.3 Speq

Metabolite quantification from NMR spectra can be done using "Speq" tool which uses CluPA (Cluster-based Peak Alignment) method for peak alignment. This R-package aligns reference spectrum with the target spectrum using top–bottom approach and builds a cluster tree. Further, the spectra get divided into small segment on the basis of the farthest cluster. It also carries out different statistical analyses like F-statistic or a one-way ANOVA to quantify the NMR data. This package does not have any data preprocessing option (Beirnaert et al. 2019).

13.6.4 BATMAN (Bayesian AuTomated Metabolite Analyzer for NMR)

"Batman" is an R-package, which can automatically quantify metabolites signals present in the NMR spectra. This package helps in the deconvolution of peaks from 1D NMR spectra and estimates concentration of specific metabolites in the target list. The Bayesian model includes metabolite characteristic peak patterns and can easily identify any shift from the native position of peaks. Peak shifting is usually very common in NMR spectra. This tool lacks any data preprocessing function (Hao et al. 2012).

13.6.5 NMRPipe

NMRPipe is a UNIX based vast software program used for the NMR spectroscopic data processing and analysis. It helps in the processing and analysis of multidimensional NMR spectra. It need support of C-shell and X11 Graphics along with terminal window. It supports both GUI and command line interface with shell and TCL scripts. This makes it very flexible in nature where user can write their own scripts and run to perform any specific action on the spectral data (Delaglio et al. 1995). This package also includes many other programs like NMRDraw, NMRWish, DYNAMO, ACME, DC, etc. These programs help to perform several different types of actions like interactive processing of data, script editing, chemical shift analysis, peak detection, etc. A conventional processing pipeline using NMRPipe is as follows:

- Varian or Bruker raw spectrometer data to nmrPipe format conversion.
- Inspection of time-dependent data using nmrDraw package.

- Initial preprocessing of low resolution data like baseline correction, phasing, etc.
- Further inspection and if required additional processing.
- Automatic peak detection.
- Data analysis.

After data processing, the peak list generated from the tool can be uploaded directly on various metabolite libraries servers for example MMC database or BMR databank for metabolite profiling to identify various metabolites present in the sample.

13.6.6 rNMR

Identification and quantification of metabolites present in different spectra can be done using rNMR tool. It is an open source R-Package which provides user friendly GUI for 1D or 2D NMR spectra visualization and processing. There are two major ways through which user can interact with this tool. It also supports command line in R console based operation. Only UCSF spectra format is supported by rNMR tool. Hence, first of all, one must convert spectral data files into UCSF format before further processing or analysis. "cf()" is the supported file conversion function which can convert any spectral file into supported format. rNMR tool uses ROIs (Regions of Interest) based analysis method in which NMR data is distributed among defined range of chemical shifts. rNMR can help in the visualization and quantification of hundreds of spectra simultaneously (Lewis et al. 2009). MMC database is linked with its package using mmcd() function for metabolite identification using generated peak list.

13.7 NMR Spectral Libraries for Metabolite Identification

NMR Spectral library is used for bio-profiling to identify various metabolites present in sample, using peak lists generated from processed spectral data. There are various NMR base metabolite libraries available freely on World Wide Web in which we can submit our data to search for metabolite present in the sample. Few metabolite spectral libraries are as follows:

13.7.1 Madison Metabolomics Consortium Database *(MMCD)*

MS and NMR spectroscopy based metabolomics researches are highly dependent on MMCD resources hosted and maintained by Magnetic Resonance facility, Madison. The goal of this database is to support identification and quantification of various metabolites from MS and NMR spectra obtained from biological samples (Cui et al. 2008).

13.7.2 Biological Magnetic Resonance Data Bank (BMRDB)

NMR data of peptides, DNA, RNA, and other biomolecules are stored in this database and serves as reference library for metabolite identification. It is a member database of PDB (Ulrich et al. 2008).

13.7.3 NMRShiftDB

It is also a web based NMR database serves as reference library for many NMR data analysis tools. It has a large collection of structures of organic compounds along with their spectra. It is an open source database and available under the GNU Free Documentation License (Steinbeck and Kuhn 2004).

13.8 Conclusion

NMRS tool has very good options for NMR data preprocessing but it is limited to just data preprocessing task where as "ChemoSpec" has a wide range of data preprocessing and analysis functions, but the problem with this package is that it only accepts raw data in csv file format having two columns of frequency and intensity. So, the integration of some other function/script to generate the peak list with two column having ppm and intensity from the FID data is required. It also has all the data preprocessing options except Fast Fourier Transformation and phase correction because these are not needed to apply on intensity table. The intensity table is generated only after applying FFT on fid data using any other tool. These tools are specific for 1D NMR data. But in metabolomics 2D data are also being generated to get more robust and high resolution result. Hence tools/R-packages which can process the multidimensional NMR data are required. NMRPipe and rNMR are the tools which can handle 2D NMR data very well. NMRPipe is an UNIX based collection of various programs which allows user to interact either with command line mode or with GUI. NMRPipe is capable in handling data from 1D to 4D. It allows user to do all data processing and generate peak list from the spectra with which one can proceed for bio-profiling using various NMR spectrum libraries like MMCD, NMRShiftDB, etc. and further analysis. rNMR is R-based 1D and 2D NMR data analysis tools. It also provides command line interface and Graphical user interface to run various data analysis steps. As it was mainly developed for multiple NMR data analysis hence it does not have any data processing function. With this tool user can generate peak list for bio-profiling and select region of interest from multiple spectra to do comparative analysis.

References

Barnes S, Benton HP, Casazza K, Cooper SJ, Cui X, Du X, Engler J, Kabarowski JH, Li S, Pathmasiri W, Prasain JK, Renfrow MB, Tiwari HK (2016) Training in metabolomics research. II. Processing and statistical analysis of metabolomics data, metabolite identification, pathway analysis, applications of metabolomics and its future. J Mass Spectrom 51(8):535–548

Beirnaert C, Vu TN, Meysman P, Laukens K, Valkenborg D (2019) speaq: Tools for Nuclear Magnetic Resonance (NMR) Spectra Alignment, Peak Based Processing, Quantitative Analysis and Visualizations. R-Package, https://cran.r-project.org/web/packages/speaq/index.html

Boccard J, Veuthey JL, Rudaz S (2010) Knowledge discovery in metabolomics: an overview of MS data handling. J Sep Sci 33:290–304

Bodenhausen G, Ruben DJ (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett 69(1):185–189

Cobas C (2011) Alignment of NMR spectra – part II: binning/bucketing. Blog on NMR analysis. http://nmr-analysis.blogspot.com/2011/01/alignment-of-nmr-spectra-part-ii.html

Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, Markley JL (2008) Metabolite identification via the Madison metabolomics consortium database. Nat Biotechnol 26(2):162–164

De Meyer T, Sinnaeve D, Van Gasse B, Rietzschel ER, Buyzere MLD, Langlois MR, Bekaert S, Martins JC, Criekinge W (2010) Evaluation of standard and advanced preprocessing methods for the univariate analysis of blood serum 1H-NMR spectra. Anal Bioanal Chem 398 (4):1781–1790

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293

Duer M (2004) Introduction to solid-state NMR spectroscopy. Blackwell Publishing, Hoboken, NJ, pp 43–58

Ellinger JJ, Chylla RA, Ulrich EL, Markley JL (2013) Databases and software for NMR-based metabolomics. Curr Metabolomics 1(1). https://doi.org/10.2174/2213235X11301010028

Forshed J, Schuppe-Koistinen I, Jacobsson SP (2003) Peak alignment of NMR signals by means of a genetic algorithm. Anal Chim Acta 487(2):189–199

Gheysen K, Mihai C, Conrath K, Martins J (2008) Rapid identification of common hexapyranose monosaccharide units by a simple TOCSY matching approach. Chem Eur J 14:8869–8878

Goodacre R, Broadhurst D, Smilde AK, Kristal BS, Baker DJ, Beger R, Bessant C, Connor S, Capuani G, Craig A, Ebbels T, Kell DB, Manetti C, Newton J, Paternostro G, Somorjai R, Sjöström M, Trygg J, Wulfert F (2007) Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics 3:231–241

Hanson BA, Bostock M, Keinsley M (2020) ChemoSpec: exploratory chemometrics for spectroscopy. R Package. https://cran.r-project.org/web/packages/ChemoSpec/index.html

Hao J, Astle W, Iorio MD, Ebbels T (2012) BATMAN--an R package for the automated quantification of metabolites from NMR spectra using a Bayesian Model. Bioinformatics 28 (15):2088–2090

Izquierdo JL (2013) Package 'NMRS'. R-Package. http://www2.uaem.mx/r-mirror/web/packages/NMRS/NMRS.pdf

Izquierdo-García JL, Rodríguez I, Kyriazis A, Villa P, Barreiro P, Desco M, Ruiz-Cabello J (2009) Metabonomic: a novel R-package graphic user interface for the analysis of metabonomic profiles. BMC Bioinformatics 10:363

Keeler J (2004) Chapter 4: Fourier transformation and data processing. In: Understanding NMR spectroscopy lecture series. The James Keeler Group, Cambridge

Keeler J (2010) Understanding NMR spectroscopy, 2nd edn. Wiley, Hoboken, NJ, pp 280–299

Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmüller E, Dörmann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D (2004) GMD@CSB.DB: the Golm metabolome database. Bioinformatics 21(8):1635–1638

Lewis IA, Schommerand SC, Markley JL (2009) rNMR: open source software for identifying and quantifying metabolites in NMR spectra. Magn Reson Chem 47(1):123–126

Oliver SG, Winson MK, Kell DB, Baganz F (1998) Systematic functional analysis of the yeast genome. Trends Biotechnol 16(9):373–378

Steinbeck C, Kuhn S (2004) NMRShiftDB – compound identification and structure elucidation support through a free community-built web database. Phytochemistry 65(19):2711–2717

Torgrip RJO, Aberg KM, Alm E et al (2008) A note on normalization of biofluid 1D 1H-NMR data. Metabolomics 4:114–121

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent WR, Yao H, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36(Database issue):D402–D408

van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7:142

Wang T, Shao K, Ch O, Ren Y, Mu Y, Qu L, He J, Jin C, Xia B (2009) Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. BMC Bioinformatics 10:83

Wen B, Zhu M (2015) Applications of mass spectrometry in drug metabolism: 50 years of progress. Drug Metab Rev 47(1):71–87

Xi Y, Rocke DM (2008) Baseline correction for NMR spectroscopic metabolomics data analysis. BMC Bioinformatics 9:324

Xia J, Psychogios N, Young N, Wishart DS (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res 37(Web Server issue):W652–W660

Next Generation Sequencing

14

Anchita Prasad, Harshita Bhargava, Ayam Gupta, Nidhi Shukla, Shalini Rajagopal, Sonal Gupta, Amita Sharma, Jayaraman Valadi, Vinod Nigam, and Prashanth Suravajhala

Abstract

The next generation sequencing (NGS) technology refers to non-Sanger based DNA sequencing methods which have replaced conventional sequencing methods. They have been vividly used for analyses of complete genome (whole genome sequencing), the coding exons within already reported genes (whole exome sequencing), and only coding regions of selected genes (targeted panel). In this chapter, we give an introduction of NGS technology as well as a gist of different types and applications of NGS. As advancements in NGS data analysis have opened up new therapeutic opportunities for disease diagnosis, the complementary approaches such as machine learning algorithms used in NGS are subtly dealt at the end.

A. Prasad · V. Nigam
Department of Bioengineering, Birla Institute of technology, Mesra, Ranchi, Jharkhand, India

H. Bhargava · A. Sharma
Department of Computer Science, IIS University, Jaipur, Rajasthan, India

A. Gupta · N. Shukla · S. Rajagopal · S. Gupta
Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research (BISR), Statue Circle, Jaipur, Rajasthan, India

J. Valadi
Department of Informatics, Shiv Nadar University, Noida, India

Department of Computer Science, Flame University, Pune, Maharashtra, India

P. Suravajhala (✉)
Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research (BISR), Statue Circle, Jaipur, Rajasthan, India

Bioclues.org, Hyderabad, India
e-mail: prash@bisr.res.in

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_14

Keywords

Next generation sequencing · Whole exome sequencing · Whole genome sequencing · RNA-Seq · Machine learning

14.1 Introduction

Increased understanding about interpreting the human genome has provided critical evidence for genetic disorders as well as development of extensive treatment and diagnostic therapy strategies. This has had a face lift after sequencing efforts burgeoned in the last few decades. After the success of the human genome sequence project in the year 2004 (Abdellah et al. 2004), the growing need to sequence a massive number of genomes was lifted from traditional Sanger sequencing method to novel DNA sequencing techniques. In 2005, the first parallel DNA sequencing method appeared, ushering the new era of next generation sequencing (NGS) technologies (Shendure 2005). The NGS involves high-throughput and massively parallel sequencing technologies which has revolutionized the biological research. Further, evolving at a faster pace over the last few decades in terms of declining sequencing cost per base and high-throughput effects. Due to its high-throughput, scalability, and speed, NGS has enabled researchers in a wide array of biological and clinical applications (Abbasi and Masoumi 2020). Through NGS, millions of DNA reads are sequenced in a single assay at much lower cost. Owing to these advantages, NGS methods have been used for a wide range of applications including variant identification using whole genome/exome resequencing, transcriptome profile analysis of tissues, microbial profiling, and detecting genetic biomarkers for disease prognosis (Schuster 2008; Suravajhala et al. 2016). Over the last few years, increased awareness about decoding the human genome has provided significant evidence for detecting rare genetic disorders as well as their diagnosis and treatment in an efficient manner. In addition, for studying germline DNA and for analysis of cancer genome "massive" or "deep" sequencing techniques are applied (Pettersson et al. 2009; Stratton et al. 2009). On the other hand, information about transmission and outbreaks encompassing microbial genomes has also been determined, for example, inferring virulence, transmission, antibiotic resistance, and molecular sub typing. In addition, with NGS tracking the outbreak of Methicillin-resistant Staphylococcus aureus (MRSA) on neonates (Chiu et al. 2008), the main utility of NGS in microbiology has steadfastly replaced the conventional characterization of pathogens on various criteria with genomic features (Deurenberg et al. 2017). The bottom-line is that NGS has been a recommended strategy for characterizing various facets of organisms, viz. bacteria, viruses, fungi, yeast, and parasites. As for NGS, there is no requirement for target-specific primers as desired for Sanger sequencing, such techniques are available to researchers, practitioners, and academicians at a very reasonable cost and with higher accuracy (Di Resta et al. 2018). The NGS process for DNA sequencing has been explained in the following general steps (Fig. 14.1).

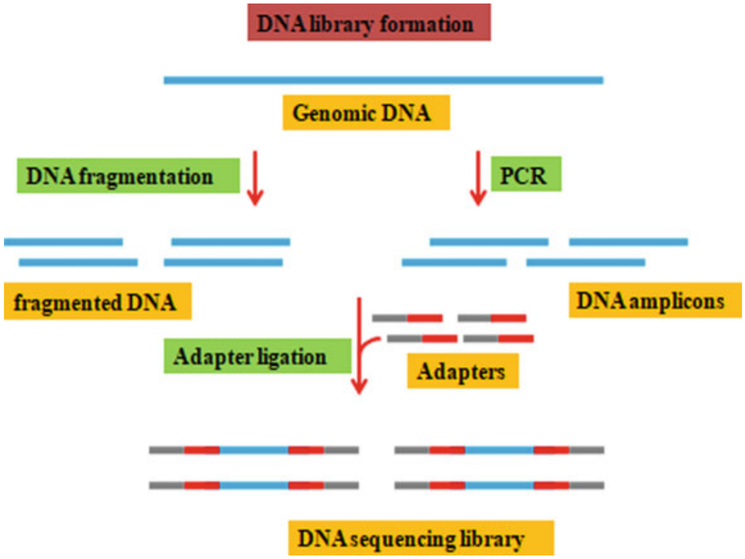


Fig. 14.1 Workflow illustrating NGS process for DNA sequencing

14.2 Fragmentation

Fragmentation can be done either by physical or enzymatic method (Thermes 2014). Physical methods include acoustic shearing, sonication (excitation using ultrasound) to create smaller strands. This depends on short-read sequencing technologies such as Illumina which cannot readily analyze long DNA strands a sin case of 10X or Pacbio. Hence the samples are fragmented into uniform pieces to make them enable to sequence.

14.3 Adapter Ligation

The adapters are introduced at the beginning and end to create known “fragments” in the form of random sequences (Heather and Chain 2016). A ligase enzyme covalently links the adapter and inserts DNA fragments, making complete library molecules. As these adapters serve as multiple functions, they are attached to the sequences for easy sample identification and multiplexing. In some cases, barcodes are also attached while for bridge amplification, the DNA fragment binds to the oligos which creates the bridge with a primer binding to this DNA sequence and amplifying vertically (Ambardar et al. 2016).

14.4 Sequencing

The polymerase adds the nucleotide into the bridge amplification, where the signals are recorded which will further generate multiple sequencing databases for the DNA sequences (Buermans and den Dunnen 2014). The sequencing steps vary from one instrument to the other depending on the capacity and flow cells. There are other parameters, viz. average depth, coverage and size of the reads and read chemistry as discussed in Table 14.1 and Fig. 14.1.

14.5 Data Analysis

The data generated by the sequencing machines can be aligned to the reference genome sequence. Basically, from the library preparation to data analysis, there are four different types of NGS methods, viz. whole genome sequencing (WGS), whole exome sequencing (WES), whole transcriptome shotgun sequencing (WTSS), and targeted/amplicon genome sequencing. If the organism is not under question, there could be metagenome sequencing where several organisms in questions could be characterized,

As dealt earlier, various modern sequencing technologies including – Illumina (Solexa) sequencing, whole genome sequencing, Targeted sequencing, Amplicon sequencing, exome sequencing, de novo sequencing and transcriptomics, etc. have been preferentially used for short-read sequencing while 10x or Pacbio (Rhoads and Au 2015) or Oxford Nanopore uses long read sequencing chemistry (Metzker 2010). The sequences are read from one end to the other based on the type of read chemistry, single in case of single read, paired end, and mate pair chemistry in the case of multiple paired end reads (Head et al. 2014). There are, however, recent technologies that have allowed us to sequence DNA and RNA more easily and cost-effectively than previously used Sanger sequencing (Pareek et al. 2011). This has not only helped in studying any alterations harbored in genetics and molecular pathways associated with mutated genes but also allowed us to identify non-coding spectrum from WGS and WTSS datasets (Ansorge 2009). In addition, NGS in recent years has made it possible to better understand the genetics behind rare diseases and implement it as a technological advancement in clinical and diagnostic practices across a wide array of genomes (Mathur et al. 2018). The NGS has allowed us to analyze diverse regions of a genome in a single reaction assay in a much better cost-effective manner and proved as an efficient tool for examining patients with genetic disorders (Depristo et al. 2011), (Deurenberg et al. 2017) (Fig. 14.2). Furthermore, the molecular and genomic data in the form of precise detection of disease biomarkers has helped understanding regulation, identifying inheritable disorders, and depicting factors governing response to therapeutic responses (Rabbani et al. 2014; Jamuar and Tan 2015). Moreover, a diverse variety of molecular tests are implemented that make use of NGS technology, such as single- and multiple-gene panel sequencing, single-cell sequencing, WGS, WES, cell-free DNA for prenatal sequencing (Van den Veyver and Eng 2015). Considering the fact that NGS is a recent and efficient

Table 14.1 Different platforms of next generation sequencing

NGS platform	Output (gb)	Read length	Run time	Advantages	Disadvantages	Applications
Roche 454	0.04	400–700 bp	23 h	High-throughput Longer sequence length	High error rate owing to homo-polymers	– Emulsion of the PCR – Pyrosequencing technology
Hi Seq 2500/ 4000	2.9	36–100 bp	11 days	Patterned flow cell; low cost per read per MB	Expensive; may not be flexible in usage when compared to its peers	High output and more rapidity
Mi-Seq	0.2	36–250 bp	4–24 h	– Long reads – Highest flexibility – Low run cost – Broad range of biological applications	Relatively few reads and higher cost per Mb compared to next seq or hi Seq	– Sequencing small genome – Amplicon sequencing – Small RNA – Metagenomic 16 s ribosomal sequencing
PacBio RS 2/bio Seq	0.1	250–10,000 bp	30 min to 6 h	– single molecules real time sequencing – Longest available read length – Ability to detect base modification	– High error rates so high coverage is necessary – Low total number of reads per run – High cost per Mb	Allow to capture long DNA molecules and sequence
Ion PGM/proton	2.5	100–200 bp	2–4 h	– Fast – Scalable – High-throughput	– Higher error rate than illumine – Higher cost per Mb relative to mi-seq	– Whole genome sequencing – Exome sequencing – RNA sequencing – mRNA sequencing
Nova Seq	134–6000	2 × 150 bp	13–44 h	– Scalable throughputs – 2-channel chemistry – Reduced imaging and data processing time	– Careful consideration of flow cells output is crucial – High yields of flow cells present difficulties in efficiently filling flow cells	– Whole genome sequencing – Exome sequencing – RNA sequencing – epigenome sequencing

(continued)

Table 14.1 (continued)

NGS platform	Output (gb)	Read length	Run time	Advantages	Disadvantages	Applications
Next Seq 500	120	150 bp	1 day	– Easy to use – Moderate instrument and run cost.	Version 2 of chemistry not as good as older chemistry (used on mi-Seq and hi-Seq)	– RNA sequencing – Whole genome – Metagenome – Whole exome sequencing
Oxford Nanopore technology	1–2 GB per flow cell Flongle	50–10,000 bp	30 min to 3 h	– Longest available read length	– Minimized error rates with less total number of reads per run – High cost per Mb	Allow to capture long DNA molecules and sequence Portable

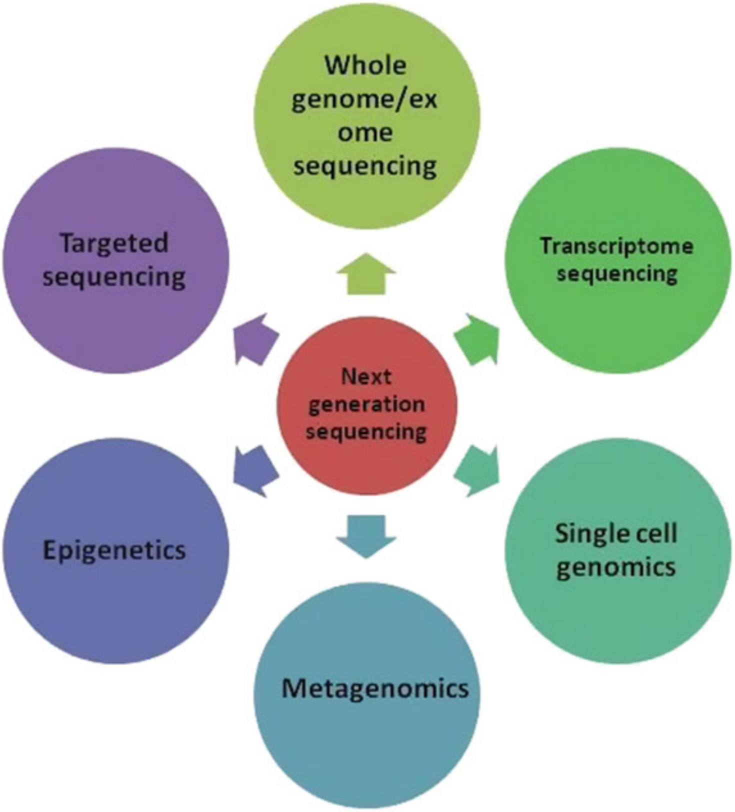


Fig. 14.2 Different applications of Next Generation sequencing

diagnostic tool in clinical genetics, there are challenges and limitations regarding how to analyze and interpret the sequencing data and communicate it to patients and family members. Thus, it is essential to properly understand the applications, strength and weakness of different approaches which we document below for each case.

14.5.1 Whole Genome Sequencing

A comprehensive method for analyzing the entire genome information which identifies inherited disorders, the mutations lead to various disease outbreaks and population genetics (Ng and Kirkness 2010). The sequencer has both single and paired end reads of data that will map with above Q30 Phred quality score. The sequences will be retrained for further analysis after removing adapters and low-quality reads with >30. The total generated reads are aligned to the reference genome with an average read depth and genome-wide coverage if the organism has the reference data otherwise, the de novo assembly step will be proceeded for data analysis (Lam et al. 2012). Thus, the predicted gene sequences will be annotated on freely available online resources such as Uniprot, NR-NCBI databases that will identify the known unknown regions. From the WGS results, we can analyze the read depth, gene density, insertion density, and SNP density and elucidate the unexplored genomic regions. The identified unique variants would reveal novel biological pathways that lead to complex disorders that provide high-resolution insights in the affected pathways (Sanders et al. 2017).

14.5.2 Whole Exome Sequencing

The whole exome sequencing (WES) is one of the broadly used NGS techniques where only protein-coding regions of the genome are sequenced. As the human exome consists of less than 2% of protein-coding genes but harbors more than 85% of the disease-causing variants, resulting in a cost-efficient sequencing approach in contrast to whole genome sequencing (WGS). The DNA libraries for WES approach could be developed in just 1 day, thereby yielding 4–5 Gb of sequenced data per exome. The WES utilizes exome enrichment methodology for deciphering coding regions which can further be applied to a wide range of clinical applications, including cancer studies, population-based studies, and genetic disorders (Gupta et al. 2017; Mueller et al. 2018; Weigelt et al. 2018). In addition, WES has been proven advantageous for identifying pathogenic variants in several Mendelian phenotypes, complex disorders as well as rare disorders (Jeste and Geschwind 2014; Mathur et al. 2018). Since the past, the WES approach has been routinely applied in clinical diagnostics as a generic test in managing various disorders (Arts et al. 2019) and has been included as an efficient genomic strategy in the 1000 Genome Project (Altshuler et al. 2012), and Exome Aggregation Consortium (ExAC) (Lek et al. 2016) to decipher population-risk variants and to predict disorders linked to rare mutations. Various pipelines to perform the analyses do exist and vary from commercial platforms to open source tools, viz. SeqMule, Interpretomics, Qiagen/CLCBio, GATK in addition to bash based pipelines that uses open source tools developed by us (Meena et al. 2018). In contrast to targeted sequencing, WES has several advantages, for example, firstly, it allows prediction of novel causal genes associated with any genetic disorder that are not included in exome-wide genetic panel and secondly, along with small polymorphic variants it also provides genome-wide data accessibility for reliable detection of larger polymorphic sites including copy number variants (CNVs) and homozygous locations (Stray-Pedersen et al. 2017; Gambin et al. 2017). Moreover, to reduce the intricacies of data analysis and accelerate the process, WES methods could be combined with computational data-driven processes of already reported cohorts of causal genes (Neveling et al. 2013). As WES delivers an extensive depth of coverage for the coding regions of the genome and yields, compact and manageable data information for faster and precise analysis are used in comparison to WGS methods (Gupta et al. 2020). The WES methods allow variant detection located in the coding exonic sites,

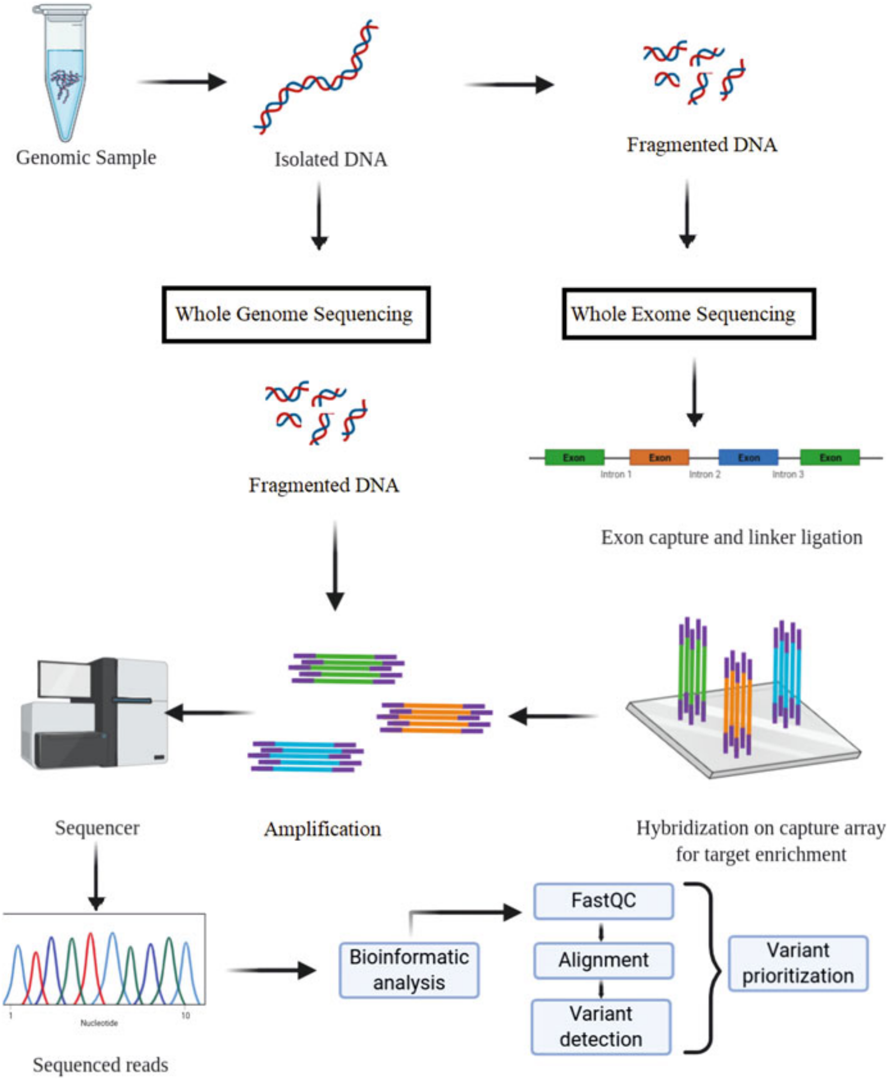


Fig. 14.3 A schematic representation of Whole Genome and Whole Exome Sequencing workflow and analysis

with an ability to extend the target regions to involve untranslated regions (UTRs), and in some cases microRNAs (de Carvalho et al. 2019) and even long non-coding RNAs to get a more detailed outlook of gene regulation in rare disorders (Gupta et al. 2018) (Fig. 14.3).

14.5.3 Whole Transcriptome Shotgun Sequencing

In a given cell at a particular stage, the entire set of RNA transcripts is known as transcriptome. To understand any particular development and disease, understanding the transcriptome is an essential element (Martin and Wang 2011). One of the most widely used methods to study differentially expressed genes is microarray technology but it has its own limitations. Advancement in sequencing technologies has revolutionized transcriptome analysis by c-DNA sequencing (RNA-Seq). Because of its higher reproducibility and better resolution, RNA-Seq is widely accepted and used for different research purposes (Wang et al. 2009). The main steps of RNA-Seq approach include (1) analysis of raw data, (2) alignment read, (3) transcriptome reconstruction, and (4) quantification and differential expression analysis (Nagalakshmi et al. 2010). Initial steps of RNA-Seq include quality check of the raw data followed by mapping to the reference genome. If there is no availability of reference genome, it can be done by using a de novo assembly approach. The last step is analysis of differentially expressed genes which can be done using different available approaches (Garber et al. 2011) (Fig. 14.4).

14.5.4 Metagenome Sequencing

Metagenomics is the sequencing and analysis of complex genomic material derived directly from clinical or environmental samples in order to investigate the population

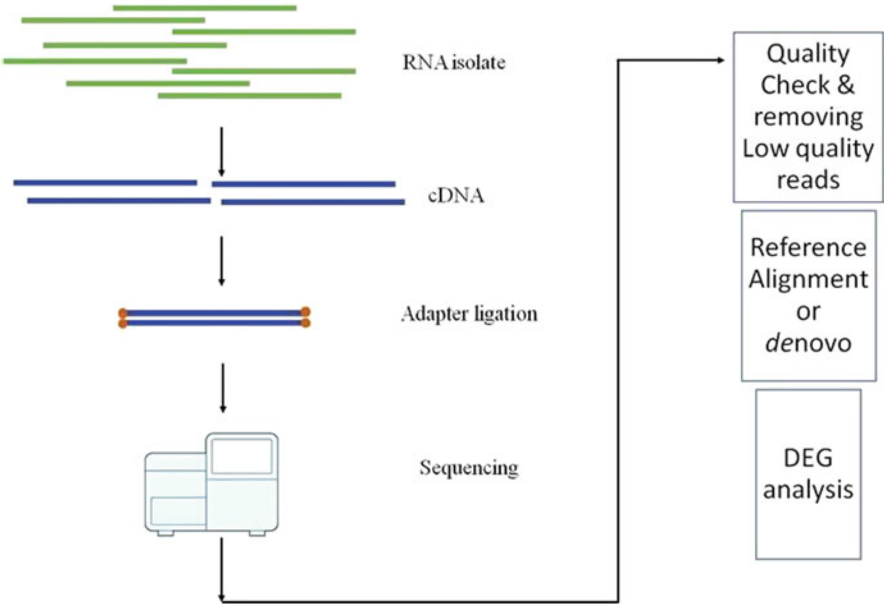

RNA isolate
cDNA
Adapter ligation
Sequencing
Quality Check & removing Low quality reads
Reference Alignment or *denovo*
DEG analysis

Fig. 14.4 A schematic representation of RNA-sequencing workflow

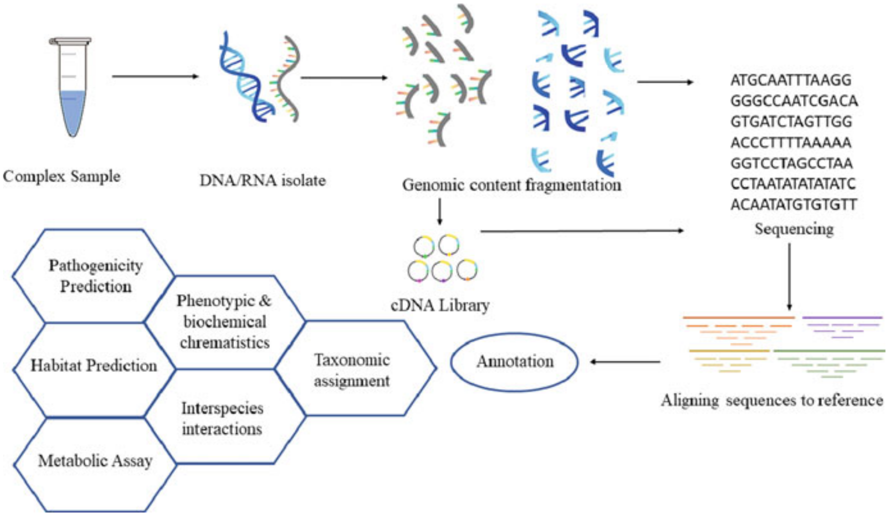


Fig. 14.5 Workflow of metatranscriptomics analysis

of microorganisms present, without the need of collecting pure cultures (Tringe 2005). Metagenomics is also known as microbial ecogenomics or community genomics or environmental genomics. Functional metagenomics facilitates high-resolution genomic analysis of unculturable microbes and the connection of genomes with different environmental functions (Lam et al. 2015). The process of metagenomics involves isolation of DNA followed by random fragmentation and sequencing which leads to the identification of different genes and metabolic processes performed by microbial communities which in turn depend on sequencing depth and complexity of the community (Charles et al. 2017). A simple workflow has been described in Fig. 14.5 with metagenomics used to analyze the functional diversity of different microorganisms present in metal-contaminated ground water, at contaminated sites and in environmental samples. Through the advent of NGS, millions of genes have been identified in the gut microbiome of European population. By mapping metagenome sequences against reference genome, different species from a sample can be identified (Morgan et al. 2010). Recently, metagenomic next generation sequencing (mNGS) has emerged as sensitive technology that is capable to detect pathologenic organisms from human sample such as blood, urine, BALF, and sputum (Wylie et al. 2013; Salzberg et al. 2016; Mai et al. 2017; Parize et al. 2017). The mNGS provides detailed sequencing of the total DNA or RNA content of the microbiome and hence has the potential to identify pathogens further facilitating easy diagnosis (Barzon et al. 2011; Chan et al. 2014; Goldberg et al. 2015). Advances in NGS technology now allow us to understand the microbial community in various niches, including respiratory tract in both physiological and pathological conditions (Hui et al. 2013).

Metatranscriptomics and metaproteomics are fairly new sub-areas of metagenomics that aid functional study of microbial communities.

Metatranscriptomics is used to study gene expression by performing whole metatranscriptome shotgun sequencing by capturing total mRNA of the microbiome (Bashiardes et al. 2016). This also helps to better understand the functional component of the entire microbial community on microbial characterization and its novel interactions, study of differential gene expression, etc. Because of high environmental diversity and enormous microbial population, metatranscriptomics cannot capture the entire metatranscriptome. Since RNA is short-lived, it interferes with the experimental designing of metatranscriptomics analysis. While these are some of the shortcomings of metatranscriptomics, metaproteomics fills the gap in studying all proteins obtained from environmental sources. Metaproteomics is also known as community proteomics, community proteogenomics or environmental proteomics (Maron et al. 2007). The different steps of metaproteomics include sample collection, protein extraction, and separation by 2D gel electrophoresis followed by mass spectrometry based identification (Heyer et al. 2017). Metagenomic investigation explores the entire genomic makeup of the microbial communities through sequencing followed by downstream analysis. It also plays a role in the assessment of the biochemical component of microbes in atypical environments and their association with other environmental features (Thomas et al. 2012). With the help of metagenomics, some of the complex ecological interactions such as phage-host dynamics, lateral gene transfer, and metabolic complementation can be studied now which can shed new insight into the environmentally and functionally important microbial communities.

14.5.5 Machine Learning in NGS

In 1947, Alan Tuning gave a statement "... what we want is a machine that can learn from experience." Today is a time where we are using a technique which learns from experience and is named as *Machine Learning* (ML). Machine Learning aims to forecast the future possibilities in a specific problem by continuous learning. In the past decade, machine learning has vastly improved understanding of the human genome. The ML is so pervasive today that we probably use it dozens of times a day without knowing it. ML has revolutionized traditional statistical techniques by supplying a new insight in data analysis. It has been deployed on vast problems of different areas like biomedical, education, commerce, engineering, aeronautics, space science, and many more. While the ML has enormous capabilities of drilling out deeply the wide variety of hidden patterns, this incredible property has made it an integral part of genomics research. Generally, genomics has problems: genome sequencing, gene-editing, drug–target interaction, molecular docking, microorganism genesis, antiviral drug study, etc. The nature of data in such problems is massive, multidimensional, viscosity, and variety. Also, data could be structure like database, or semi structures as log files or completely unstructured like videos, images, and audios. The ML algorithms have been devised to penetrate in the structured or unstructured perplexing data and get insight from it. While ML does not only provide predictive analysis, it also assists in decision making. These algorithms have capabilities to prescribe a highly possible or successful decision in all sets of problems. Precision medicine and treatment recommendation systems are well known applications of ML's prescriptive analysis. ML is an umbrella term used for a different set of algorithms as these are classified based on learning mechanisms. Learning is a characteristic of an algorithm that helps it to acquire knowledge from data. If an algorithm is not trained about the data before it concludes to any decision, then such algorithm is known as an unsupervised learning algorithm. Whereas when an algorithm that generates results only after proper training using labelled data, it is termed as supervised learning. Algorithms with intrinsic nature of self-training and self-learning are called as reinforcement learning algorithms. Some machine learning algorithms are discussed in Table 14.2.

The results supplied by ML algorithms are validated using different types of metrics. These metrics are selected after identifying the algorithm's task from given five categories: association, clustering, dimension reduction, classification, and prediction.

- Association rule algorithms find the relationships between the given data items. The support and confidence are commonly defined measures for association rule algorithms.
- Clustering algorithms split the datasets in different clusters using a mathematical formula (Brun et al. 2007). These are validated using three types of measures, i.e. internal, external, and relative. Internal validation assesses the clustering quality using intrinsic data (Hämäläinen et al. 2017).
- External validation works with extrinsic data to evaluate level of cohesion in grouping. Relative validation is a hybrid of above two and judges the overall quality. Internal indexes are Dunn index, Calinski–Harabasz, Gamma index, C-Index, Davies–Bouldin index, Silhouette index, etc. External are Jaccard Coefficient, Goodman–Kruskals Coefficient, f measure, Rand statistics, purity, Fowlkes and Mallows Index, Entropy, etc. (Vendramin et al. 2010).
- Dimensionality reduction (DR) algorithms are essential for reducing the dimensional complexity of data (Sarwar et al. 2000). The quality assessment for such algorithms is classified into two categories: local preservation criteria and global preservation criteria. Local approach focuses on local-neighborhood preservation, whereas global approach focuses on overall structure-holding preservation (Gracia et al. 2014).
- Classification algorithms classify a dataset based on the class of output variable/s. These are assessed using a confusion matrix or error matrix. A confusion matrix is a report card of classification model providing details of its performance with sensitivity and specificity. It depicts a number of classified data in the form of true positive, true negative, false positive, and false negative which reveals the correctly and incorrectly classified data. Apart from this accuracy and precision are also calculated using confusion matrix. These two metrics define the robustness of the model. A *receiver operating characteristic* (ROC) curve is also primarily used to accept or reject the classification model (Cai and Dodd 2008; Greiner et al. 2000). It is drawn using specificity and sensitivity provided by the

Table 14.2 Machine learning algorithm for NGS

S. no.	Category	Name of algorithm	Concept
1.	Unsupervised	K-means	• Clustering algorithms • Divide data into k clusters • Euclidean distance is used to form clusters • It takes numeric data as input
		Apriori	• Association rule-based algorithms • Generate frequent itemset after finding relationship between the given items in a set of transactions • Relationships are based on occurrence of items in a transaction set • It takes numeric input
		FP-growth	• Association rule-based algorithms • Generates frequent pattern trees after finding relationships between the given items • Frequency of item in each transaction is considered • Takes numeric input
2.	Supervised	Linear regression	• A linear approach to model the relationship between dependent and independent variables • Two types: Simple and multiple linear regression • Takes numeric input • Used for real valued prediction
		Logistic regression	• Models the data using the sigmoid function • Three types: binomial, multinomial and ordinal • Takes numeric input • Used for classification
		Decision tree	• Decision trees are formed • Identifies ways to split a data set based on different conditions • It takes numeric input • Used for classification and prediction
		SVM	• Creates a hyperplane that separates the dataset into classes • Hyperplane is defined in N dimensional space or N number of features used for classification • Kernel function is used to design hyperplane • Different type of kernel functions • Takes numeric input • Used for classification and prediction
		Naive Bayes	• Assumes the independence between features which is commonly not found with real life datasets • Requires less training data • Used for classification

(continued)

Table 14.2 (continued)

S. no.	Category	Name of algorithm	Concept
		Random Forest	• Uses an ensemble of decision trees • Takes feature importance into consideration • Used for both classification and regression
		Dimensionality reduction algorithms	• Aim to reduce the no of features while preserving the prediction capability of the model • Improves the performance of the ML model while handling big data sets with high dimensionality
		Gradient boosting algorithms	• Uses an ensemble of decision trees as weak learners • Uses gradient descent to minimize the loss while adding trees to the model • Used for both regression and classification
		Multilayer perceptron	• Used to learn the non-linear relationships between the inputs and outputs • Weights are updated using a backpropagation algorithm
3.	Reinforcement learning		• Q learning and SARSA are two algorithms wherein the former uses a greedy approach that takes action based on maximum utility of the next state while the latter uses a stated policy to take the next action

confusion matrix. Other metrics for decision tree algorithms are Gini Index, and classification error and entropy (Silahtaroğlu 2009).

- Prediction algorithms are used to forecast the possibility of object or event. The measures for evaluating prediction are accuracy, precision; mean absolute error, root mean square error, etc. These are also used for measuring the performance of prescriptive analysis.

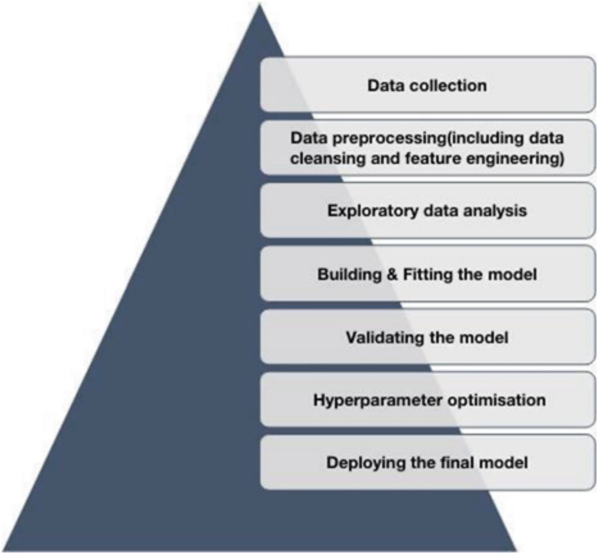
These metrics are behaviors that measure the progress of algorithms. They guide researchers to finalize and verify the ML based model for their study as different NGS applications require different models (Marceddu et al. 2019). Common application of machine learning algorithms in NGS are screening of compounds, fragment-based de novo design, computational screening of molecular fragments, and fragment linking to design novel inhibitors, molecular docking analysis with virtual screening, Construction of homology models, designing of linear discriminant analysis model, and designing of analogs (van den Akker et al. 2018). The new areas of genomic research are soil fertility study with microorganisms, newborn genetic screening, precision medicine, gene-based prescription, energy healing and genetic transformation, genetic disease and vibrational therapies, and many more.

14.5.6 ML and NGS Data Analysis

NGS data generated using different platforms has inherent applications in identification of genes, variants including copy number variants (CNVs) and single nucleotide polymorphisms (SNVs), exomes, RNA, and small RNAs (Tripathi et al. 2016). The enormous data generated using NGS is one of the most promising instances of "big data" which is evident from the fact that the space needed to store 1000 genomes is approximately three terabytes. The storage space and data processing requirements are often addressed using cloud computing solutions with Apache Hadoop framework. The sequence compression algorithms also contribute towards better transmission, analysis, and storage of such data (Wandelt et al. 2012). The cost effectiveness of NGS in terms of time and throughput as compared to Sanger sequencing has posed challenges as well as opportunities in terms of effective data analysis and storage. With the advent of NGS techniques the focus gradually shifted from data generation to methods which could assist in gaining insights out of this data. The basic workflow of machine learning can be summarized as shown in Fig. 14.6.

The first step in analyzing the NGS data is the curation of the relevant dataset with respect to the underlying problem (Tripathi et al. 2016). The next step is to preprocess the data which itself is a combination of several substeps including data cleaning, data integration, data transformation, handling data imbalance issues, and dimensional reduction. Data cleaning includes removal of out of date data, handling or removal of missing values/features, respectively, and identifying the outliers. The data integration step may involve the handling of different heterogeneous sources of data while converting them into a uniform format. Some features in the dataset have extreme values falling in different ranges which can be handled

Fig. 14.6 Machine learning workflow


Data collection
Data preprocessing(including data cleansing and feature engineering)
Exploratory data analysis
Building & Fitting the model
Validating the model
Hyperparameter optimisation
Deploying the final model

using different normalization schemes for transforming the data. Data imbalance is one of the most prominent issues while handling NGS datasets. This issue is very inherent while identifying the trait/disease related non-coding variants (Schubach et al. 2017) wherein the machine learning model tends to learn the majority class thereby generating wrong predictions/classification for the instances of the minority class. The purpose of dimensionality reduction is to reduce the number of features/ variables which thereby affects the storage and time complexity involved in processing such data. However, feature selection differs from dimensionality reduction in the way that the former deals with selecting the relevant and important features while the latter aims at projecting the existing features to a lower dimensional space. This further simplifies the data exploratory analysis step wherein the visualization can be done in the form of plots to study or analyze the data prior to modeling or hypothesis testing (He et al. 2017). As the next step the ML model is chosen and trained on the training dataset and further evaluated using the test set. The evaluation results can be further improved by tuning the hyper parameters associated with the ML model. The hyperparameters are the structural/architectural parameters say the no of estimators or depth to be taken for a random forest model. The final learned/tuned model should have the generalization ability to predict for any kind of input sample.

14.6 Tools for NGS Data Analysis

The low cost and high-throughput NGS techniques have encouraged the application of genetic tests for studying and identifying the variants, mutations associated with the rare diseases, Mendelian disorders (Wadapurkar and Vyas 2018). This has also paved the way to diagnose and cure the human genetic disorders and diseases on the basis of individual genome profile. There are various computational tools that are used for NGS data analysis from preprocessing, sequence alignment, post alignment processing, variant (structural variants and copy number variants) calling, variant functional annotation stages. A list of tools used for NGS data analysis is shown in Table 14.3.

14.6.1 Current Challenges of NGS

Over the last few decades, the NGS technique and its applications has increased by leaps and bounds (Levy and Myers 2016). The outcomes have shown rise while lowering down the sequencing costs per sample run, both by orders of magnitude and precision. The majority of sequencing platform companies have spent a couple of years since the past, mainly focusing on improving it in terms of accessibility and user-friendliness. Illumina's newly launched sequencing systems such as HiSeq (Illumina 2015), MiSeq (Schirmer et al. 2015), NextSeq systems, all making use of reagent cartridges for operations and reducing the hands-on-time for sample library preparation. The Ion Torrent sequencing platforms have been observed to

Table 14.3 List of tools used for NGS data analyses, all URLs accessed on September 27, 2020

Tools	Function	Category	References/URL
FastQC	QC report for high-throughput sequencing data	Quality check	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FASTX toolkit	Performs some of the preprocessing tasks such as conversion from FASTQ to FASTA format, renaming the sequence identifiers in FASTA/FASTQ files, collapsing identical sequences in FASTA/FASTQ files into a single sequence, etc.	Preprocessing	http://hannonlab.cshl.edu/fastx_toolkit/index.html
Cutadapt/sickle/PRINSEQ/Trimmomatic	Removal of adapters and low-quality reads	Preprocessing	Martin (2011), Joshi and Fass (2011), Bolger et al. (2014); http://prinseq.sourceforge.net/index.html
BWA	BWA is a software package that is based on burrows wheeler transform algorithm and is fast and efficient for short and long reads.	Sequence alignment	Li and Durbin (2009)
mrFAST	Micro read fast alignment search tool which maps short reads from the ILLUMINA platform.	Sequence alignment	http://mrfast.sourceforge.net/
Bowtie	Faster than BWA, memory efficient short-read aligner	Sequence alignment	Langmead et al. (2009)
HISAT2	A graph-based alignment algorithm. It enables fast and sensitive alignment	Sequence alignment	http://daehwankimlab.github.io/hisat2/
ABySS	De novo assembler	Sequence alignment	Simpson et al. (2009)
SAM tools	Allows for manipulation of alignments in the SAM (sequence alignment/map) format	Post alignment processing	http://samtools.sourceforge.net/
VarScan	Variant detection	Post alignment processing	http://varscan.sourceforge.net/
Genome analysis toolkit (GATK)	Suite of tools including depth of coverage analyzers, unified genotype inference, haplotype mapping a	Post alignment processing	https://gatk.broadinstitute.org/hc/en-us

(continued)

Table 14.3 (continued)

Tools	Function	Category	References/URL
	quality score recalibrator, SNP/indel caller, etc.		
Breakdancer	PERL package used for detection of structural variants and is based on paired end mapping strategy.	Variant calling	Fan et al. (2014)
PEMer	Package to analyze and construct structural variants.	Variant calling	Korbel et al. (2009)
SeattleSeq	Provides annotation of SNVs and small indels.	Variant functional annotation	https://snp.gs.washington.edu/SeattleSeqAnnotation153/
ANNOVAR	Software tool to functionally annotate the variants from several genomes of different organisms	Variant functional annotation	Wang et al. (2010)
snpEff	Annotates and specifies the effects of variants on genes	Variant functional annotation	Cingolani et al. (2012)
TopHat	It is a fast splice junction mapper for RNA reads.	Transcriptome analysis	Trapnell et al. (2009)
Orange	Powerful tool for interactive workflows and visualizations	Transcriptome analysis	Demšar et al. (2013)
Cuffmerge	Merge transcripts	Transcriptome analysis	http://cole-trapnell-lab.github.io/cufflinks/cuffmerge/
Cufflinks	Assembles and estimates the abundance of transcripts	Transcriptome assembly and differential expression analysis for RNA-Seq	http://cole-trapnell-lab.github.io/cufflinks/
Cuffdiff	Estimate differential expression at gene and transcript level	Reports at FPKM level	http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/
CummeRbund/DESeq2/edgeR	Visualization tools	Differentially expressed genes	Goff et al. (2012), Love et al. (2017), Robinson et al. (2009)
Blast2GO	Annotation, visualization and analysis in functional genomics research	Gene ontology	Conesa et al. (2005)
KAAS server	KEGG automatic annotation server provides functional annotation of genes by BLAST	Pathway analysis	Moriya et al. (2007)

(continued)

Table 14.3 (continued)

Tools	Function	Category	References/URL
PRODIGAL	Prokaryotic dynamic programming gene finding algorithm	Metagenome analysis	Hyatt et al. (2010)
Mothur	Analyzing and comparing microbial communities	Metagenome analysis	Schloss et al. (2009)
QIIME	Quantitative insights into microbial ecology	Metagenome analysis	Caporaso et al. (2010)

be more difficult to operate than Illumina platforms. Nevertheless, the Ion-S5 system from Thermo Fisher Scientific has been explicitly engineered to shorten the entire sequencing methodology, from sample library preparation to data production and interpretation (Quail et al. 2012). After looking at such sequencing improvements of higher outcomes, reduced sequencing costs and greater accessibility to users, one can imagine that all the barriers to progression have been secluded. But the hard-work has just begun and there remain various challenges for NGS that need to be resolved. One of the major challenges is about data storage, as there is a huge amount of data generated by NGS and storing them can be a herculean task. For every single sample, the generated raw files are in gigabytes depending on its further application, which makes this process cumbersome. For example, the raw reads for whole genome sequencing can go up to 250GB whereas for deep RNA-seq, the raw reads in fastq formats range between 20 and 25 GB. After obtaining and filtering initial data, it can be streamlined for further downstream analysis which requires alignment to the reference genome or transcriptome of the organism of origin (Langmead et al. 2009)

This is the most time-consuming step of the analysis which requires different algorithms but this poses another problem. Choosing an appropriate algorithm tool from an existing set is not that easy. Some of the important criteria to consider while choosing an algorithm should be its performance (in publications) or in-house benchmarking studies. Another important point to consider for NGS analysis is smooth and fast functioning of the workflows and the instruments involved in analysis. Another challenge is about data analysis as there is multiple software available so choosing the right one according to one's need is not that easy. RNA-seq has revolutionized our understanding of the entire transcriptome to better analyze differential gene expression in different experimental groups (Trapnell et al. 2010). Each RNA-seq experiment basically consists of several steps such as experimental design, mapping short reads, quality control, estimating transcript abundance, and analyzing differential expression. All these steps have their own challenge (Waern et al. 2011). For example, proper normalization of read counts is required for estimating transcript abundance but due to RNA fragmentation, longer transcripts generate more reads compared to shorter transcripts and both are present at the same abundance in the sample which causes variation in the data analysis. Another important point to consider during RNA-seq is non-uniformity of coverage which is produced as a result of biases introduced at different steps such as cDNA synthesis, amplification, and sequencing (Finotello et al. 2012). There could be transcript length bias as well which is due to several fragmentation steps during the entire process. This further poses a problem in downstream analysis. One of the major areas that have been disregarded most of the time is the quality of samples that need to be sequenced. Since the first step in NGS is library preparation, knowing the quality of nucleic acid (DNA/RNA) before the process is crucial (Luthra et al. 2015). Another major challenge in NGS is to successfully process low input samples with accuracy and precision (Head et al. 2014). Since, these samples are precious, proper handling has to be done to minimize losses and retain data quality.

14.7 Conclusions and Future Perspectives

Application of NGS technique has allowed the researchers to predict large amounts of genomic, transcriptomic, and metagenomic data in an efficient manner. However, only a small fraction of data has been applied for clinical and diagnostic purposes, so far. Understanding the rest of unsupported data could be managed with the help of introduction of novel and powerful NGS approaches and algorithms. The potential that NGS technique holds for diagnosing any disorder and their clinical implementation is enormous which could be further enhanced by careful application of NGS data analysis. Moreover, expeditious usage of divergent methods for big data analysis might improve patient diagnosis and treatment. Implementation of computational diagnostic methods such as machine and deep learning algorithms are more consistent, unbiased and less prone to human errors. Utilizing machine learning algorithms as a diagnostic modeling tool are more advantageous in clinical practices. However, such models are required to be properly trained for understanding biological systems to have proper insights into the disease mechanisms. In such a manner, these modeled algorithms could be used in agreement with the analyzed disorder, and concurrently could be trained and altered according to the disease evolution.

References

Abbasi S, Masoumi S (2020) Next-generation sequencing (NGS). Int J Adv Sci Technol. https://doi.org/10.1007/978-3-662-49054-9_3542-1

Abdellah Z, Ahmadi A, Ahmed S et al (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945. https://doi.org/10.1038/nature03001

Altshuler DM, Durbin RM, Abecasis GR et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. https://doi.org/10.1038/nature11632

Ambardar S, Gupta R, Trakroo D et al (2016) High throughput sequencing: an overview of sequencing chemistry. Indian J Microbiol 56:394–404

Ansorge WJ (2009) Next-generation DNA sequencing techniques. N Biotechnol 25:195–203

Arts P, Simons A, AlZahrani MS et al (2019) Exome sequencing in routine diagnostics: a generic test for 254 patients with primary immunodeficiencies. Genome Med 11:38. https://doi.org/10.1186/s13073-019-0649-3

Barzon L, Lavezzo E, Militello V et al (2011) Applications of next-generation sequencing technologies to diagnostic virology. Int J Mol Sci 12:7861–7884. https://doi.org/10.3390/ijms12117861

Bashiardes S, Zilberman-Schapira G, Elinav E (2016) Use of metatranscriptomics in microbiome research. Bioinform Biol Insights 10:19. https://doi.org/10.4137/BBI.S34610

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brun M, Sima C, Hua J et al (2007) Model-based evaluation of clustering validation measures. Pattern Recogn 40:3. https://doi.org/10.1016/j.patcog.2006.06.026

Buermans HPJ, den Dunnen JT (2014) Next generation sequencing technology: advances and applications. Biochim Biophys Acta – Mol Basis Dis 1842:1932–1941

Cai T, Dodd LE (2008) Regression analysis for the partial area under the ROC curve. Stat Sin 18:817

Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335

Chan BK, Wilson T, Fischer KF, Kriesel JD (2014) Deep sequencing to identify the causes of viral encephalitis. PLoS One 9:e93993. https://doi.org/10.1371/journal.pone.0093993

Charles TC, Liles MR, Sessitsch A (2017) Functional metagenomics: tools and applications. Springer, Cham

Chiu RWK, Chan KCA, Gao Y et al (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci 105:20458–20463. https://doi.org/10.1073/pnas.0810641105

Cingolani P, Platts A, Wang LL et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80–92. https://doi.org/10.4161/fly.19695

Conesa A, Götz S, García-Gómez JM et al (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674. https://doi.org/10.1093/bioinformatics/bti610

de Carvalho JB, de Morais GL, Vieira TCDS et al (2019) miRNA genetic variants alter their secondary structure and expression in patients with RASopathies syndromes. Front Genet 10:1144. https://doi.org/10.3389/fgene.2019.01144

Demšar J, Curk T, Erjavec A et al (2013) Orange: data mining toolbox in python. J Mach Learn Res 14:2349–2353

Depristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. https://doi.org/10.1038/ng.806

Deurenberg RH, Bathoorn E, Chlebowicz MA et al (2017) Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol 243:16–24. https://doi.org/10.1016/j.jbiotec.2016.12.022

Di Resta C, Galbiati S, Carrera P, Ferrari M (2018) Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities. Electron J Int Fed Clin Chem Lab Med 29:4–14

Fan X, Abbott TE, Larson D, Chen K (2014) BreakDancer: identification of genomic structural variation from paired-end read mapping. Curr Protoc Bioinformatics 45:15. https://doi.org/10.1002/0471250953.bi1506s45

Finotello F, Lavezzo E, Barzon L et al (2012) A strategy to reduce technical variability and bias in RNA sequencing data. EMBnet J 18:5. https://doi.org/10.14806/ej.18.b.552

Gambin T, Akdemir ZC, Yuan B et al (2017) Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. Nucleic Acids Res 45:1633–1648. https://doi.org/10.1093/nar/gkw1237

Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8:469–477

Goff LA, Trapnell C, Kelley D (2012) CummeRbund: visualization and exploration of cufflinks high-throughput sequencing data. R Packag version

Goldberg B, Sichtig H, Geyer C et al (2015) Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. MBio 6:e01888. https://doi.org/10.1128/mBio.01888-15

Gracia A, González S, Robles V, Menasalvas E (2014) A methodology to compare dimensionality reduction algorithms in terms of loss of quality. Inf Sci (Ny) 270:1–27. https://doi.org/10.1016/j.ins.2014.02.068

Greiner M, Pfeiffer D, Smith RD (2000) Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev Vet Med 45:23–41. https://doi.org/10.1016/S0167-5877(00)00115-X

Gupta S, Chatterjee S, Mukherjee A, Mutsuddi M (2017) Whole exome sequencing: uncovering causal genetic variants for ocular diseases. Exp Eye Res 164:139–150

Gupta S, Gupta N, Tiwari P et al (2018) Lnc-EPB41-protein interactions associated with congenital pouch colon. Biomol Ther 8:95. https://doi.org/10.3390/biom8030095

Gupta A, Shukla N, Nehra M et al (2020) A pilot study on the whole exome sequencing of prostate cancer in the Indian phenotype reveals distinct polymorphisms. Front Genet 11:874. https://doi.org/10.3389/fgene.2020.00874

Hämäläinen J, Jauhiainen S, Kärkkäinen T (2017) Comparison of internal clustering validation indices for prototype-based clustering. Algorithms 10:105. https://doi.org/10.3390/a10030105

He KY, Ge D, He MM (2017) Big data analytics for genomic medicine. Int J Mol Sci 18:412

Head SR, Kiyomi Komori H, LaMere SA et al (2014) Library construction for next-generation sequencing: overviews and challenges. Biotechniques 56:61–77. https://doi.org/10.2144/000114133

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107:1–8

Heyer R, Schallert K, Zoun R et al (2017) Challenges and perspectives of metaproteomic data analysis. J Biotechnol 261:24–36

Hui AWH, Lau HW, Chan THT, Tsui SKW (2013) The human microbiota: a new direction in the investigation of thoracic diseases. J Thorac Dis 5:127–131

Hyatt D, Chen GL, LoCascio PF et al (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119

Illumina (2015) HiSeq 3000/HiSeq 4000 sequencing systems. In: Illumina

Jamuar SS h, Tan E-C (2015) Clinical application of next-generation sequencing for Mendelian diseases. Hum Genomics 9:10. https://doi.org/10.1186/s40246-015-0031-5

Jeste SS, Geschwind DH (2014) Disentangling the heterogeneity of autism spectrum disorder through genetic findings. Nat Rev Neurol 10:74–81. https://doi.org/10.1038/nrneurol.2013.278

Joshi N, Fass J (2011) Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. https://github.com/najoshi/sickle

Korbel JO, Abyzov A, Mu XJ et al (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10:23. https://doi.org/10.1186/gb-2009-10-2-r23

Lam HYK, Clark MJ, Chen R et al (2012) Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30:78–82. https://doi.org/10.1038/nbt.2065

Lam KN, Cheng J, Engel K et al (2015) Current and future resources for functional metagenomics. Front Microbiol 6:1196. https://doi.org/10.3389/fmicb.2015.01196

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:25. https://doi.org/10.1186/gb-2009-10-3-r25

Lek M, Karczewski KJ, Minikel EV et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291. https://doi.org/10.1038/nature19057

Levy SE, Myers RM (2016) Advancements in next-generation sequencing. Annu Rev Genomics Hum Genet 17:95–115

Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Love M, Anders S, Huber W (2017) Analyzing RNA-seq data with DESeq2. Bioconductor

Luthra R, Chen H, Roy-Chowdhuri S, Singh RR (2015) Next-generation sequencing in clinical molecular diagnostics of cancer: advantages and challenges. Cancers (Basel) 7:14

Mai NTH, Phu NH, Nhu LNT et al (2017) Central nervous system infection diagnosis by next-generation sequencing: a glimpse into the future? Open Forum Infect Dis 4:046. https://doi.org/10.1093/ofid/ofx046

Marceddu G, Dallavilla T, Guerri G et al (2019) Analysis of machine learning algorithms as integrative tools for validation of next generation sequencing data. Eur Rev Med Pharmacol Sci 23:8139. https://doi.org/10.26355/eurrev_201909_19034

Maron PA, Ranjard L, Mougel C, Lemanceau P (2007) Metaproteomics: a new approach for studying functional microbial ecology. Microb Ecol 53:486–493

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10. https://doi.org/10.14806/ej.17.1.200

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12:671

Mathur P, Medicherla KM, Chaudhary S et al (2018) Whole exome sequencing reveals rare variants linked to congenital pouch colon. Sci Rep 8:6646. https://doi.org/10.1038/s41598-018-24967-y

Meena N, Mathur P, Medicherla K, Suravajhala P (2018) A bioinformatics pipeline for whole exome sequencing: overview of the processing and steps from raw data to downstream analysis. Bio-Protocol 8:e2805. https://doi.org/10.21769/BioProtoc.2805

Metzker ML (2010) Sequencing technologies the next generation. Nat Rev Genet 11:31–46

Morgan JL, Darling AE, Eisen JA (2010) Metagenomic sequencing of an in vitro-simulated microbial community. PLoS One 5:e10209. https://doi.org/10.1371/journal.pone.0010209

Moriya Y, Itoh M, Okuda S et al (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:2. https://doi.org/10.1093/nar/gkm321

Mueller JJ, Schlappe BA, Kumar R et al (2018) Massively parallel sequencing analysis of mucinous ovarian carcinomas: genomic profiling and differential diagnoses. Gynecol Oncol 150:127–135. https://doi.org/10.1016/j.ygyno.2018.05.008

Nagalakshmi U, Waern K, Snyder M (2010) RNA-seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol 89:4.11.1–4.11.13

Neveling K, Feenstra I, Gilissen C et al (2013) A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. Hum Mutat 34:1721–1726. https://doi.org/10.1002/humu.22450

Ng PC, Kirkness EF (2010) Whole genome sequencing. Methods Mol Biol 628:215–226

Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52:413–435

Parize P, Muth E, Richaud C et al (2017) Untargeted next-generation sequencing-based first-line diagnosis of infection in immunocompromised adults: a multicentre, blinded, prospective study. Clin Microbiol Infect 23:574. https://doi.org/10.1016/j.cmi.2017.02.006

Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. Genomics 93:105–111. https://doi.org/10.1016/j.ygeno.2008.10.003

Quail MA, Smith M, Coupland P et al (2012) A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13:341. https://doi.org/10.1186/1471-2164-13-341

Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. J Hum Genet 59:5–15. https://doi.org/10.1038/jhg.2013.114

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289

Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139. https://doi.org/10.1093/bioinformatics/btp616

Salzberg SL, Breitwieser FP, Kumar A et al (2016) Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. Neurol - Neuroimmunol Neuroinflammation 3: e251. https://doi.org/10.1212/NXI.0000000000000251

Sanders SJ, Neale BM, Huang H et al (2017) Whole genome sequencing in psychiatric disorders: the WGSPD consortium. Nat Neurosci 20:1661–1668. https://doi.org/10.1038/s41593-017-0017-9

Sarwar B, Karypis G, Konstan J, Riedl J (2000) Application of dimensionality reduction in recommender system—a case study. ACM WebKDD 2000 Web Min ECommerce Work. https://doi.org/10.3141/1625-22

Schirmer M, Ijaz UZ, D'Amore R et al (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 43:37. https://doi.org/10.1093/nar/gku1341

Schloss PD, Westcott SL, Ryabin T et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537. https://doi.org/10.1128/AEM.01541-09

Schubach M, Re M, Robinson PN, Valentini G (2017) Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. Sci Rep 7:2959. https://doi.org/10.1038/s41598-017-03011-5

Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5:16–18

Shendure J (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732. https://doi.org/10.1126/science.1117389

Silahtaroğlu G (2009) An attribute-centre based decision tree classification algorithm. World Acad Sci Eng Technol 36:11282

Simpson JT, Wong K, Jackman SD et al (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123. https://doi.org/10.1101/gr.089532.108

Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458:719–724. https://doi.org/10.1038/nature07943

Stray-Pedersen A, Sorte HS, Samarakoon P et al (2017) Primary immunodeficiency diseases: genomic approaches delineate heterogeneous Mendelian disorders. J Allergy Clin Immunol 139:232–245. https://doi.org/10.1016/j.jaci.2016.05.042

Suravajhala P, Kogelman LJA, Kadarmideen HN (2016) Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. Genet Sel Evol 48:38. https://doi.org/10.1186/s12711-016-0217-x

Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30:418–426. https://doi.org/10.1016/j.tig.2014.07.001

Thomas T, Gilbert J, Meyer F (2012) Metagenomics—a guide from sampling to data analysis. Microb Inform Exp 2:3. https://doi.org/10.1186/2042-5783-2-3

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105. https://doi.org/10.1093/bioinformatics/btp120

Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515. https://doi.org/10.1038/nbt.1621

Tringe SG (2005) Comparative metagenomics of microbial communities. Science 308:554–557. https://doi.org/10.1126/science.1107851

Tripathi R, Sharma P, Chakraborty P, Varadwaj PK (2016) Next-generation sequencing revolution through big data analytics. Front Life Sci 9:119–149. https://doi.org/10.1080/21553769.2016.1178180

van den Akker J, Mishne G, Zimmer AD, Zhou AY (2018) A machine learning model to determine the accuracy of variant calls in capture-based next generation sequencing. BMC Genomics 19:263. https://doi.org/10.1186/s12864-018-4659-0

Van den Veyver IB, Eng CM (2015) Genome-wide sequencing for prenatal detection of fetal single-gene disorders. Cold Spring Harb Perspect Med 5:23077. https://doi.org/10.1101/cshperspect.a023077

Vendramin L, Campello RJGB, Hruschka ER (2010) Relative clustering validity criteria: a comparative overview. Stat Anal Data Min 3:209. https://doi.org/10.1002/sam.10080

Wadapurkar RM, Vyas R (2018) Computational analysis of next generation sequencing data and its applications in clinical oncology. Informatics Med Unlocked 11:75–82. https://doi.org/10.1016/j.imu.2018.05.003

Waern K, Nagalakshmi U, Snyder M (2011) RNA sequencing. Methods Mol Biol 3:209–235. https://doi.org/10.1007/978-1-61779-173-4_8

Wandelt S, Rheinländer A, Bux M et al (2012) Data management challenges in next generation sequencing. Datenbank-Spektrum 12:161–171. https://doi.org/10.1007/s13222-012-0098-2

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:164. https://doi.org/10.1093/nar/gkq603

Weigelt B, Bi R, Kumar R et al (2018) The landscape of somatic genetic alterations in breast cancers from ATM germline mutation carriers. JNCI J Natl Cancer Inst 110:1030–1034. https://doi.org/10.1093/jnci/djy028

Wylie KM, Weinstock GM, Storch GA (2013) Virome genomics: a tool for defining the human virome. Curr Opin Microbiol 16:479–484. https://doi.org/10.1016/j.mib.2013.04.006

Bioinformatics in Personalized Medicine 15

G. Sunil Krishnan, Amit Joshi, and Vikas Kaushik

Abstract

Genomics delivers purposeful biological information, as it is a part of life science dealing about the comprehension and planning of genomes. A genome is the complex arrangement of genetic sets present in a cell or a whole living being. It is a useful measure of information when you consider that the human genome has in excess of 3 billion DNA base sets. It is a stunning measure of data that people have experienced difficulty in wielding, despite the fact that nature figured out how to pack everything into each cell in the human body. Customized medication is clinical consideration related to every patient's hereditary cosmetics. It implies mass, mechanical production system like medication reaches a conclusion, and medication intended to convey greatest advantage to the individual turns into the standard. This would kill a great deal of awful side-effects related with standard medicines presently, decrease or dispose of hypersensitive responses, diminish the expense of medical care, and lessen patient sufferings, as enduring more successful therapies. So as to really perform customized medication, every patient's genome should initially be converted into advanced information which is then handled, put away, and recovered varying. Accordingly the triple play of genomics, bioinformatics and customized medication is vital. Everything sounds so basic yet it is so confounded. Numerous medications and preventive therapies are neglected to convey ideal reaction to wide populace. PM is a combinational way to deal with individual specific heath care. The patient specific medication development advanced through bioinformatics tools. Bioinformatics devices may furnish better conclusions in genomic level with prior identification of infection and better focused on treatment through productive PM improvement. Variety in

G. Sunil Krishnan · A. Joshi · V. Kaushik (✉)
Lovely Professional University, Phagwara, Punjab, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_15

sex, racial, ethnic, genetic polymorphisms, and other ecological variables influence the in resistant reaction to a specific therapy.

Keywords

Personalized medicine · Bioinformatics · Genome · Health care · Human body

15.1 Introduction

Medication and medicines are changing due to the advancement of technology development. Many people are suffering in the world due to rare diseases and adverse effect of a therapy. This was because of the unavailability of efficacious drug or personalized medicine (PM). PM is an individual targeted tailor made medication to reduce the drug or disease associated risk this can leads to provide more efficacious treatment. PM is also known as precision or stratified medicine. In this specific approach individual patient's genomics profile plays an important role to find safe and efficacious treatment. PM has the prospective tools to manage different incurable disease stages from detection to prevention. Next-Gen Sequencing (NGS) innovation, frequently observed as the establishment of personalized medication, has been effectively applied in oncology diagnostics and immunotherapy. With propels in quality diagnostics and immunotherapy, there might be an opportunity to control the advancement of malignant growths and mitigate the enduring of patients going through chemotherapy. To advance the interpretation of exactness medication from seat to alongside and from utilization of hereditary testing to customized medication, new investigation techniques for NGS and hereditary information should be created. For instance, the NGS board is very unique in relation to entire or whole-genome sequencing (WGS), focus on less gene sets or locales yet requiring more noteworthy exactness and effectiveness. For complex maladies, for example, malignant growths, the driver genetic elements are normally a bunch of qualities in a positive or negative regulatory organization. Chart speculations, for example, briefest way examination and irregular walk calculations, will help dismember entire genomic communications into key modules or ways whose brokenness is related with infectious propagation. The genomics technologies enabling the search and filter genes and their variants from the whole genome and the pharmacogenomics identify the patient specific drug associated variant genes. In the practical approach of personalized medicine patient's genomic and proteomic data processed to digital data. Then the stored data retrieved and analyzed through bioinformatics tool for this tailor made genomics-based drug discovery and therapy. This selects right drug and dose based on individual's genomic data processed, analyzed, clinical, and along with disease environmental data. This field is young and evolving in the healthcare system. The pharmaco-dynamics, genomics, and kinetics involvement has an important role in the succession of PM proceedings. The information and communication technology (ICT) manage patient's personal and medical data (Louca 2012). Bioinformatics and data mining are combined to

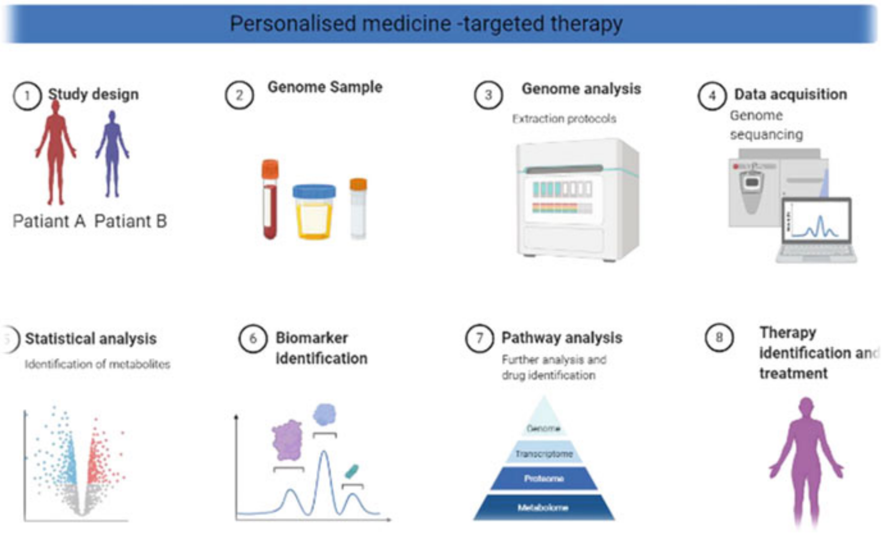


Fig. 15.1 Bioinformatics and genomics protocol in personalized medicine

create tools and procedures for the prediction of emerging, recurrence, progression, response of disease to treatment. Individualizing drug or vaccine therapy with the use of bioinformatics and pharmacogenomics tools have the prospective to transform health care system (Mancinelli et al. 2000). The whole-exome sequencing have proven to be valuable methods for the discovery of the genetic causes of rare and complex diseases (Gonzaga-Jauregui et al. 2012). Bioinformatics and genomics protocol in personalized medicine summarized in Fig. 15.1.

15.2 Significance of Personal Medicine and Bioinformatics

Many drugs and vaccines are failing to deliver optimum response to broad population. PM is a combinational approach to individual health care. This is required for the improvement of early disease diagnosis and treatment at individual level. Each individual's pre- or post-disease clinical, genomic, and environmental information are not unique. Genome-wide association studies helped to identify genes important in serious adverse drug reactions (Daly and Day 2012). In the most recent decade, biochemical science has made numerous advances to personalized medication, including the Human Genome venture, International HapMap task, and genome-wide affiliation contemplates (GWASs). Single nucleotide polymorphisms (SNPs) are currently perceived as the fundamental driver of human hereditary fluctuation and are as of now a significant asset for planning complex hereditary characteristics. A great many DNA variations have been distinguished that are related with ailments and attributes. By joining this hereditary relationship with phenotypes and medication reaction, customized medication will tailor medicines to the patients' particular

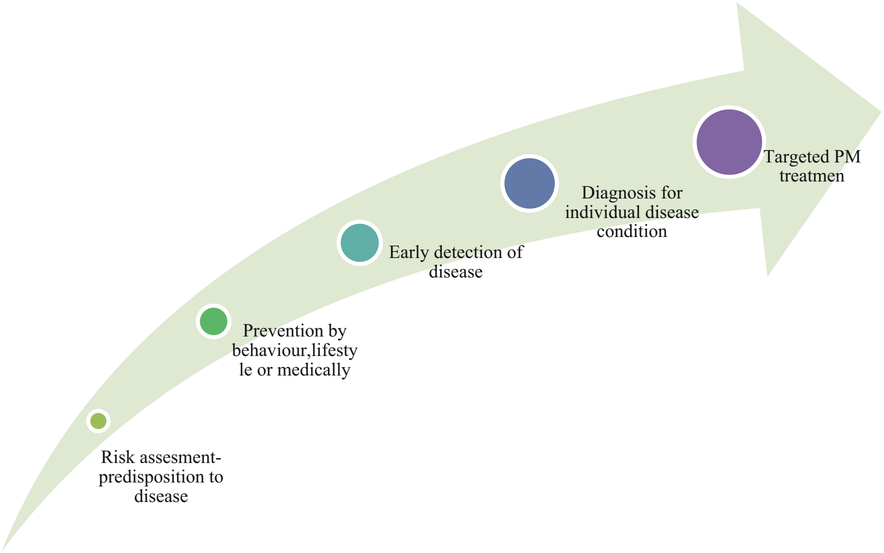

Risk assesment-predisposition to disease
Prevention by behaviour,lifestyle or medically
Early detection of disease
Diagnosis for individual disease condition
Targeted PM treatmen

Fig. 15.2 Stages of personal medicine design

genotype. Albeit entire genome groupings are not utilized in ordinary practice today, there are as of now numerous instances of customized medication in current practice. Figures 15.2 and 15.3 explain various stages and steps of personal medicine design, respectively. Chemotherapy prescriptions, for example, trastuzumab and imatinib target explicit diseases, a focused on pharmacogenetic dosing calculation is utilized for warfarin and the frequency of unfavorable occasions is decreased by checking for powerless genotypes for drugs like abacavir, carbamazepine and clozapine.

Customized medication is required to profit by consolidating genomic data with customary checking of physiological states by different high-throughput methods. Over the previous decade, upgrades in instrument affectability, speed, exactness, and throughput, combined with the improvement of innovations, for example, various responses observing. Under the direction of the Human Proteome Organization over 80% of the proteins anticipated by the human genome have now been recognized utilizing either mass spectrometric or immunizer based procedures, and the staying "missing proteins" are as a rule consistently represented. Assets, for example, the Human MRM Atlas, a far reaching asset intended to empower researchers to perform quantitative examination of every human protein, are being created to encourage reproducible exchange of quantitative tests between labs. Such turns of events and activities currently empower both top to bottom disclosure and focused on/quantitative work processes, making the way for the clinical analytic field. Combined with this, the foundation of exhaustive information bases and the improvement of amazing in silico methods is empowering viable information mining. Specifically this has empowered interactome examines permitting the recognizable proof of key flagging pathways prompting potential new medication

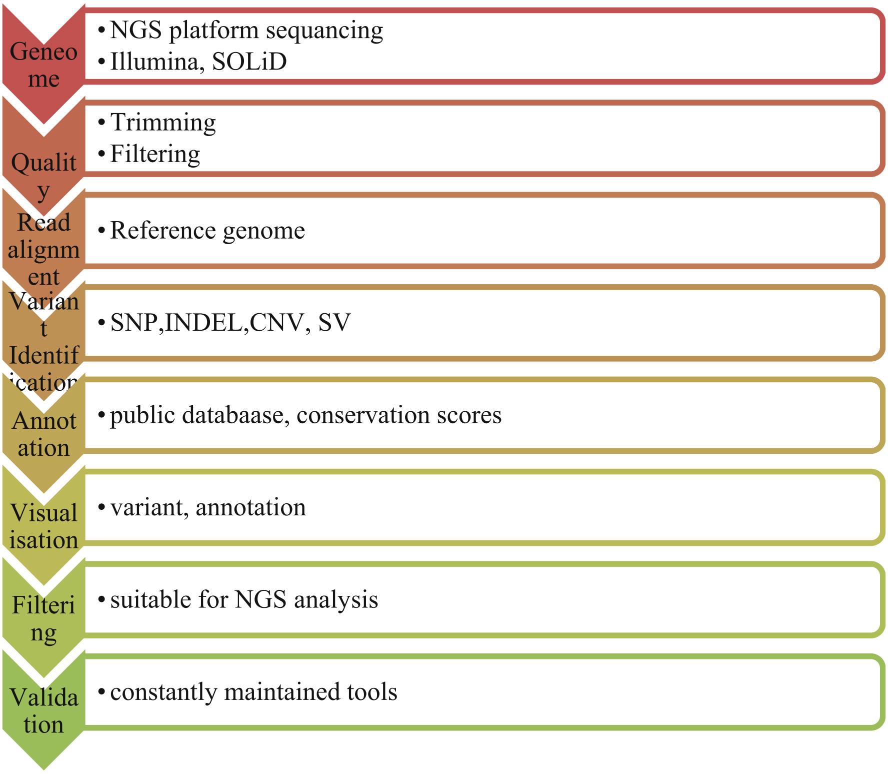


Fig. 15.3 Bioinformatics steps in genomic data processing for PM

targets, despite the fact that to date it has been assessed that under 20% of the protein communications in people, not including dynamic, tissue-or infection explicit associations, have been distinguished (Chen et al. 2012).

15.3 Application of Bioinformatics in Personal Medicines and Vaccines

Bioinformatics tools may provide better diagnoses in genomic level with earlier detection of disease and better targeted therapy through efficient personalized medicine development. Omics analysis provides a great assistance in the development of personalized medicine (Fig. 15.4). Bioinformatics tools helps in diagnosis, intervention, drug development, therapy, and personalized vaccination. Personalized vaccine means for an optimized prevention of disease with minimized reactogenicity and side effect. Personalized vaccines are developed to take care of haplotypes and polymorphism can become risk of an adverse vaccine reaction. Variation in gender, racial, ethnic, gene polymorphisms and other environmental factors affect the in

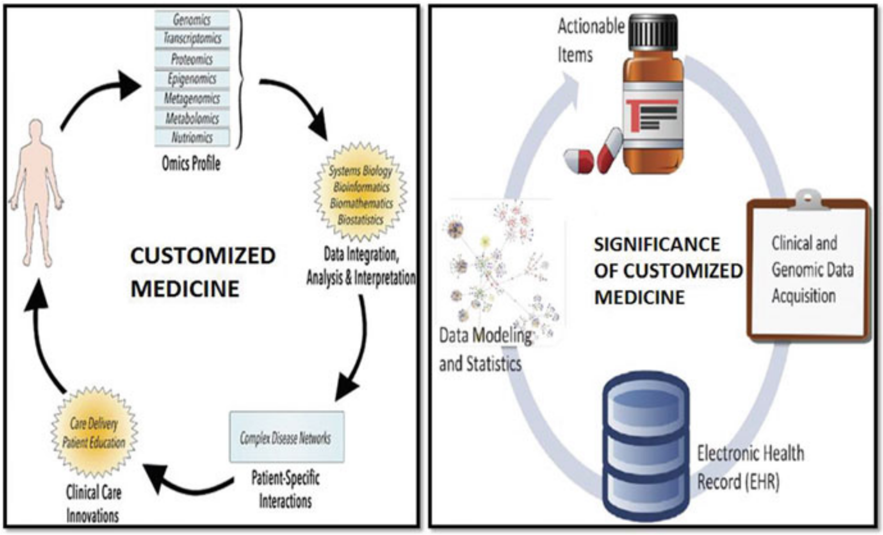


Fig. 15.4 Customized or personal medicine based on omics analysis

immune response to a particular vaccine where we required a personalized vaccine and drugs (Atsaves et al. 2019; Poland et al. 2011). Insilico designing helps vaccines and vaccine adjuvants design for immunologically different groups (Piasecka et al. 2018; Oli et al. 2020). The identification of Human Leukocyte Antigen (HLA) complex and stable polymorphism and effective vaccine for the each individual is possible through bioinformatics analysis of HLA class I and II molecules and predict suitable peptide for HLA binding (Gfeller and Bassani-Sternberg 2018). Due to the expression differences of MHC- HLA allele to viral proteins the T cell responses varies (Clemens et al. 2018; Auladell et al. 2019).The artificial neural network algorithms and datasets made possible to develop different epitope predicting tools. The tools help to predict the epitope peptides for a particular MHC-HLA binding of an individual (Chandra and Yadav 2016). Peptide motifs and MHC ligands databases obtained from epitope peptide prediction servers (Lundegaard et al. 2010; Glutting and Reinherz 2002). Immunoinformatics prediction of Immunodominant epitopes (SSNLYKGVY) from AA41-49 of glycoprotein 1 of *Lassa fever* virus can induce of humoral and cell-mediated immunity African populations and endemic country (Hossain et al. 2018) and in *Oropouche* virus (Adhikari et al. 2018). These approaches help individualized vaccination and prevent endemic diseases. The diversity of HLA regions suspected for the generation vaccine immune response in each individual (Kaifu and Nakamura 2017). The genetic variation in male and female may leads to differences in immune responses against influenza (Voigt et al. 2019; Fink et al. 2018), rubella (Mitchell et al. 1992), and measles immunization (Fischinger et al. 2019). The new cutting edge technology like vaccinomics a combination of immunogenomics, bioinformatics and immunogenetics could be helpful in the personalized vaccine development (Majumder

2015). The personalized vaccinology and medicine developed through international HapMap and that of the Human Genome Project. The variation in gene level, linkage disequilibrium maps, and single nucleotide polymorphism (SNP), have significant roles in immune responses (Brodin and Davis 2017; Cotugno et al. 2019). The sequencing technologies, bioinformatics analysis tools, genotypic and phenotypic data bases advances the immune response prediction of drugs, vaccines, insecticides and diagnostics (Gunawardena and Karunaweera 2015). Immunoinformatics studies were found to be successful in predicting epitope based vaccines for SARS-Cov2 (Joshi and Kaushik 2020; Akhtar et al. 2020), Dengue virus (Sunil Krishnan et al. 2020), and Nipah virus (Kaushik 2019), even the rarest bacterium like *Tropheryma whipplei* causing lipodystrophy could also be successfully targeted by epitope based vaccine formulations (Joshi et al. 2020).

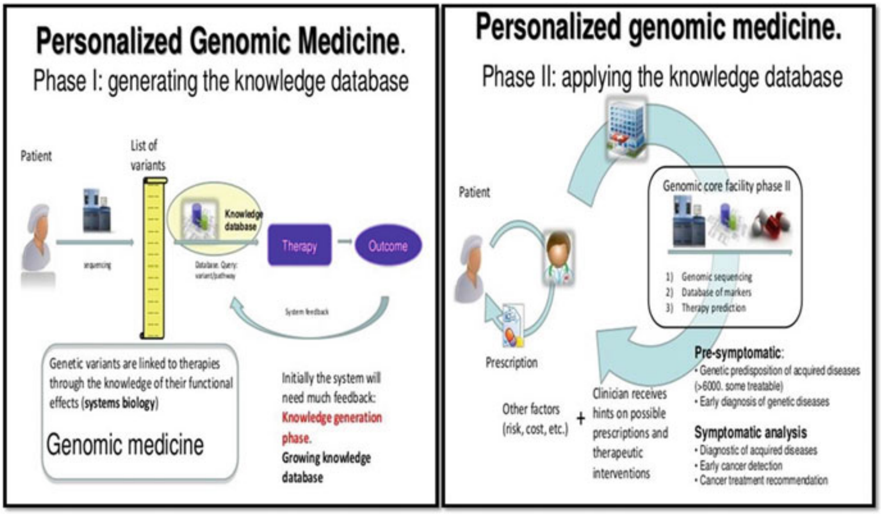
15.4 Advantages and Disadvantages of PM

The upsides of PM would be relevant in the uncommon and complex ailment the board by refining patients and care suppliers, quicken exploration, and supporting vital changes in strategy and guideline. The new bioinformatics explores have been planning apparatuses and test pipelines to investigate singular affliction circumstance. The progressing customized medication has been in understanding consideration relevant to cardiovascular sicknesses, Mendelian problems, malignant growths, Kabuki condition and hereditarily heterogeneous issues (schizophrenia, irregular mental imbalance and range issues) (Table 15.1).

Phases involved in genomics analysis (Fig. 15.5) are sequencing by deploying next-generation approach like illumina solexa, 454 pyrosequencing, Ion torrent, etc. After sequencing genetic sets are analyzed to detect epigenetic relationships, to determine phylogenetic expressions involved to accumulate information in databases that can be used for personalized medication formulations.

Table 15.1 Personal medicine's advantage and disadvantages

S. no.	Advantages	Disadvantages
1	Minimize the incidence of adverse effect of treatment	Expensive and not accessible to everyone
2	Understanding of the individual patient or population treatment need	Economically impossible to target small patient populations
3	Interpret genetic information	Technology not licensed
4	Advancing personalized medicine in patient care	Fear of data leakage
5	Greater precision in diagnosis and more targeted drug development	Service providers are not common
6	For rare and complex diseases	Not in all cases successive
7	Increasing the accuracy of diagnosis	Need more tools to be developed to interpret the data


Personalized Genomic Medicine.

Phase I: generating the knowledge database

Genetic variants are linked to therapies through the knowledge of their functional effects (**systems biology**)

Genomic medicine

Initially the system will need much feedback: Knowledge generation phase. **Growing knowledge database**

Personalized genomic medicine.

Phase II: applying the knowledge database

1) Genomic sequencing
2) Database of markers
3) Therapy prediction

Other factors (risk, cost, etc.) + Clinician receives hints on possible prescriptions and therapeutic interventions

Pre-symptomatic:
- Genetic predisposition of acquired diseases (>6000. some treatable)
- Early diagnosis of genetic diseases

Symptomatic analysis
- Diagnostic of acquired diseases
- Early cancer detection
- Cancer treatment recommendation

Fig. 15.5 Phases involved in genome based personalized medicine development

15.5 Bioinformatics Prerequisites Challenges for Personal Medicine Design

Computational and sequencing Infrastructure, availability of individual genomic data, Data quality, bioinformatics data analysis tools, computational pipelines, interpretation, and validation of biomarkers. The processed data analysis has five analytical steps like quality assessment, alignment, variant identification, variant annotation, and visualization. For the advancement of PM many challenges have to be overcome. The availability patient's genomic data are consulted only for a little treatment plans and hardly few medical centers used for treatment (Yngvadottir et al. 2009). Bioinformatics tools would help the diverse genomics data for PM design for individual patients. The challenge includes availability of computational or sequencing infrastructure, error rate in individual genomic data (1000 Genomes Project Consortium et al. 2010), data quality, bioinformatics data analysis tools, computational pipelines for large data processing, and validation of biomarkers. The processing such large amounts of genetic data obtained from next-generation sequencing (NGS) requires bioinformatics dataprocessing . High amount of data and its accuracy challenges for the analysis and interpretation. Through NGS the detection of Copy number variants (CNVs) and structural variants (SVs) are more difficult (Shendure and Ji 2008). Bioinformaticians have developed new algorithms for tools like BLAST (Altschul et al. 1990), BLAT (Kent 2002), BWA (Li and Durbin 2009), and MOSAIK, developed by the Marth Lab (Michael Stromberg, Boston University) to address these problems in different times.

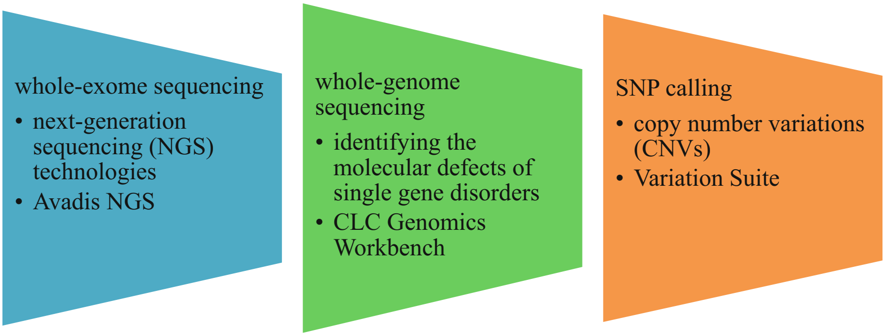

whole-exome sequencing
• next-generation sequencing (NGS) technologies
• Avadis NGS
whole-genome sequencing
• identifying the molecular defects of single gene disorders
• CLC Genomics Workbench
SNP calling
• copy number variations (CNVs)
• Variation Suite

Fig. 15.6 General Sequencing and bioinformatics tools used in the PM design

The whole-genome and whole-exome data interpretation has an important role in the experimental success (Schadt et al. 2010). General Sequencing tools used in the PM design summarized in Fig. 15.6. The appropriate choice of tools, data handling and tool compatibility programs for variant analysis of NGS data. Intellectual property rights, reimbursement policies, patient privacy, data bases and confidentiality as well as regulatory oversight. The Vaccine, therapeutics or drug development process, and regulatory requirement need to be changed for targeting smaller patient populations with rare diseases.

15.6 Advanced Methods Involved in Personalized Medicine Designing

Coordinating a lot of information originating from high-throughput advances towards customized medication and diagnostics cannot be conceivable without utilizing computational ways to deal with sort out the unpredictability of handling and associating numerous factors at the "omics" level. Bioinformatics is an interdisciplinary field of science that is centered on applying computational methods for the investigation and separating data from information originating from biomolecules. Typically, it coordinates methods from the fields of informatics, software engineering, sub-atomic science, genomics, proteomics, arithmetic, and measurements. In spite of the fact that it began as a field completely committed to essential exploration in advancement and hereditary qualities, it has been advancing in corresponding with high-throughput strategies bringing about the improvement of numerous techniques and apparatuses that encourage the translation of "omics" information. In bioinformatics, high-throughput information is prepared and broke down methodically from crude information to the outcomes utilizing pipelines of examination utilizing the maximum capacity of PCs. Bioinformatics pipelines typically contain various strides for information quality evaluation, include extraction, measurement decrease, biomarker recognition, and results age. This arrangement of examination is completely robotized where the client has no obstruction except for can assume the part of "caretaker" to check the approval of the yields (results).

With the advancement of the computational force, bioinformatics picked up the possibility to handle huge information and incorporate a lot of information a lot quicker than it is delivered, turning into an answer applying high-throughput methods in clinical diagnostics and customized medication. For instance, a few examinations have shown that bioinformatics pipelines produced for the investigation of MALDI-ToF mass spectra can extricate symptomatic data from pee, blood, and undeveloped organism culture media quicker than its ability of being created. In genomics, a few bioinformatics pipelines of examination for NGS, RNAseq, and microarrays have been additionally evolved to remove analytic data out of sequencing of infection, obsessive microorganisms, and malignant growth biopsies. In addition, bioinformatics instruments for preparing "omics" have likewise been fruitful in the revelation of novel medication focuses for malignant growth treatment. Bioinformatics can additionally improve clinical research facilities proficiency and expenses by sparing time and HR on the investigation and answering to centers and patients. This should be possible by creating pipelines of examination with computerized revealing and APIs completely committed to giving constant online access, encouraging the correspondence between labs, clinicians, and patients. Also, persistent chronicled information and metadata ought to be secure and sorted out in an organized manner (information "stockrooms") with the end goal that it very well may be additionally pulled efficiently to bioinformatics pipelines. This would permit going past in integrative examination of patients by having their information as a component of time permitting a more customized checking of patients indicative and permitting better prognostics.

Apparatuses with direct significance to customized medication

1. Biomarker-driven medication: multi-omics, IT, approval, reproducibility, clinical utility.
2. Genomics information translation, in addition to phenotypes.
3. Man-made consciousness, Machine Learning, Simulation.
4. Resident Science, Biobanks, Health Data Cooperatives.
5. European frameworks for customized medication (for example, open science cloud).

The advancement of numerical models and calculations that produce strong expectations is a hard undertaking and requires thorough approval techniques before an indicator is fit to be dispatched into the market. Relatively few indicators for diagnostics are accessible to be utilized or can be adjusted to a given clinical lab setting. Accordingly, model turn of events and enhancement for every lab would be the ideal situation. Incorporating prescient displaying work processes in bioinformatics pipelines likewise encourage model advancement by organizing the cycle of approval and model determination utilizing the information and metadata. A few kinds of models can be utilized to settle on indicative expectations and the decision relies upon the information accessible, innovation, and the idea of the issue.

Measurable models dependent on known appropriations of biomarkers are normal to be utilized in the demonstrative of a specific illness. These are anything but difficult to actualize in bioinformatics pipelines and fill in as correlative data for clinicians. Execution of example acknowledgment, AI, and man-made brainpower (AI) calculations into bioinformatics pipelines are critical to enhance numerical models towards meeting more exact forecasts. Critically, the utilization of AI and AI calculations are fundamental for customized medication since they empower the fitting of conventional models of illness to every patient situation and body science. Deterministic models, for example, the coherent and dynamic demonstrating structures can likewise be utilized for reenactment of physiological situations and making powerful forecasts with clinical applications.

For instance, reproduction of the tumor miniature condition utilizing a consistent organization model of the guideline of cell attachment properties permitted to build up relations between malignancy de-guidelines and the metastatic potential. This has a tremendous potential for the future improvement of bioinformatics apparatuses that permit the expectation of the metastatic potential and propose the best treatment for each case dependent on the tumor biopsy. Dynamic models, then again, can possibly be more exact and produce a persistent scope of expectation esteems. Notwithstanding, their boundary assessment is perplexing and requires AI calculations to adjust them to a specific physiological framework. These kinds of models are phenomenal for portraying the digestion and can be valuable in as future apparatuses in customized medication (Dodin 2017).

15.7 Conclusions

The enhancement of bioinformatics tools and databases development for new diseases would be helpful for the advancement of PM. In future looking for more coverage, affordability in genome data processing, accuracy in data interpretation, fast genetic data processing, development of more bioinformatics tools the understanding of disease at the molecular level, and bioinformatics advancement for data interpretation. This would help the elevated success rate in this new young health care field. Discovery of bioinformatics tools would help to integrate the huge genomics data analysis and speed up the PM research. As the environment of partners attempts to progress customized medication, cooperation with government controllers and policymakers is important to empower inescapable utilization of these new devices and technology advances. The administrative cycle must develop in light of advances that are focused to littler patient populaces dependent on hereditary profiles, and arrangements and enactment must be sanctioned that give motivating forces to inventive exploration and selection of new advances. Together, progress in the exploration, clinical concern, and strategy empowering customized personal therapy can possibly improve the nature of patient consideration and to help contain medical services costs.

References

Adhikari UK, Tayebi M, Rahman MM (2018) Immunoinformatics approach for epitope-based peptide vaccine design and active site prediction against polyprotein of emerging oropouche virus. J Immunol Res 2018:1–22

Akhtar N, Joshi A, Singh B, Kaushik V (2020) Immuno-informatics quest against COVID-19/ SARS-COV-2: determining putative T-cell epitopes for vaccine prediction. Infect Disord Drug Targets 20:32957905. https://doi.org/10.2174/1871526520666200921154149

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Atsaves V, Leventaki V, Rassidakis GZ, Claret FX (2019) AP-1 transcription factors as regulators of immune responses in cancer. Cancer 11(7):1037

Auladell M, Jia X, Hensen L, Chua B, Fox A, Nguyen TH, Kedzierska K (2019) Recalling the future: immunological memory toward unpredictable influenza viruses. Front Immunol 10

Brodin P, Davis MM (2017) Human immune system variation. Nat Rev Immunol 17(1):21–29. https://doi.org/10.1038/nri.2016.125

Chandra H, Yadav JS (2016) Human leukocyte antigen (HLA)-binding epitopes dataset for the newly identified T-cell antigens of mycobacterium immunogenum. Data Brief 8:1069

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R et al (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148(6):1293–1307

Clemens EB, Van de Sandt C, Wong SS, Wakim LM, Valkenburg SA (2018) Harnessing the power of T cells: the promising hope for a universal influenza vaccine. Vaccine 6(2):18

Cotugno N, Ruggiero A, Santilli V, Manno EC, Rocca S, Zicari S, Amodio D, Colucci M, Rossi P, Levy O, Martinon-Torres F, Pollard AJ, Palma P (2019) OMIC technologies and vaccine development: from the identification of vulnerable individuals to the formulation of invulnerable vaccines. J Immunol Res 2019:8732191. https://doi.org/10.1155/2019/8732191

Daly AK, Day CP (2012) Genetic association studies in drug-induced liver injury. Drug Metab Rev 44(1):116–126

Dodin G (2017) Personal genomics: new concepts for future community data banks. bioRxiv. https://doi.org/10.1101/230516

Fink AL, Engle K, Ursin RL, Tang WY, Klein SL (2018) Biological sex affects vaccine efficacy and protection against influenza in mice. Proc Natl Acad Sci 115(49):12477–12482

Fischinger S, Boudreau CM, Butler AL, Streeck H, Alter G (2019) Sex differences in vaccine-induced humoral immunity. In: Seminars in Immunopathology. Springer, Berlin Heidelberg, pp 239–249

Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061

Gfeller D, Bassani-Sternberg M (2018) Predicting antigen presentation—what could we learn from a million peptides? Front Immunol 9:1716

Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012) Human genome sequencing in health and disease. Annu Rev Med 63:35–61

Gunawardena S, Karunaweera ND (2015) Advances in genetics and genomics: use and limitations in achieving malaria elimination goals. Pathogens Global Health 109(3):123–141. https://doi.org/10.1179/2047773215Y.0000000015

Hossain MU, Omar TM, Oany AR, Kibria KK, Shibly AZ, Moniruzzaman M, Islam MM (2018) Design of peptide-based epitope vaccine and further binding site scrutiny led to groundswell in drug discovery against Lassa virus. 3 Biotech 8(2):81

Joshi A, Kaushik V (2020) In-silico proteomic exploratory quest: crafting T-cell epitope vaccine against Whipple's disease. Int J Pep Res Therapeut 18:1–11. https://doi.org/10.1007/s10989-020-10077-9

Joshi A, Joshi BC, Mannan MA, Kaushik V (2020) Epitope based vaccine prediction for SARS-COV-2 by deploying immuno-informatics approach. Inform Med 19:100338. https://doi.org/10.1016/j.imu.2020.100338

Kaifu T, Nakamura A (2017) Polymorphisms of immunoglobulin receptors and the effects on clinical outcome in cancer immunotherapy and other immune diseases: a general review. Int Immunol 29(7):319–325

Kaushik V (2019) Silico identification of epitope-based peptide vaccine for Nipah virus. Int J Pept Res Ther 26(2):1147–1153. https://doi.org/10.1007/s10989-019-09917-0

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12(4):656–664

Li H, Durbin R (2009) Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics 25(14):1754–1760

Louca S (2012) Personalized medicine–a tailored health care system: challenges and opportunities. Croat Med J 53(3):211–213

Lundegaard C, Lund O, Buus S, Nielsen M (2010) Major histocompatibility complex class I binding predictions as a tool in epitope discovery. Immunology 130(3):309–318

Majumder PP (2015) Genomics of immune response to typhoid and cholera vaccines. Philos Trans Royal Socf London *370*(1671):20140142. https://doi.org/10.1098/rstb.2014.0142

Mancinelli L, Cronin M, Sadée W (2000) Pharmacogenomics: the promise of personalized medicine. AAPS PharmSci 2(1):29–41

Mitchell LA, Zhang T, Tingle AJ (1992) Differential antibody responses to rubella virus infection in males and females. J Infect Dis 166(6):1258–1265

Oli AN, Obialor WO, Ifeanyichukwu MO, Odimegwu DC, Okoyeh JN, Emechebe GO, Ibeanu GC (2020) Immunoinformatics and vaccine development: an overview. ImmunoTargets Ther 9:13

Piasecka B, Duffy D, Urrutia A, Quach H, Patin E, Posseme C, Hasan M (2018) Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. Proc Natl Acad Sci 115(3):E488–E497

Poland GA, Kennedy RB, Ovsyannikova IG (2011) Vaccinomics and personalized vaccinology: is science leading us toward a new path of directed vaccine development and discovery? PLoS Pathog 7(12):e1002344

Reche PA, Glutting JP, Reinherz EL (2002) Prediction of MHC class I binding peptides using profile motifs. Hum Immunol 63(9):701–709

Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11(9):647–657

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10):1135–1145

Sunil Krishnan G, Joshi A, Kaushik V (2020) T cell epitope designing for dengue peptide vaccine using docking and molecular simulation studies. Mol Simul 46(10):787–795. https://doi.org/10.1080/08927022.2020.1772970

Voigt EA, Ovsyannikova IG, Kennedy RB, Grill DE, Goergen KM, Schaid DJ, Poland GA (2019) Sex differences in older adults' immune responses to seasonal influenza vaccination. Front Immunol 10:180

Yngvadottir B, MacArthur DG, Jin H, Tyler-Smith C (2009) The promise and reality of personal genomics. Genome Biol 10(9):237

Bioinformatics Tools for Gene and Genome Annotation Analysis of Microbes for Synthetic Biology and Cancer Biology Applications

16

Ekene Emmanuel Nweke, Essa Suleman, Morne Du Plessis, and Deepak B. Thimiri Govinda Raj

Abstract

In this book chapter, we focus on application of genome annotation and analysis of microbes for synthetic biology and cancer biology research. We particularly emphasised on application of microbial genomics in synthetic biology and cancer biology. Finally, we delineated future perspective and potential route map for improving the microbial genome annotation and microbial genomics analysis. We infer that our future perspective strategies would assist in reshaping the genome annotation of microorganisms along with microbial genomics analysis. In addition, with better understanding on microbial genome annotation and microbial genomics analysis, we believe this would better enable synthetic biology and cancer biology applications.

E. Emmanuel Nweke
Department of Surgery, University of Witwatersrand, Johannesburg, South Africa

E. Suleman
Veterinary Molecular Diagnostics and Vaccines, Medical Devices and Diagnostics Impact Area, CSIR, Pretoria, South Africa

M. Du Plessis
Bioinformatics and Comparative Genomics, South African National Biodiversity Institute, Pretoria, Gauteng, South Africa

D. B. Thimiri Govinda Raj (✉)
Synthetic Nanobiotechnology and Biomachines, Centre for Synthetic Biology and Precision Medicine, CSIR, Pretoria, South Africa

Biotechnology Innovation Centre, Rhodes University, Grahamstown, South Africa

Department of Clinical Medicine, WITS University, Johannesburg, South Africa

University of Pretoria, Pretoria, South Africa
e-mail: dgovindaraj@csir.co.za

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_16

Keyword

Gene annotation · Microbial genomics · Synthetic biology · Systems biology · Cancer biology

16.1 Introduction

Computational biology has played the key role in recent advancement of biological sciences and biotechnology research (Mulder et al. 2016). Bio-IT has supported rapid progress across all the fields due to its ability to store and analyse larger datasets. This approach enabled use of computational approach across all technology innovation and development. Some of recent applications for computational biology in biological sciences research and development include human genome project, human epigenome project, Cancer genome and human microbiome atlas (Hood and Rowen 2013). With advancement for data analytics, high performance statistical analysis, block chain analysis and artificial intelligence, computational biology has empowered and enabled research innovation advancement at higher level (Özdemir et al. 2017). Particularly, pharmaceutical companies are applying the computational biology tools such as biostatistics, artificial intelligence and pharmacogenomics for its drug discovery development pipeline. During the recent COVID19 pandemic, rapid progress in decoding structural and functional relevance of the SARS-CoV2 protein was achieved through high performance cloud computing based Cryo-EM studies (Wintjens et al. 2020).

One example of computational biology applications in the microbial synthetic biology research space is where Govindaraj et al. designed and constructed a semi-synthetic baculovirus genome called SynBac using synthetic biology techniques such as homologous recombination and Cre-loxP. SynBac is currently sold in the market by Geneva Biotech, an EMBL spin-out company (SynBac™ 2020). SynBac was build based on computational analysis, gene annotation and genome map of *Autographa californica multiple nucleopolyhedrovirus* (ACMNPV) (Berger and Raj 2014). Briefly, gene annotation and comparative genome analysis were performed on baculovirus genome across the various genome families. Using the genome annotation and data mining on functional role of the baculovirus genes, Govindaraj et al. were able to delineate the essential and non-essential gene annotation in AcMNPV genome (Vijayachandran et al. 2013). Further, they used cutting-edge synthetic biology technologies such as homologous recombination and Cre-loxP to rewire AcMNPV genome and invented SynBac (SynBac 1.0) which is currently applied for recombinant protein production (FEBS-EMBO Conference proceedings 2014). Similarly, several teams are using cutting-edge bioinformatics technologies to assist with the development of synthetic microbes for biotechnology and biological applications such as cancer biology. For example, the Synthetic Nanobiotechnology and Biomachines team at CSIR, South Africa is currently establishing bioinformatics driven synthetic biology approach to produce synthetic microbes for recombinant protein production in industrial scale. Another application of computational biology is in precision medicine research where the statistics, artificial intelligence and data analysis play key role in clinical analysis of various patient cohort. Some of the computational biology applications in precision medicine include use of computational science in genomics, RNA-seq analysis, proteomics, pharmacogenomics and drug sensitivity screening. In this regard, a cancer drug sensitivity screening platform that combines drug repurposing platform with computational analysis such as drug sensitivity score (DSS) and automation is currently established at CSIR (Raj et al. 2018a, b).

In this book chapter, we have a focused review on the gene and genome annotation analysis of microorganisms along with its applications in synthetic biology and cancer biology applications. Our focus review is delineated with several subsections that includes the current status of gene annotation (automated and manual) in microbes, gene features on the prokaryotic genome, comparative analysis on gene annotation between virus, bacteria and other microbes, gene ontology along with community annotation in cancer biology, application of microbial genomics in cancer biology and application of synthetic microbes for cancer treatment. Finally, we report in-depth future perspective analysis with the applications for genome annotation of microbes for synthetic biology and cancer biology both in applied research and clinical setting.

16.2 Current Status of Gene Annotation (Automated and Manual) in Microbes

The annotation of microbial genes is an ever evolving process in terms of the implementation of specific annotation algorithms (which has reached a degree of stability in terms of approach) as well as the integration of a number of often freestanding tools into more complex and diverse pipelines. There are of course the proponents of manual annotation that argue that the automated annotation models have the consequence of often introducing errors into an annotation system that is continually perpetuated. As such they emphasise the continued need to grow accurate model databases (Danchin et al. 2018).

The classification of gene annotation tools, in terms of the approach, continues to involve the following: (a) structure and function (e.g. Proteogenomics), (b) sequence similarity (e.g. transfer of functional information between highly similar sequences) and (c) ab initio based approaches (which relates to identifying signals associated with gene features, such as the implementation of neural networks and machine learning approaches).

In order to increase the confidence in accuracy of annotation, there is a continued need to accumulate and develop experimentally derived resources which serves as a basis for validation of the accuracy of all of these approaches. These include the continued growth in the deposit of microbial sequences (gene sequences, genome sequences, protein sequences, transcript derived sequences and protein structures). The latest developments in the field have resulted in the existences of a number of specialised databases, rather than the standard broad database approach which has lower levels of curation accuracy, as a result of its diverse nature. The nature of the specialisation more often relates to grouping the annotation references by related taxa, whilst others rather focus on various biological categories, which are broadly termed as subject-specific databases.

Given the development of advances and decreasing costs of next generation sequencing much of the gene annotation algorithms have now been incorporated as part of genome annotation workflows. Whilst they are developed with genome scale data in mind, the principles around annotating individual genes or shorter gene regions remain valid. As such, an aspect of the annotation pipeline involves the annotation of all features of a genome, including the genes. The most widely used tools in this context are the PGAP (Tatusova et al. 2016) and Prokka (Seemann 2014) as observed from genome publications in current journals. As well as serving the role of genome annotation tools, they also serve as gene annotation tools, pending the length of the input sequence.

Often the genome annotation workflows have a version of the most recognisable gene annotation tools integrated in their workflows. In terms of frequency of usage, the most utilised gene annotation tools are Glimmer, GeneMark and Prodigal. Glimmer and GeneMark use various Markov models, whilst Prodigal, in turn, uses a log-likelihood function, which particularly performs better with high GC content genomes. Collectively, these tools are capable of identifying about of 97% of genes (Hyatt et al. 2010; Lomsadze et al. 2018) in a general annotation setting. When looking at the manual curation environment, again, not much has changed in terms of the general approach. The most significant advance in manual annotation has been through utilising more annotators in the same space of time, or in the same sample annotation environment. This as such relates to decreasing the annotation time, for manual curation by including more physical annotators. In addition, it functions in increasing the accuracy of the annotation as it represents several layers of manual checking. The approach has been validated in a study by (Rödelsperger et al. 2019). The work clearly reflected that the novel approach of community/crowd participation is a viable approach for improved manual curation. It furthermore serves to actually grow the community along with the skills-set as a significant training aspect in order to promote uniformity is required.

16.3 Gene Features on the Prokaryotic Genome

Prokaryotes are simple, unicellular (although some form biofilms) microorganisms (bacteria and archaea) that do not possess a nucleus or any membrane bound organelles. A brief overview of prokaryotic genome structure and key features is outlined below. These structures and features are important for bioinformatics analysis and comparative genomics of prokaryote genomes, which are also discussed briefly below.

The prokaryotic genome is unevenly distributed within the cell, aggregating into a dense, viscous region known as the nucleoid (Teif and Bohinc 2011; Youssef et al. 2019). Prokaryotic genomes generally comprise of a single circular double-stranded DNA molecule varying in length but usually at least a few million bases (Mb) encoding several thousand genes. For example, *Escherichia coli* has a circular genome of 4.64 Mb that codes for approximately 4400 genes (Blattner et al. 1997). In contrast, *Mycoplasma genitalium* has a genome size of only 0.58 Mb and only 500 genes (Fraser et al. 1997). Some prokaryotes may have genomes with multiple circular molecules (e.g. *Vibrio cholerae* and *Deinococcus radiodurans* (Heidelberg et al. 2000; White et al. 1999)) or a combination of circular and linear and circular molecules (e.g. *Borrelia burgdorferi* (Fraser et al. 1997)). Prokaryotes may also have additional genetic material on small, circular (or linear) DNA molecules called plasmids, that are independent of the larger genome and have genes that confer additional properties (e.g. antibiotic resistance, ability to use other nutrients as carbon sources, etc.). Plasmids are not essential for prokaryotic survival but they are very beneficial and can be transferred from one prokaryote to another via horizontal gene transfer.

Replication of prokaryotic genomes occurs at a highly conserved sequence known as the *origin of replication* (*oriC*). Prokaryotes with circular chromosome have a single origin site while linear chromosomes generally have an origin site in the middle of the chromosome. Bacterial *oriC* sites can be very diverse in size (250–2000 bp), sequence composition and organisation depending on species (Mackiewicz et al. 2004). Replication in both circular and linear bacterial genomes generally occurs bidirectionally from the *oriC* site, but termination of replication in linear chromosomes requires telomeres (multiple tandem repeats of noncoding nucleotide sequences at the ends of the linear chromosome) to protect the ends of the chromosome from gradual degradation since the replication enzymes are unable to synthesise new DNA at the ends of the linear chromosome. The *oriC* site and associated sequences are very useful for annotation and assembly of prokaryotic genomes generated by Next Generation Sequencing (NGS). For example, DoriC (Luo and Gao 2019) was initially developed as a database of bacterial replication origins (*oriC*), which were determined either experimentally or as predicted by Ori-Finder (Raj et al. 2018a). Since the initial development and launch of DoriC in 2007, the database has undergone several upgrades and improvements. The current version, DoriC 10, has 7580, 226 and 1209 oriCs of bacteria, archaea and plasmids, respectively, and is the most complete and scalable database of prokaryotic replication origins (Luo and Gao 2019). This database facilitates enhanced understanding of the structure and functions of prokaryotic replication origins with many predictions verified experimentally in the laboratory (Luo and Gao 2019; Gao and Zhang 2008).

Generally, prokaryotic genes are clustered together as operons, i.e. genes that are required for a specific cellular or metabolic function are grouped together. Furthermore, aside from the coding sequences in an operon, there are additional sequences that are important for regulation expression of genes within an operon. These include (1) a promoter region that is located at the start ($5'$ end) of the operon and contains sequences for initiation of transcription by RNA polymerase, (2) the open reading frame (ORFs) which contain the polycistronic gene sequences, i.e. the coding sequences for the individual genes comprising the operon, (3) the terminator region located at the end ($3'$) of the operon which regulates termination of transcription. The first bacterial operon discovered was the *lac* operon, in *E. coli*, which contains all the genes required for utilising lactose as a carbon (energy) source (Jacob and Monod

1961). This ground breaking discovery earned Francois Jacob and Jacques Monod the Nobel Prize in Physiology or Medicine in 1965. Several bioinformatics tools have been developed to identify prokaryotic operons, promoters and other regulatory sequences. More recently, Operon-mapper (http://biocomputo.ibt.unam.mx/operon_mapper/) is the first publically available web server that predicts the operons of any bacterial or archaeal genome that only requires the genomic sequence as input data (Taboada et al. 2018). Another novel method for detecting and predicting operons called Operon Hunter uses visual representations of genomic fragments and a neural network architecture (i.e. machine learning) to achieve highly accurate predictions outperforming other state of the art tools (Assaf et al. 2020). Most of the current bioinformatics tools for operon prediction require genomic sequence data. However, the Rockhopper platform (https://cs.wellesley.edu/~btjaden/Rockhopper/) combines genome sequencing data with RNA-seq data (which is becoming more prevalent) to improve accuracy and specificity of computationally identified prokaryotic operons (Tjaden 2020). Another useful tool developed for visualisation and analysis of prokaryotic genomes, particularly identification of features such as analysis of transcription factors (TFs), regulatory motifs (promoters, ribosome binding sites, terminators), transcriptional regulation, etc. is BAC-Browser (Garanina et al. 2018). BAC-Browser (http://smdb.rcpcm.org/tools/index.html) also incorporates a variety of other free tools for primer design, visualisation and analysis.

Prokaryotic genomes consist of double-stranded DNA molecules encoding numerous genes which could be transcribed from either the sense ($5'$ to $3'$) or antisense ($3'$ to $5'$) strand. The sequence between the promoter and terminator that contains no stop codons, and which can be transcribed into mRNA and subsequently translated into protein is referred to as the *open reading frame* (ORF). Generally, ORFs begin with the start codon (ATG encoding methionine) and one of three possible stop codons (TAA, TAG, TGA). Since genes may encoded on either the sense or antisense strands and depending on the starting point chosen for a particular gene sequence, there are six possible reading frames for translating any DNA sequence into an amino acid sequence. Generally, lengthy ORFs are used, in conjunction with other empirical or predicted evidence, to identify possible protein coding sequences. There are many bioinformatics tools that have been developed to address these challenges and to accurately identify the correct ORFs. One of the most commonly used tools for prediction of prokaryotic ORFs is ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/). However, metagenome sequencing generates significant amounts of sequencing information, including from many prokaryotic species that cannot be cultured. Most computational tools for finding ORFs become computationally bottlenecked when used with metagenome data, particularly when this data consists of unassembled reads. To address these challenges, OrfM, a bioinformatics tool to identify ORFs in metagenomics sequence data has recently been developed (Woodcroft et al. 2016). This tool is four to five times faster than similar tools such as GetOrf (Tringe et al. 2005) and Translate (unpublished, http://eddylab.org/software.html) without any decrease in accuracy. This increase in performance significantly reduces the bottlenecks associated with ORF identification from large datasets which are typical of metagenome studies.

16.4 Gene Ontology and Community Annotation in Cancer Biology

The advent of cheaper second generation/next generation sequencing, along with the availability of single cell long read technologies, has significantly changed the field of cancer community cell heterogeneity. Along with that there has also been some development around digital cell identification technologies.

There are two processes that are currently typically used, from a physiological perspective, to evaluate variability in cancer cell communities. In the first instance researchers are looking at sectioning cancer tumours into different regions and then to conduct sequencing on the regional scale. This approach is termed multi-region sequencing.

The alternate approach, which allows for a finer scale analysis, involves utilising a combination of either single cell with long read sequencing technology or with transcriptome, sequencing. However, the error rates associated with long read technology at current (while it is improving) points to transcriptome sequencing leading the charge in terms of current applications. Ultimately, regardless of the approach, in the final analysis what is sought is to clearly define the differences in cell types and abundances of the cell types (e.g. Normal vs cancerous, and within cancerous to capture the differences between those cells). This knowledge-base in turn feeds into understanding the drivers of tumour evolution.

The value of understanding the diversity in cancer cells, as defined above, is that studies have now started revealing that there are correlations between the variations and overall patient survival in relation to liver cancer. The study by Ma et al. (2019) clearly demonstrates a correlation between transcriptomic diversity, which links to genomic diversity, and predicts patient prognosis. This is achieved through a unique bioinformatics workflow which involved the utility of Seurat package (version 2.3.0) (Butler et al. 2018) in R (version 3.4.3). The computational tool takes in single cell RNA seq data, and allows for the identification of shared populations across varying datasets. This provides input for modelling the overlaps and similarities versus the differences of cancer cell signatures within the overall population. Outside of the scope of sequence based approaches there have also been significant developments in the implementation of digital cell identification technologies, which have found a home in a discipline termed computational pathology. This is concerned with the development of algorithms that facilitate the evaluation and analysis of digital pathology images. The implementation of such an approach uses deep neural network training models, which forms part of a semi-supervised approach for successfully distinguishing different community cells (Javed et al. 2020). Ultimately it translates into the construction of a library of digital imagery of cancerous cells which feeds into a model in order to identify the cancerous cells in a novel setting, or when scientists are exposed to classifying a novel dataset. In terms of advancing the area of cancer research, both in the context of community variation and generally understanding the disease, there is an obvious need for a well-represented reference databases with a shared vocabulary which all researchers in the cancer/general biology space recognise. One of the older yet, still most relevant resources relating to ontologies remain the Gene Ontology database, which have been continuously updated to keep trend with the latest development in the field of ontologies (Ashburner 2000; Gene Ontology Consortium 2019).

An obvious development within the cancer research community has subsequently been the development of ontologies with greater specificity in the field of cancer. These represent a diverse array of areas, which is too broad to entirely define within the scope of this work. As an example, there has been the development of ontologies that speak to aspects such as hematologic malignancies (Serra et al. 2019), where the basis of their defining criteria involves using the classification of immunophenotypes as the basis for differentiation.

16.5 Application of Microbial Genomics in Cancer Biology

Using technologies such as 16S rRNA gene sequencing and high throughput array panels, the elucidation of microbial genomics has been applied primarily in cancer diagnosis, prognosis and in deciphering mechanisms of disease progression. Elevated salivary microbiota (*Prevotella melaninogenica*, *Streptococcus mitis* and *Capnocytophaga gingivalis*) were found to be a potential diagnostic indicators of oral cancer (Mager et al. 2005). Another study using 16S rRNA gene sequencing showed that oral tumours had an increased abundance of *Veillonella*, *Dialister*, and *Streptococcus* species (Guerrero-Preston et al. 2016). Börnigen et al. compared oral bacteria of 121 oral cancer patients to 242 controls and found significant changes in the microbial abundance and diversity between the two groups (Börnigen et al. 2017).

In lung cancer research, there has been several studies conducted to identify microbial communities present in lung cancer patients utilising varying sample types (Yan et al. 2015; Yang et al. 2018; Xu et al. 2020). Yan et al. showed through deep sequencing of 20 lung cancer patient samples compared to ten healthy controls, the significant abundance of *Neisseria*, *Veillonella* and *Capnocytophaga* in the tumour samples (Yan et al. 2015). Consequently they proposed that these bacteria could be potential diagnostic markers for lung cancer. In another study, 16S rRNA gene sequencing was used to determine microbial abundance and diversity in 75 lung cancer patients. They found that compared to 127 healthy individuals, this group of non-smoking female cancer patients had increased levels of *Blastomonas* and *Sphingomonas*. Furthermore, they showed that the expression of Napsin A, a well-known immunohistochemical marker of lung adenocarcinoma, was positively correlated to *Blastomonas* occurrence in the samples (Yang et al. 2018). Metagenomic sequencing was also used to identify plausible bacterial biomarkers for lung cancer in a study conducted by Cameron et al. The authors evaluated sputum samples from ten suspected lung cancer patients and determined that *Streptococcus viridans* was significantly present in the lung cancer positive patients compared to those that were negative (Cameron et al. 2017). Bronchoalveolar fluid samples from 20 lung cancer patients and 8 benign diseased patients were analysed using 16S rRNA sequencing. It was found that *Veillonella* and *Megasphaera* were significantly abundant in lung cancer patients (Lee et al. 2016).

The role of the microbiome in the study of colorectal cancer pathogenesis is evident (Saus et al. 2019; Song et al. 2020; Cho et al. 2014; Lin et al. 2019). 16S rRNA gene sequencing and real-time polymerase chain reaction were applied to faecal and mucosal samples obtained from colorectal cancer (CRC) patients. It was observed that the abundance and diversity of microbiota, and the expression of inflammation-associated genes, were distinct between CRC patients and healthy individuals (Flemer et al. 2017). Another study utilised 16S rRNA gene sequencing and gas-chromatography mass spectrometry to evaluate the microbiome and metabolome, respectively, of faecal samples obtained from 50 CRC patients versus 50 healthy individuals. It was observed that 76 operational taxonomic units differentiated the two groups (Yang et al. 2019a). Yu et al. conducted a metagenome-wide association utilising faecal samples obtained from CRC Chinese patients and 50 control samples and determined that *Peptostreptococcus stromatis* and *Fusobacterium nucleatum* were associated with CRC (Yu et al. 2017). The alteration of intestinal microbiota was shown in 15 CRC patients, specifically *Fusobacterium*, *Selemonas* and *Peptostreptococcus* increased in abundance and diversity (Hibberd et al. 2017). The analysis of oral microbiome was conducted using 16S rRNA gene sequencing in a prospective cohort which included participants from African-American and low-income groups. They observed that *Prevotella intermedia* and *Treponema denticola* were associated with increased CRC risk. Additionally, Bifidobacteriaceae was more abundant in CRC patients compared to controls (Yang et al. 2019b).

In breast cancer patients, microbial dysbiosis has also been observed. One study found that *Methylobacterium radiotolerans* was increased in tumours compared to normal tissues (Xuan et al. 2014). Utilising a pan-pathogen array, distinct microbial signature has been observed in triple negative breast cancer patients (Banerjee et al. 2015). The presence of distinct microbial signatures in breast tumours was also confirmed by Hieken et al. (2016). Similar trends can also be observed in other cancers such as pancreatic, oesophageal, gastric, head and neck cancers. A study involving 361 pancreatic cancer (PC) and 371 controls was analysed and was shown that while *Aggregatibacter actinomycetemcomitans* and *Porphyromonas gingivalis* were associated with increased risk of PC, *Fusobacteria* was linked to decreased PC risk (Fan et al. 2018). The widely studied *Helicobacter pylori* has been associated with gastrointestinal carcinogenesis (Inamura 2020; Meng et al. 2015; Trikudanathan et al. 2011). In gastric cancer, it promotes carcinogenesis by delivering CagA protein in epithelial cells (Hatakeyama 2004). Coker et al. applied 16S rRNA gene sequencing on 81 gastric mucosal samples and found that as the diseased progressed there was an abundance of *Streptococcus anginosus, Peptostreptococcus stomatis, Parvimonas micra, Dialister pneumosintes* and *Slackia exigua* hinting on their potential roles in gastric cancer progression (Coker et al. 2018).

The oral microbiome was assessed in a prospective cohort of oesophageal adenocarcinoma patients and *Tannerella forsythia* was found to be associated with increased risk of the cancer (Peters et al. 2017). In the same study, it was also observed that abundant *Porphyromonas gingivalis* was linked to increased oesophageal squamous cell carcinoma. One study showed that in head and neck squamous cell (HNSC) cancer patients, increased abundance of *Corynebacterium* and *Kingella* was linked to reduced HNSC cancer risk (Hayes et al. 2018). Likewise, in a cohort of 169 HNSC cancer patients, Wang et al. observed through 16S rRNA gene sequencing that Parvimonas was elevated in tumours compared to normal tissues (Wang et al. 2017).

Microbial genomics has also been critical in understanding the initiation and progression of cancer by modulating several cellular and biological processes such as immune response, inflammation and cell proliferation and death mechanisms (Xu et al. 2020; Inamura 2020; Chattopadhyay et al. 2019). Using in vivo models, Jin et al. showed that in lung cancer, inflammation can be caused by the microbiota through activation of γδ T cells resident in the lungs (Jin et al. 2019). In oral squamous cell carcinoma (OSCC), *Porphyromonas gingivalis* induced the upregulation of pro-inflammatory molecules such as IL1, 6, 8 and metalloproteinases 1, 9, 10 (Chattopadhyay et al. 2019). *Porphyromonas gingivalis* also enhanced Epithelial-to-Mesenchymal Transition pathway crucial in cellular migration and metastasis (Chattopadhyay et al. 2019). The bacteria was also found to inactivate Bad, a pro-apoptotic protein, through Akt (Yao et al. 2010). DNA sequencing of 20 OSCC tumours compared to 20 controls showed that *Fusobacterium nucleatum* and *Pseudomonas aeruginosa* was abundant in tumours (Al-Hebshi et al. 2017). The study also predicted that their abundance was linked to genes responsible for processes such as lipopolysaccharide synthesis which plays a key role in inflammatory response.

Over the past years the study of microbial communities present in tumours has helped better understand and predict tumourigenesis. Undoubtedly, the application of microbial genomics would provide future novel ways to aid in cancer diagnosis, management and treatment.

16.6 Application of Microbes for Cancer Treatment and Cancer Precision Medicine

Recently synthetic biology has enabled cancer precision medicine, particularly with respect to recent development with respect to application of synthetic microbes as an enabling tool and technologies for cancer therapy (Courbet et al. 2016). Despite the recent development with respect to applications of synthetic biology for cancer therapy, bacterial systems have applied in cancer treatment for more than century (Felgner et al. 2016). In this section of the book chapter, we will focus on historical relevance and current developments on applying synthetic microbes for cancer precision medicine. In this section, we focus on microbial applications for cancer treatment and how synthetic microbes enabled for cancer precision medicine and therapy.

For a long period in the past, controversial approach was to use live bacteria for prophylactic vaccination and cancer therapy (Payette and Davis 2001). In the nineteenth century, Coley WB reported spontaneous tumour regression in patients with streptococcal infections. Further, Coley generated a variety of “anti-tumour vaccines” by combining heat-killed *S. pyogenes* with heat-killed *S. marcescens*. These vaccines are known as Coley’s toxins that were administered to cancer patients. Coley’s toxins were one of the first cancer immunotherapy and Coley’s contributions were one of the first applications for bacterial-based therapeutics in cancer treatment. In 1970s, Bacillus Calmette–Guerin (BCG) is the only FDA approved bacterial agent that is employed for the treatment of non-muscle invasive bladder cancer (NMIBC). In the last decade, there were a number of live attenuated bacteria applied for destroying the cancer cells in vitro, in rodents to destroy tumours and inhibit tumour growth in the organism (Hoffman 2012). Several research groups have published genetically engineered stains of E.coli and Salmonella bacteria that showed effective targeting of tumour and delivery of drugs both in vitro and in animal model (Min et al. 2008). Similarly, bacteria such as *Clostridium novyi* and *Listeria monocytogenes* were applied for cancer treatment (Roberts et al. 2014; Wood et al. 2008). In addition, microbes such as *Streptococcus pyogenes* OK-432 were used in cancer treatment based on sclerotherapy where *S. pyogenes* OK-432 were injected for lymphangiomas treatment. Several anaerobic microbes have shown potential to be applied in anticancer treatment due to its ability to grow under hypoxia conditions. *Magnetococcus marinus* MC1 has unique bacteria structure based on the presence of magnetosomes and has negative aerotaxis capabilities that would enable as a tool for destroying cancer cells. In addition, using MRI, there is high possibility to redirect bacteria containing magnetosomes to the target site in tumour system (Loshitskiy and Nikolov 2015). Another interesting microbes that are applied in cancer treatment is *Toxoplasma gondii* which is an obligatory intracellular protozoan. It has been reported that Toxoplasma lysate antigen (TLA) contains microbe’s antigen that can be applied for neurodegenerative disease and cancer. Particularly, Toxoplasma gondii carbamoyl phosphate synthase mutant is being applied in several aggressive cancer (melanoma, pancreatic, ovarian and lung) (Bzik et al. 2013). There are reports on applying Plasmodium falciparum in cancer treatment due to connections with malaria (Nordor et al. 2018).

Gut microbiome for cancer treatment: Applications of gut microbiome for cancer treatment are recently focused by the cancer researcher with great interest. There has been tremendous interest to explore the application and pivotal role of microbiome in cancer immunotherapy treatment. Some of the applications of gut microbiome for cancer treatment include drug metabolism as the gut microbiome influences mode of actions of drugs, its efficacy and antibody therapy. Several research teams are currently studying the gut microbiota and its impact on effective cancer treatments such as Chemo, Radio and Immunotherapy (Inamura 2020; Vivarelli et al. 2019). One of the examples of gut microbiome applications includes development of probiotics for cancer treatment. Besides gut microbiome, skin, nasal and vaginal microbiome also play a critical role in influencing cancer treatment at the associated human organs. Skin microbiome has been reported to impact skin cancer treatment. Similarly nasal/lung microbiota has been reported to be linked to lung cancer and vaginal microbiome has been reported to ovarian cancer. Several research groups are studying the skin, nasal and vaginal microbiome in healthy and cancer patient cohort in order to better understand the relationship between microbiome and cancer (Hieken et al. 2016).

Furthermore, with better understanding of gut, skin, nasal and vaginal microbiome, the scientists are proposing tailor-made cancer treatment for individual patients. To elaborate in details, by using microbiome information, there is potential approach to apply for cancer precision medicine in clinical setting. With the recent advancement of next generation sequencing and synthetic biology technologies, microbiome dataset has been integrated with cancer precision medicine in order to provide clinically relevant drug treatments for various patient cohorts. With the better understanding on the microbiome, there might be potential approach to understand and address problem of adverse drug reaction in African patient cohort.

16.7 Conclusion

The above sections provide a brief summary of prokaryotic genome architecture and more recently developed bioinformatics tools which are very useful for studying, identifying and annotating various features of prokaryotic genomes. With the recent developments of synthetic biology tools and technologies, several designer microbes have been invented for various applications. As a future perspective, we should focus on synthetic biology driven designer microbes enabled improved cancer treatment and its application in cancer precision medicine.

References

Al-Hebshi NN, Nasher AT, Maryoud MY, Homeida HE, Chen T, Idris AM et al (2017) Inflammatory bacteriome featuring *Fusobacterium nucleatum* and *Pseudomonas aeruginosa* identified in association with oral squamous cell carcinoma. Sci Rep 7:1834. https://doi.org/10.1038/s41598-017-02079-3

Ashburner M (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29. https://doi.org/10.1038/75556

Assaf R, Xia F, Stevens R (2020) Detecting operons in bacterial genomes via visual representation learning. bioRxiv 2020:860221. https://doi.org/10.1101/860221

Banerjee S, Wei Z, Tan F, Peck KN, Shih N, Feldman M et al (2015) Distinct microbiological signatures associated with triple negative breast cancer. Sci Rep 5:15162. https://doi.org/10.1038/srep15162

Berger I, Raj DBTG (2014) Improved baculoviral expression system and methods of producing the same. CA2898696A1. https://patents.google.com/patent/CA2898696A1/tr

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M et al (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462. https://doi.org/10.1126/science.277.5331.1453

Börnigen D, Ren B, Pickard R, Li J, Ozer E, Hartmann EM et al (2017) Alterations in oral bacterial communities are associated with risk factors for oral and oropharyngeal cancer. Sci Rep 7:17686. https://doi.org/10.1038/s41598-017-17795-z

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36:411–420. https://doi.org/10.1038/nbt.4096

Bzik DJ, Fox B, Fiering SN, Conejo-Garcia JR, Baird J, Sanders KL (2013) Method for treating cancer with *Toxoplasma gondii* vaccine. WO2013059411A1. https://patents.google.com/patent/WO2013059411A1/en

Cameron SJS, Lewis KE, Huws SA, Hegarty MJ, Lewis PD, Pachebat JA et al (2017) A pilot study using metagenomic sequencing of the sputum microbiome suggests potential bacterial biomarkers for lung cancer. PLoS One 12:e0177062. https://doi.org/10.1371/journal.pone.0177062

Chattopadhyay I, Verma M, Panda M (2019) Role of oral microbiome signatures in diagnosis and prognosis of oral cancer. Technol Cancer Res Treat 2019:18. https://doi.org/10.1177/1533033819867354

Cho M, Carter J, Harari S, Pei Z (2014) The interrelationships of the gut microbiome and inflammation in colorectal carcinogenesis. Clin Lab Med 34:699–710. https://doi.org/10.1016/j.cll.2014.08.002

Coker OO, Dai Z, Nie Y, Zhao G, Cao L, Nakatsu G et al (2018) Mucosal microbiome dysbiosis in gastric carcinogenesis. Gut 67:1024–1032. https://doi.org/10.1136/gutjnl-2017-314281

Courbet A, Renard E, Molina F (2016) Bringing next-generation diagnostics to the clinic through synthetic biology. EMBO Mol Med 8:987–991. https://doi.org/10.15252/emmm.201606541

Danchin A, Ouzounis C, Tokuyasu T, Zucker J (2018) No wisdom in the crowd: genome annotation in the era of big data—current status and future prospects. Microb Biotechnol 11:588–605. https://doi.org/10.1111/1751-7915.13284

Fan X, Alekseyenko AV, Wu J, Peters BA, Jacobs EJ, Gapstur SM et al (2018) Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. Gut 67:120–127. https://doi.org/10.1136/gutjnl-2016-312580

FEBS-EMBO (2020) Concurrent session (lecture): CS III-6-3 SynBac: designer minimal baculovirus genome for drug discovery. In: ResearchGate [Internet]. https://www.researchgate.net/publication/265140755_FEBS-EMBO_Concurrent_Session_lecture_CS_III-6-3_SynBac_designer_minimal_baculovirus_genome_for_drug_discovery

Felgner S, Kocijancic D, Frahm M, Weiss S (2016) Bacteria in cancer therapy: renaissance of an old concept. Int J Microbiol 2016:e8451728. https://doi.org/10.1155/2016/8451728

Flemer B, Lynch DB, Brown JMR, Jeffery IB, Ryan FJ, Claesson MJ et al (2017) Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut 66:633–643. https://doi.org/10.1136/gutjnl-2015-309595

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R et al (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390:580–586. https://doi.org/10.1038/37551

Gao F, Zhang C-T (2008) Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. BMC Bioinfo 9:79. https://doi.org/10.1186/1471-2105-9-79

Garanina IA, Fisunov GY, Govorun VM (2018) BAC-BROWSER: the tool for visualization and analysis of prokaryotic genomes. Front Microbiol 9:2827. https://doi.org/10.3389/fmicb.2018.02827

Gene Ontology Consortium (2019) The gene Ontology resource: 20 years and still going strong. Nucleic Acids Res 47:D330–D338. https://doi.org/10.1093/nar/gky1055

Guerrero-Preston R, Godoy-Vitorino F, Jedlicka A, Rodríguez-Hilario A, González H, Bondy J et al (2016) 16S rRNA amplicon sequencing identifies microbiota associated with oral cancer, human papilloma virus infection and surgical treatment. Oncotarget 7:51320–51334. https://doi.org/10.18632/oncotarget.9710

Hatakeyama M (2004) Oncogenic mechanisms of the *Helicobacter pylori* CagA protein. Nat Rev Cancer 4:688–694. https://doi.org/10.1038/nrc1433

Hayes RB, Ahn J, Fan X, Peters BA, Ma Y, Yang L et al (2018) Association of oral microbiome with risk for incident head and neck squamous cell cancer. JAMA Oncol 4:358–365. https://doi.org/10.1001/jamaoncol.2017.4777

Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ et al (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477–483. https://doi.org/10.1038/35020000

Hibberd AA, Lyra A, Ouwehand AC, Rolny P, Lindegren H, Cedgård L et al (2017) Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention. BMJ Open Gastroenterol 4:e000145. https://doi.org/10.1136/bmjgast-2017-000145

Hieken TJ, Chen J, Hoskin TL, Walther-Antonio M, Johnson S, Ramaker S et al (2016) The microbiome of aseptically collected human breast tissue in benign and malignant disease. Sci Rep 6:30751. https://doi.org/10.1038/srep30751

Hoffman RM (2012) The preclinical discovery of bacterial therapy for the treatment of metastatic cancer with unique advantages. Expert Opin Drug Discovery 7:73–83. https://doi.org/10.1517/17460441.2012.644534

Hood L, Rowen L (2013) The human genome project: big science transforms biology and medicine. Genome Med 5:79. https://doi.org/10.1186/gm483

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinfo 11:119. https://doi.org/10.1186/1471-2105-11-119

Inamura K (2020) Gut microbiota contributes towards immunomodulation against cancer: new frontiers in precision cancer therapeutics. Sem Cancer Biol 5:29. https://doi.org/10.1016/j.semcancer.2020.06.006

Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356. https://doi.org/10.1016/S0022-2836(61)80072-7

Javed S, Mahmood A, Fraz MM, Koohbanani NA, Benes K, Tsang Y-W et al (2020) Cellular community detection for tissue phenotyping in colorectal cancer histology images. Med Image Anal 63:101696. https://doi.org/10.1016/j.media.2020.101696

Jin C, Lagoudas GK, Zhao C, Bullman S, Bhutkar A, Hu B et al (2019) Commensal microbiota promote lung cancer development via γδ T cells. Cell 176:998–1013. https://doi.org/10.1016/j.cell.2018.12.040

Lee SH, Sung JY, Yong D, Chun J, Kim SY, Song JH et al (2016) Characterization of microbiome in bronchoalveolar lavage fluid of patients with lung cancer comparing with benign mass like lesions. Lung Cancer 102:89–95. https://doi.org/10.1016/j.lungcan.2016.10.016

Lin C, Cai X, Zhang J, Wang W, Sheng Q, Hua H et al (2019) Role of gut microbiota in the development and treatment of colorectal Cancer. Digestion 100:72–78. https://doi.org/10.1159/000494052

Lomsadze A, Gemayel K, Tang S, Borodovsky M (2018) Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Res 28:1079–1089. https://doi.org/10.1101/gr.230615.117

Loshitskiy PP, Nikolov NA (2015) Magnetothermia utilization in the curing of malignancies. Part 1. Radioelectron Commun Syst 58:49–60. https://doi.org/10.3103/S0735272715020016

Luo H, Gao F (2019) DoriC 10.0: an updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. Nucleic Acids Res 47:D74–D77. https://doi.org/10.1093/nar/gky1014

Ma L, Hernandez MO, Zhao Y, Mehta M, Tran B, Kelly M et al (2019) Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer. Cancer Cell 36:418–430. https://doi.org/10.1016/j.ccell.2019.08.007

Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek MR, Cebrat S (2004) Where does bacterial replication start? Rules for predicting the oriC region. Nucleic Acids Res 32:3781–3791. https://doi.org/10.1093/nar/gkh699

Mager DL, Haffajee AD, Devlin PM, Norris CM, Posner MR, Goodson JM (2005) The salivary microbiota as a diagnostic indicator of oral cancer: a descriptive, non-randomized study of cancer-free and oral squamous cell carcinoma subjects. J Transl Med 3:27. https://doi.org/10.1186/1479-5876-3-27

Meng W, Bai B, Sheng L, Li Y, Yue P, Li X et al (2015) Role of helicobacter pylori in gastric cancer: advances and controversies. Discov Med 20:285–293

Min J-J, Kim H-J, Park JH, Moon S, Jeong JH, Hong Y-J et al (2008) Noninvasive real-time imaging of tumors and metastases using tumor-targeting light-emitting *Escherichia coli*. Mol Imaging Biol 10:54–61. https://doi.org/10.1007/s11307-007-0120-5

Mulder NJ, Christoffels A, de Oliveira T, Gamieldien J, Hazelhurst S, Joubert F et al (2016) The development of computational biology in South Africa: successes achieved and lessons learnt. PLoS Comput Biol 12:e1004395. https://doi.org/10.1371/journal.pcbi.1004395

Nordor AV, Bellet D, Siwo GH (2018) Cancer–malaria: hidden connections. Open Biol 8:180127. https://doi.org/10.1098/rsob.180127

Özdemir V, Dove ES, Gürsoy UK, Şardaş S, Yıldırım A, Yılmaz ŞG et al (2017) Personalized medicine beyond genomics: alternative futures in big data-proteomics, environtome and the social proteome. J Neural Transm (Vienna) 124:25–32. https://doi.org/10.1007/s00702-015-1489-y

Payette PJ, Davis HL (2001) History of vaccines and positioning of current trends. Curr Drug Targets Infect Disord 1:241–247. https://doi.org/10.2174/1568005014606017

Peters BA, Wu J, Pei Z, Yang L, Purdue MP, Freedman ND et al (2017) Oral microbiome composition reflects prospective risk for esophageal cancers. Cancer Res 77:6777–6787. https://doi.org/10.1158/0008-5472.CAN-17-1296

Raj TG, Balaji DB, Giliberto M, Cremaschi A, Skånland SS, Gade A, Tjønnfjord GE et al (2018a) Drug sensitivity screening on multiple myeloma for precision cancer therapy. Blood 132:4677–4677. https://doi.org/10.1182/blood-2018-99-110669

Raj TG, Balaji DB, Cremaschi A, Skånland SS, Gade A, Schjesvold FH, Tjønnfjord GE et al (2018b) In-vitro drug sensitivity screening in chronic lymphocytic leukemia (CLL) primary patient samples identifies drug candidates for precision cancer therapy. Blood 132:4676–4676. https://doi.org/10.1182/blood-2018-99-110357

Roberts NJ, Zhang L, Janku F, Collins A, Bai R-Y, Staedtke V et al (2014) Intratumoral injection of *Clostridium novyi*-NT spores induces antitumor responses. Sci Transl Med 6:249. https://doi.org/10.1126/scitranslmed.3008982

Rödelsperger C, Athanasouli M, Lenuzzi M, Theska T, Sun S, Dardiry M et al (2019) Crowdsourcing and the feasibility of manual gene annotation: a pilot study in the nematode Pristionchus pacificus. Sci Rep 9:18789. https://doi.org/10.1038/s41598-019-55359-5

Saus E, Iraola-Guzmán S, Willis JR, Brunet-Vega A, Gabaldón T (2019) Microbiome and colorectal cancer: roles in carcinogenesis and clinical potential. Mol Asp Med 69:93–106. https://doi.org/10.1016/j.mam.2019.05.001

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Serra LM, Duncan WD, Diehl AD (2019) An ontology for representing hematologic malignancies: the cancer cell ontology. BMC Bioinfo 20:181. https://doi.org/10.1186/s12859-019-2722-8

Song M, Chan AT, Sun J (2020) Influence of the gut microbiome, diet, and environment on risk of colorectal cancer. Gastroenterology 158:322–340. https://doi.org/10.1053/j.gastro.2019.06.048

SynBac™ (2020) In: Geneva Biotech [cited 2 Nov 2020]. https://geneva-biotech.com/product_category/insect-cell-expression/synbac/

Taboada B, Estrada K, Ciria R, Merino E (2018) Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. Bioinformatics 34:4118–4120. https://doi.org/10.1093/bioinformatics/bty496

Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L et al (2016) NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569

Teif VB, Bohinc K (2011) Condensed DNA: condensing the concepts. Prog Biophys Mol Biol 105:208–222. https://doi.org/10.1016/j.pbiomolbio.2010.07.002

Tjaden B (2020) A computational system for identifying operons based on RNA-seq data. Methods 176:62–70. https://doi.org/10.1016/j.ymeth.2019.03.026

Trikudanathan G, Philip A, Dasanu CA, Baker WL (2011) Association between Helicobacter pylori infection and pancreatic cancer. A cumulative meta-analysis. JOP 12:26–31

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW et al (2005) Comparative metagenomics of microbial communities. Science 308:554–557. https://doi.org/10.1126/science.1107851

Vijayachandran LS, Thimiri Govinda Raj DB, Edelweiss E, Gupta K, Maier J, Gordeliy V et al (2013) Gene gymnastics: synthetic biology for baculovirus expression vector system engineering. Bioengineered 4:279–287

Vivarelli S, Salemi R, Candido S, Falzone L, Santagati M, Stefani S et al (2019) Gut microbiota and cancer: from pathogenesis to therapy. Cancer 11:38. https://doi.org/10.3390/cancers11010038

Wang H, Funchain P, Bebek G, Altemus J, Zhang H, Niazi F et al (2017) Microbiomic differences in tumor and paired-normal tissue in head and neck squamous cell carcinomas. Genome Med 9:14. https://doi.org/10.1186/s13073-017-0405-5

White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ et al (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286:1571–1577. https://doi.org/10.1126/science.286.5444.1571

Wintjens R, Bifani AM, Bifani P (2020) Impact of glycan cloud on the B-cell epitope prediction of SARS-CoV-2 spike protein. Vaccine 5:1–8. https://doi.org/10.1038/s41541-020-00237-9

Wood LM, Guirnalda PD, Seavey MM, Paterson Y (2008) Cancer immunotherapy using listeria monocytogenes and listerial virulence factors. Immunol Res 42:233–245. https://doi.org/10.1007/s12026-008-8087-0

Woodcroft BJ, Boyd JA, Tyson GW (2016) OrfM: a fast open reading frame predictor for metagenomic data. Bioinformatics 32:2702–2703. https://doi.org/10.1093/bioinformatics/btw241

Xu N, Wang L, Li C, Ding C, Li C, Fan W et al (2020) Microbiota dysbiosis in lung cancer: evidence of association and potential mechanisms. Transl Lung Cancer Res 9:1554–1568. https://doi.org/10.21037/tlcr-20-156

Xuan C, Shamonki JM, Chung A, Dinome ML, Chung M, Sieling PA et al (2014) Microbial dysbiosis is associated with human breast cancer. PLoS One 9:e83744. https://doi.org/10.1371/journal.pone.0083744

Yan X, Yang M, Liu J, Gao R, Hu J, Li J et al (2015) Discovery and validation of potential bacterial biomarkers for lung cancer. Am J Cancer Res 5:3111–3122

Yang J, Mu X, Wang Y, Zhu D, Zhang J, Liang C et al (2018) Dysbiosis of the salivary microbiome is associated with non-smoking female lung cancer and correlated with immunocytochemistry markers. Front Oncol 8:520. https://doi.org/10.3389/fonc.2018.00520

Yang Y, Misra BB, Liang L, Bi D, Weng W, Wu W et al (2019a) Integrated microbiome and metabolome analysis reveals a novel interplay between commensal bacteria and metabolites in colorectal cancer. Theranostics 9:4101–4114. https://doi.org/10.7150/thno.35186

Yang Y, Cai Q, Shu X-O, Steinwandel MD, Blot WJ, Zheng W et al (2019b) Prospective study of oral microbiome and colorectal cancer risk in low-income and African American populations. Int J Cancer 144:2381–2389. https://doi.org/10.1002/ijc.31941

Yao L, Jermanus C, Barbetta B, Choi C, Verbeke P, Ojcius DM et al (2010) Porphyromonas gingivalis infection sequesters pro-apoptotic bad through Akt in primary gingival epithelial cells. Mol Oral Microbiol 25:89–101. https://doi.org/10.1111/j.2041-1014.2010.00569.x

Youssef N, Budd A, Bielawski JP (2019) Introduction to genome biology and diversity. Methods Mol Biol 1910:3–31. https://doi.org/10.1007/978-1-4939-9074-0_1

Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y et al (2017) Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 66:70–78. https://doi.org/10.1136/gutjnl-2015-309800

Bioinformatics for Human Microbiome

17

Abhirupa Ghosh, Shazia Firdous, and Sudipto Saha

Abstract

The dynamics of human metabolism and physiology is governed by the complex microbial communities present in different body sites. Advances in sequencing technologies and computational methods have boosted the microbiome analysis towards better resolution. Presently, microbiome research field has bloomed with generation of massive datasets and development of huge number of analysis tools. However, the complexity of the workflows and diversity of the tools in the repertoires make the field difficult. In this chapter we systematically discuss the metataxonomics, metagenomics and metatranscriptomics approaches, pipelines and the recommended tools. Further, the state-of-the-art downstream analysis techniques and visualisation tools were discussed. This chapter will help the researchers in computational analysis considering their biological questions related to human microbiome.

Keywords

Metabolism · Microbial community · Metagenomics · Downstream analysis · Microbiome · Computational analysis

17.1 Introduction

Development of the sequencing technology towards massively parallel high-throughput sequencing (HTS) in the last decade has prompted big projects such as Metagenomics of the Human Intestinal tract (MetaHIT), Human Microbiome Project (HMP), Belgian Flemish Gut Flora Project, Dutch LifeLines-DEEP study,

A. Ghosh · S. Firdous · S. Saha (✉)
Division of Bioinformatics, Bose Institute, Kolkata, West Bengal, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_17

integrative HMP (iHMP) and Chinese Academy of Sciences Initiative of Microbiome (CAS-CMI) (Qin et al. 2010; Human Microbiome Project 2012a, b; The Integrative HMP (iHMP) Research Network Consortium 2019; Tigchelaar et al. 2015; Valles-Colomer et al. 2019; Shi et al. 2019). These projects led to large amounts of data generation, increase in the number of reference genomes, computational resources and analysis pipelines. Moreover, large population based microbiome studies have encouraged a plethora of projects exploring the 'normal' microbiota of various human body sites; the perturbation of microbiota during various diseases, over time period or with age; the comparison of microbiota from different geographical populations, different diet groups and various medication exposed groups (Costello et al. 2009; Grice et al. 2009; Caporaso et al. 2011; Rajilic-Stojanovic et al. 2012; Mehta et al. 2018; Sommer et al. 2017; Tamburini et al. 2016; Robertson et al. 2019; Karlsson et al. 2013; Zimmermann et al. 2019; Yatsunenko et al. 2012; Pasolli et al. 2019; Bay et al. 2020; Sun et al. 2020a, b; Stennett et al. 2020; Huey et al. 2020; Susic et al. 2020; Nishijima et al. 2016; Das et al. 2018; Nayfach et al. 2019). Recent studies also explore the human microbiome to address the antibiotic resistance (Relman and Lipsitch 2018). Altogether, these studies have enhanced the understanding of human microbiome diversity and the functional roles of human microbiome in health and diseases.

The growth in the sequencing technology has been accompanied by the development of computational tools and resources dedicated to microbiome analysis. In recent times, a large number of tools and software are freely available for each analysis steps from quality control to visualisation. However, the rapidly evolving computational techniques and standards are modifying the analysis pipelines. The High-Throughput Sequencing (HTS) based human microbiome studies can be broadly categorised into three different sequencing methods as shown in Fig. 17.1. The metataxonomic analysis package, QIIME (Quantitative Insights Into Microbial Ecology), a breakthrough in the field of bioinformatics tools is recently upgraded to QIIME 2 with all new features for better statistical analysis and visualisation (Caporaso et al. 2010; Bolyen et al. 2019). QIIME 2 has also incorporated the concept of amplicon sequence variants (ASVs) over the traditional operational taxonomic units (OTUs). Likewise, metagenomic analysis has evolved from direct use of BLAST for mapping reads to more computationally intensive techniques such as mapping k-mers and use of de Bruijn graphs (Altschul et al. 1990; Compeau et al. 2011; Namiki et al. 2012; Li et al. 2015; Ounit et al. 2015; Qiao et al. 2018; Wood and Salzberg 2014). While metataxonomics and metagenomics are widely used to identify the microbial community and their functional profile, microbiome research is forwarding towards metatranscriptomics to detect the active microbiome features. Similarly, there are promising improvements in the taxonomic and functional analysis with the integration of machine learning techniques and multiomics analyses (Morton et al. 2017; Subramanian et al. 2014; Knights et al. 2011; Oh and Zhang 2020; Vangay et al. 2019; Qian et al. 2020). Galaxy is a web-based open-source popular platform with large repertoire of tools and option for custom workflow and has emerged as valuable data intensive analysis platform (Afgan et al. 2018; Batut et al. 2018; Thang et al. 2019). The major difficulty of microbiome data analysis is

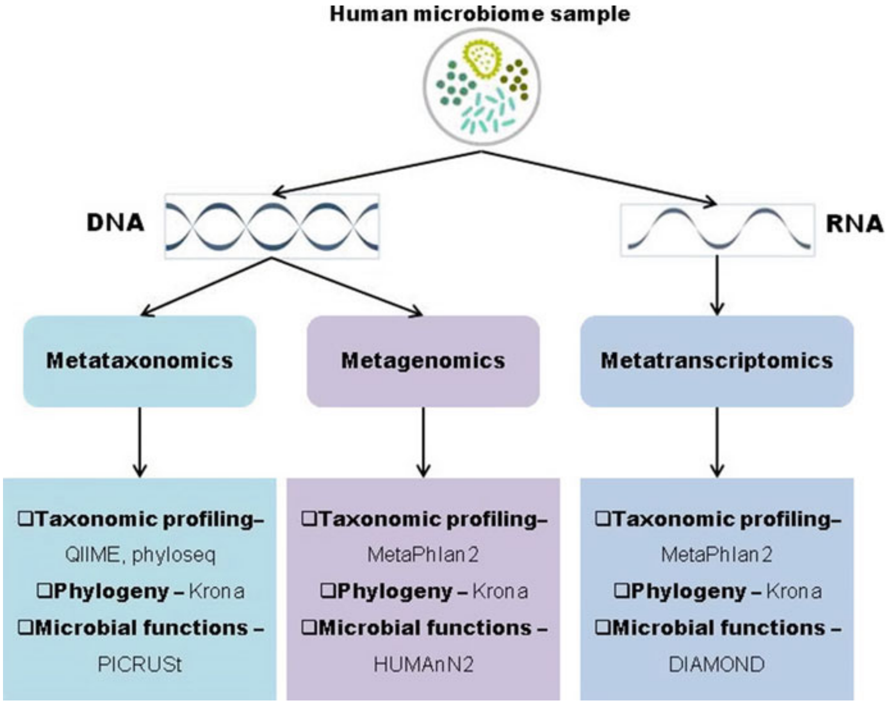

Human microbiome sample
DNA
RNA
Metataxonomics
Metagenomics
Metatranscriptomics
❑Taxonomic profiling– QIIME, phyloseq
❑Phylogeny – Krona
❑Microbial functions – PICRUSt
❑Taxonomic profiling– MetaPhlan2
❑Phylogeny – Krona
❑Microbial functions – HUMAnN2
❑Taxonomic profiling– MetaPhlan2
❑Phylogeny – Krona
❑Microbial functions – DIAMOND

Fig. 17.1 High-throughput sequencing and computational analysis allow to explore the human microbiome and to understand its role in human normal physiology and diseases

choosing the right tool from the array of computational tools and methods while maintaining the standards of the study. The design and environmental factors of experimental methods, the analysis workflow can affect the final outcome.

In this chapter, the bioinformatics aspects of microbiome data analyses have been briefly discussed, focusing on the state-of-the-art tools and workflows of both primary and downstream analyses. The chapter also discusses the known and putative applications of human microbiome analyses.

17.2 Overview of Sequencing Methods and Bioinformatics Analysis

A brief workflow of HTS based human microbiome is shown in Fig. 17.2a–c. Depending on the scientific questions and budget, different HTS approaches are chosen. Metataxonomics is preferred for identification of microbial composition. Metagenomics is used to study the total DNA for detection of microbial genes and strain level identification, whereas metatranscriptomics is used to detect microbial gene expression.

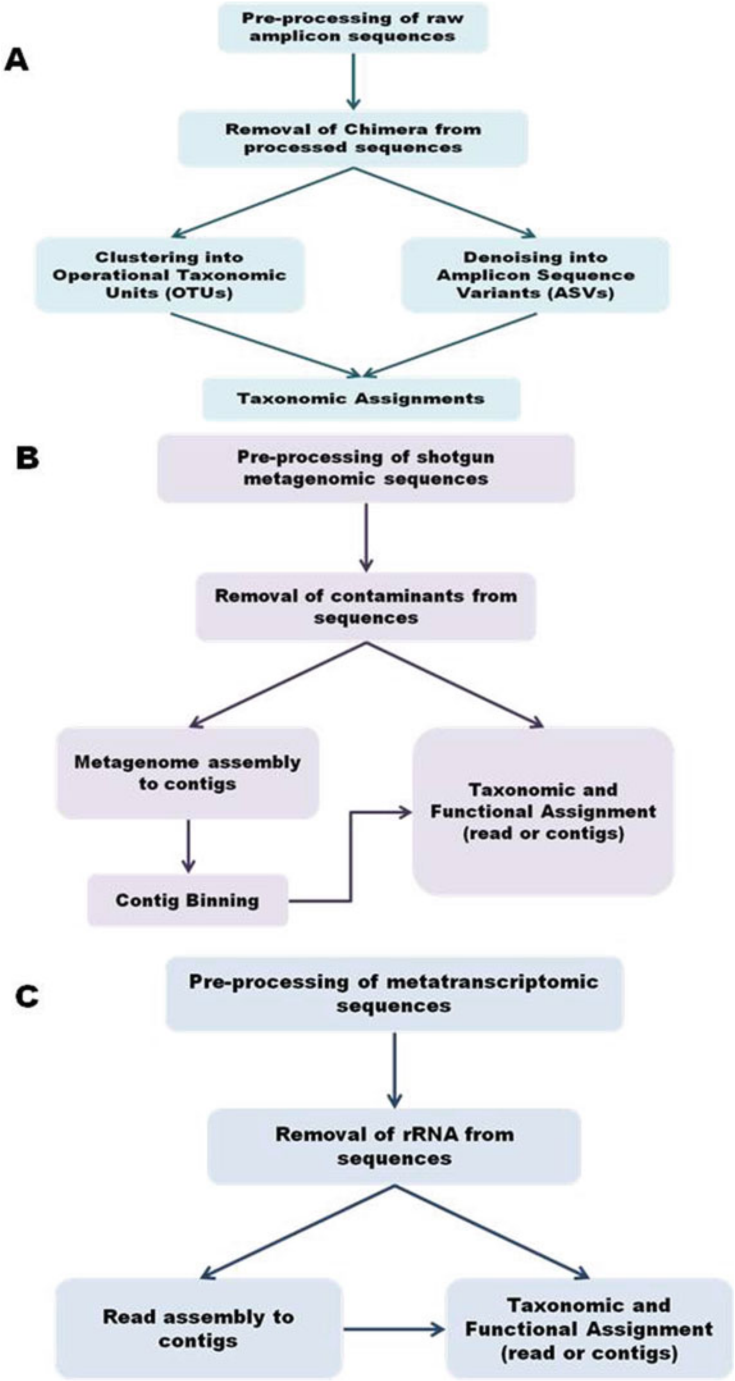

A
Pre-processing of raw amplicon sequences
Removal of Chimera from processed sequences
Clustering into Operational Taxonomic Units (OTUs)
Denoising into Amplicon Sequence Variants (ASVs)
Taxonomic Assignments
B
Pre-processing of shotgun metagenomic sequences
Removal of contaminants from sequences
Metagenome assembly to contigs
Taxonomic and Functional Assignment (read or contigs)
Contig Binning
C
Pre-processing of metatranscriptomic sequences
Removal of rRNA from sequences
Read assembly to contigs
Taxonomic and Functional Assignment (read or contigs)

Fig. 17.2 Commonly used workflow for (**a**) metataxonomic analysis, (**b**) metagenomic analysis and (**c**) metatranscriptomic analysis. The first step in each analysis is pre-processing of the raw reads

17.2.1 Pre-Processing of Sequencing Data

The initial HTS data come as raw reads in fastq format. The foremost step of analysis is the quality control of the sequence reads using a variety of computational tools to quality check, identify and remove low-quality bases and reads, low complexity reads, artefacts like primers, adapters or barcodes and remove host contamination. FastQC is the most popular pre-processing tool that provides quality control report and MultiQC is used for merging quality control report from multiple samples into a single report for easy comparison (Ewels et al. 2016). Read trimming and filtering tools like Trimmomatic, Trim Galore! and Cutadapt are used widely for DNA or RNA HTS data (Bolger et al. 2014). There are clusters of pre-processing tools such as FAST-X toolkit and BBTools for format conversion, quality report, quality trimming, filtering or masking nucleotides and removal of artefacts. Human metagenome has a big share of host nucleotide contamination that affects the microbial profiling analysis. To overcome this, KneadData, a well-designed contaminants removal tool is used often. Pre-processing is a crucial step and requires a trade-off between the sequence quality and amount of information it can provide.

17.2.2 Metataxonomics

Metataxonomics is based on amplicon sequencing of well conserved marker genes. 16S rRNA gene is highly conserved across bacteria and archaea with 9 hypervariable regions (V1-V9) to distinguish up to genus level thus it is easily targeted and amplified for identification of bacterial composition of any microbiome sample. 18S rRNA genes and ITS (Internal Transcribed Spacer) of non-transcriptional region of rRNA genes are used to identify fungi in microbiome sample. Single gene targeted amplicon sequencing is cost efficient and faster. Although with time it has become a well-established technique to study microbial diversity, their abundance and phylogenetic profile, this approach is unable to detect viruses, which are a major part of human microbiome. PCR amplification of conserved rRNA gene prevents contamination from host DNA but introduces PCR duplicates. The choice of variable region induces bias and the conserved region in rRNA gene makes it harder to differentiate the identified microbial genus into exact species and strains.

The workflow for metataxonomic analysis includes quality control and filtering, removal of artefacts like chimera, picking the representative sequences either by de-noising into Amplicon Sequence Variants (ASVs) or clustering the reads into Operational Taxonomic Units (OTUs) and finally classification of ASVs/OTUs. The initial pre-processing of the raw amplicon reads is mostly done using USEARCH or QIIME (Caporaso et al. 2010; Edgar 2010). Further removal of PCR artefacts mainly

Fig. 17.2 (continued) including removal of poor quality reads, primers and barcodes. Detailed methods are described in Sect. 17.2

chimera sequences is done using VSEARCH algorithm (Rognes et al. 2016). Mothur and QIIME use sequence alignment approach to improve clustering of the reads to OTUs (Schloss et al. 2009). OTU based clusters are the representatives of a taxonomic unit with 97% identity. Moreover, this approach fails to detect species or strain level taxa, includes sequencing errors and ignores the SNPs. More recent algorithms such as Deblur and DADA2 use alternative approach to include error profiles, sequence variations like SNPs and ability to differentiate closely related taxa (Amir et al. 2017; Callahan et al. 2016). This de-noising algorithm of developing ASVs has also been incorporated in QIIME 2 (Bolyen et al. 2019). Either of the approaches can be implemented to obtain the feature table of OTU or ASV with the quantitative frequency of features in each sample. As a final step of analysis, taxonomic assignment is done based on NCBI taxonomy databases, SILVA, RDP or Greengenes (Schoch et al. 2020; Quast et al. 2013; Cole et al. 2014; DeSantis et al. 2006).

17.2.3 Metagenomics

Metagenomics is mainly whole metagenome shotgun sequencing of all the DNA content of a microbiome sample. It is not restricted to a single gene and provides information for all genes that helps in identification of microbes up to species or strain level as well as understanding the major microbial pathways and metabolites active in that sample. This method covers all the microbes including the bacteria, archaea, fungi and viruses with little abundances. It also identifies novel and uncultured species. This approach comes with huge contamination from host DNA that requires an additional data filtration step. The complex data analysis using high performance computers along with whole genome sequencing make metagenomics costlier compare to amplicon sequencing. Shotgun sequencing is broadly sub-divided into two approaches depending upon the sequencing platforms; short-read produced by Illumina sequencers uses reference genomes for assembly and further analysis and another is long-read produced by Oxford Nanopore MinION or Pacific Biosciences Sequel can be used for de novo assemblies to identify novel genomes.

Shotgun metagenome sequence reads are often assembled into longer continuous sequences called contigs. The metagenomic assembly is done using a variety of de Bruijn graph based assemblers such as MetaVelvet, MetaSPAdes, IDBA (Afiahayati and Sakakibara 2015; Nurk et al. 2017; Peng et al. 2011, 2012). Next step is to group the sequence reads or assembled contigs from related or same organisms, known as binning. It also helps in recovery of partial or complete genomes from the metagenomic sequence data. Binning is widely classified into homology-based supervised binning and nucleotide composition based unsupervised binning. Homology-based binning methods like MetaPhlAn2 uses reference genes to cluster reads (Truong et al. 2015). Composition based methods like PhyloPythiaS and MetaCluster uses nucleotide features such as k-mer patterns (Gregor et al. 2016). There are binning tools like MaxBin, MaxBin 2, AMPHORA2, MetaBAT and

MetaBAT2 that combine both composition and homology-based approach (Wu et al. 2014, 2016; Wu and Scott 2012; Kang et al. 2015, 2019). The clustering of the sequences is possible to visualise for evaluation of binning using tools like VizBin (Laczny et al. 2015). Elviz is another tool used for visualisation of both metagenome assembly and binning (Cantor et al. 2015). As an optional step, reassembly of the reads in each bin leads to production of longer contigs and helps in metagenomic genome reconstruction. The assembly and binning are evaluated using tools like MetaQUAST, CheckM and BUSCO (Mikheenko et al. 2016; Parks et al. 2015; Seppey et al. 2019). The taxonomic assignment is another crucial step in the metagenomic analysis process. The taxonomic classification is done either using the raw reads or assembled contigs. The most primitive approach is using BLAST to match each read with the sequences of GenBank, however, the method is not computationally feasible with increasing data amount. With the development of computational techniques, various programs have been developed with strategies like aligning reads to marker genes or protein sequences, k-mer mapping or genome assembly. Marker gene based approaches such as MetaPhlAn2 use customised clade-specific genes database and GOTTCHA uses a unique strategy of creating a database of genome signatures for taxonomic profiling. Alignment requires high computational resources therefore new approaches like the k-mer mapping algorithms that built a simple lookup table which requires lesser computational work are developed. Kraken and CLARK are two popular k-mer based classifiers used for faster identification of metagenomic reads. k-mers are also represented using de Bruijn graphs as implemented in Kallisto to find strain level abundances. For more sensitive metagenomic classification, translated reads are compared with protein sequence databases as done by DIAMOND, Kaiju and MEGAN.

17.2.4 Metatranscriptomics

Metatranscriptomics is another next generation sequencing technology-based approach that uses mRNA content from the microbiome sample to detect the active functional genes, their expression profiles and related pathways. It requires sophisticated methodology and experts to perform the experiment as well as analysis. For better understanding the diversity and comparison of active microbial pathways across microbiome samples, the transcriptomics data is merged with metagenomics for referencing and diminishing the noise and contaminations.

Similar to other two HTS analysis, the pre-processing step includes quality assessment, removal of sequencing errors, poor quality reads or bases and removal of adapter and primers. Since metatranscriptomics aim to sequence any RNA, huge amount of rRNA sequences are found. Most commonly SortMeRNA is used to filter rRNA sequences that are mainly used for taxonomic assignment instead of functional information (Kopylova et al. 2012). The non-rRNA reads are further used for both community profiling and functional profiling same as metagenomics analysis.

17.2.5 Databases for Microbial Taxonomic Assignments

The heart of the microbiome data analysis is assignment of correct taxonomy to the reads. Both marker gene sequencing and shotgun sequencing provide information on the microbial composition of the microbiome sample using references from public databases. The analysis of metataxonomic data is assisted by few comprehensive resources such as SILVA, RDP, Greengenes and UNITE (Quast et al. 2013; Cole et al. 2014; DeSantis et al. 2006; Balvociute and Huson 2017; Nilsson et al. 2019). SILVA (http://www.arb-silva.de) is an updated and non-redundant database of aligned small and large subunit rRNA gene sequences from Bacteria, Archaea and Eukaryota. Ribosomal Database Project (RDP; http://rdp.cme.msu.edu/) contains aligned and annotated bacterial and archaeal small subunit rRNA genes and fungal large subunit rRNA genes. Greengenes (http://greengenes.lbl.gov) is a 16S rRNA database that includes chimera screening as an exclusive feature. UNITE (https://unite.ut.ee/) is a database of fungal ribosomal internal transcribed spacer (ITS) region. Metagenomics is not restricted to single gene thus metagenomic classifiers use large number of genes for taxonomic assignments. A widely used taxonomic assignment resource is NCBI taxonomy database that comprises of all the organism names associated with sequence submission in NCBI. For identification of genes from metagenomic reads or merged contigs, NCBI non-redundant (nr) database is used as reference database. Metagenomic Phylogenetic Analysis2 (MetaPhlAn2) is a widely used tool that has its own database of clade-specific marker genes identified from bacteria, archaea, viruses and eukaryotes for taxonomic profiling of the reads (Segata et al. 2012).

17.3 Downstream Analysis

The final outputs from HTS analysis are microbial taxonomic and active gene feature tables. These output files are further analysed to answer scientific questions such as diversity of microbial composition in samples or across sample groups, identification of pathogen species, the significant microbial genes and functional pathways specific to sample groups, identification of variants, genome structure and phylogeny. These sample groups are usually representation of diseases; drug or antibiotic exposed or control individuals. The downstream analyses from the taxonomy and gene tables are done using various statistical analysis and visualisation tools as tabulated in Table 17.1.

17.3.1 Microbial Taxonomic Analysis

Downstream analysis from taxonomic profiles involves finding alpha and beta diversity and comparing them among sample groups, finding differential abundance of taxa and correlation between taxa and metadata. Alpha diversity is measured using Shannon diversity index and Shannon evenness index that defines the species

Table 17.1 List of some computational tools for downstream analysis and visualisation

Tool/software	Description	URL
Anvi'o (Eren et al. 2015)	A platform to analyse and visualise microbial assembly and binning	http://merenlab.org/software/anvio/
BURRITO (McNally et al. 2018)	An interactive visualisation tool of multiomic microbiome data to pair taxonomic and functional information	https://github.com/borenstein-lab/burrito
Elviz (Cantor et al. 2015)	An interactive web tool for visualisation of assembled metagenomes along with metadata and sequence parameters.	https://genome.jgi.doe.gov/viz/
FragGeneScan (Rho et al. 2010)	A method to predict genes using hidden Markov model from metagenomic data	https://sourceforge.net/projects/fraggenescan/
GraPhlAn (Asnicar et al. 2015)	Tool for visualisation of microbial genomes and metagenomes along with phylogenies, metadata and abundances.	https://huttenhower.sph.harvard.edu/graphlan
HUMAnN2 (Franzosa et al. 2018)	A pipeline for profiling microbial pathways and their abundance from metagenomic or metatranscriptomic sequencing data	https://huttenhower.sph.harvard.edu/humann
Krona (Ondov et al. 2011)	An interactive tool to visualise hierarchies of metagenomic classifications along with the relative abundances and confidences.	https://github.com/marbl/Krona/wiki
LEfSe (Segata et al. 2011)	A program for biomarker discovery and identification of genomic features such as taxa, gene or pathway to differentiate between classes	https://huttenhower.sph.harvard.edu/lefse/
MaAsLin2 (Himel Mallick et al. 2021)	A program for determining multivariable association between phenotypes, environments, exposures, covariates and microbial metaomic features	https://huttenhower.sph.harvard.edu/maaslin/
MetaCHIP (Song et al. 2019)	A pipeline to predict horizontal gene transfer from metagenomic data.	https://github.com/songweizhi/MetaCHIP
MetaGeneMark (Zhu et al. 2010)	Ab initio prediction of gene from shotgun sequences	http://exon.gatech.edu/meta_gmhmmp.cgi
MetaProdigal (Hyatt et al. 2012)	A program to identify genes from short sequences with high accuracy and ability to identify sequences with alternate genetic codes	https://github.com/hyattpd/prodigal
mmvec (Morton et al. 2019)	A program to predict microbe–metabolite interactions from multiomic microbiome data	https://github.com/biocore/mmvec
Phinch (Holly M Bik 2014)	An interactive web-based framework to explore multiomic microbiome data	http://phinch.org/
Phyloseq (McMurdie and Holmes 2013)	A R package to explore microbiome phylogenetic profiles	https://joey711.github.io/phyloseq/
PICRUSt (Douglas et al. 2018)	A package to predict microbial functions from 16S rRNA analysis	http://picrust.github.io/picrust/
shortBRED (Kaminski et al. 2015)	A program for profiling protein families of interest from shotgun metaomic sequencing data with high specificity	https://huttenhower.sph.harvard.edu/shortbred

(continued)

Table 17.1 (continued)

Tool/software	Description	URL
Tax4Fun (Asshauer et al. 2015)	A package to predict functional profile of microbial communities from 16S rRNA data	http://tax4fun.gobics.de/
TIME (Baksi et al. 2018)	A web server to analyse microbiome time series data	https://web.rniapps.net/time/
VizBin (Laczny et al. 2015)	A de novo visualisation, inspection and binning of metagenomic datasets from single samples	https://github.com/claczny/VizBin

richness, diversity and evenness within a sample. The comparison of alpha diversity among or between groups is statistically determined using ANOVA, Mann–Whitney U test and Kruskal–Wallis test. The alpha diversity is visually represented using box-plots, Venn diagrams and rarefaction curves. Beta diversity finds variation in microbial composition between samples using Bray–Curtis dissimilarity, Jaccard distance and weighted, unweighted UniFrac. It is paired with principal coordinate analysis (PCoA), non-metric multi-dimensional scaling (NMDS) and constrained PCoA (CPCoA) to obtain visual outputs. Beta diversity is visually compared among samples or groups using scatter-plots and dendograms. Several programs are available for calculating the alpha and beta diversity such as QIIME, phyloseq, vegan and USEARCH (Edgar 2010; McMurdie and Holmes 2013). Volcano plots, Manhattan plots and tools like LefSe are used to find differential abundances of taxa and significant determinant taxon between groups of samples (Segata et al. 2011). Correlation coefficient curve, linear fitting curve and heatmaps are used to find correlation between taxonomic profile and metadata. The phylogenetic tree and cladogram are used to understand the phylogenetic and taxonomic hierarchy. GraPhlAn is a software that provides attractive publication-ready phylogenetic trees (Asnicar et al. 2015). Other popular visualisation tools are Krona an interactive visualisation tools to explore the relative abundances along with hierarchical classifications and TIME (Temporal Insights into Microbial Ecology) that allows prediction of taxonomical markers for different sample groups (Ondov et al. 2011; Baksi et al. 2018). Horizontal gene transfer (HGT) is a crucial phenomenon in bacteria especially involved in the spread of antibiotic resistance and human microbiome, the reservoirs of microbes is explored using tool such as MetaCHIP that helps in the identification of HGT from metagenomic datasets (Song et al. 2019).

17.3.2 Microbial Functional Analysis

The downstream analysis and metabolic pathway information can be drawn from functional analysis of microbiome study. The marker gene analysis only gives insight in to the microbial composition, but there are tools like PICRUSt and Tax4Fun that assign metabolic functions to the samples by mapping the 16S reads to annotated genomes (Douglas et al. 2018; Asshauer et al. 2015). However, for detailed and accurate functional profiling, shotgun and transcriptome sequencing are highly recommended. Genes are identified from assembled contigs of shotgun reads using tools like MetaGeneMark, FragGeneScan and MetaProdigal (Zhu et al. 2010; Rho et al. 2010; Hyatt et al. 2012). The identified genes or ORF are further analysed to predict the function. BLASTn and BLASTp use NCBI GenBank databases or UniProt databases for homology-based search to annotate the contigs (UniProt 2019). Hidden Markov model based HMMER and support vector machine model based PhyloPythiaS(+) are also used as similarity search tools (Eddy 2008). The widely used databases for prediction of gene function, pathways or functional domains are PFAM, COG, SEED, eggNOG, KEGG and TIGRFAM (El-Gebali et al. 2019; Galperin et al. 2019; Huerta-Cepas et al. 2019; Overbeek et al. 2014; Kanehisa et al. 2017; Haft et al. 2013). There are few tools to visualise the annotated functional gene information and compare among sample groups such as HUMAnN2, LEfSe and shortBRED (Franzosa et al. 2018; Kaminski et al. 2015).

17.4 Integrating Multiomic Data of Microbiome Samples

Merging the different omics data to understand the contribution of microbiome in human biology is an advanced approach and the computational methods to integrate multi-dimensional data are emerging. A recently published tool, mmvec uses neural networks to merge multiomics microbiome data for prediction of microbe–metabolite interactions that helps in determining the microbial origin of a particular metabolite (Morton et al. 2019). Burrito is also an interactive multiomic visualisation tool to merge taxonomic and functional data (McNally et al. 2018). MaAsLin2 is another recent approach that merges metadata such as human health details, diet, environmental conditions or other features to microbial community profile using linear models. There are visualisation tools that are used to explore genes, proteins and microbes data such as Phinch. Anvi'o is a platform for multiomic data re-analysis and visualisation (Eren et al. 2015).

17.5 Pre-Designed Pipelines and Web-Analysis Platforms

Microbiome data analysis is a complicated process with many overlapping steps and pressure points. Several analysis pipelines have been developed to facilitate full or partial analysis. QIIME, QIIME 2 and mothur are comprehensive amplicon sequencing analysis pipelines with integrated scripts to perform steps from quality control to diversity visualisation (Caporaso et al. 2010; Bolyen et al. 2019; Schloss 2020). Similarly, MetAMOS and ANASTASIA are pipelines for metagenomic assembly and gene annotation (Treangen et al. 2013; Koutsandreas et al. 2019). Another mentionable pipeline, SqueezeMeta does real time analysis of metagenomic data from nanopore technology (Tamames and Puente-Sanchez 2018). Metatranscriptomics is relatively new in the field and have very few tools and pipelines. SAMSA2 is a dedicated metatranscriptomic data analysis tool that handles from quality control to visualisation (Westreich et al. 2018). MetaQUBIC is a pipeline that pairs multiomics for analysis to detect gene module from metagenomic and metatranscriptomic data (Ma et al. 2019). Apart from these pipelines, web services such as MG-RAST, EBI Metagenomics, IMG/M and Qiita provide automated end-to-end processing of the data (Meyer et al. 2019; Mitchell et al. 2020; Chen et al. 2019; Gonzalez et al. 2018). An alternative approach is using Galaxy, an open-source workflow system comprising numerous tools for each step of analysis and allows customisation of frameworks such as ASaiM (Afgan et al. 2018; Batut et al. 2018).

17.6 Challenges and Best Practices for Microbiome Analysis

The biggest challenge of computational analysis of Microbiome data is not about the chosen pipeline for analysis instead the experimental factors have more influence on the outcome. Since the field is evolving rapidly, it is essential that the analyses are reproducible. To tackle such challenges, the metadata of the experimental procedures including the host phenotype, sample site, sampling technique, nucleotide extraction method, primers and barcodes must be available with the sequence data. The recommended data collection procedures should be followed as mentioned in minimum information about a marker genes (MIMARKS) and metagenomes (MIMS), minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea and uncultivated virus genome (MIUViG) (Field et al. 2008; Yilmaz et al. 2011; Bowers et al. 2017; Roux et al. 2019). The experimental handling introduces contaminations that modify the microbial composition of the sample. Therefore, usage of proper controls and considering consistency across all similar samples should be kept in mind as recommendations such as International Human Microbiome Standards and the Microbiome Quality Control (MBQC) project (Sinha et al. 2017; Costea et al. 2017). Equal efforts are also given for raw data storage and standardisation and reproducibility of the data analyses using cloud and open-source resources such as Qiita, EBI Metagenomics, Sequence Read Archive (SRA) and MG-RAST (Mitchell et al. 2020; Gonzalez et al. 2018; Kodama et al. 2012; Keegan et al. 2016).

17.7 Application of Human Microbiome Research in Human Diseases

The insight study of human microbiome manifest that the microbes of a healthy individual greatly differs from a non-healthy or diseased individual. Thereby, characterising microbiome based on their differential trait could serve as a potential tool in identifying disease risk, prognosis, phenotype and response to treatment. In diseases like inflammatory bowel disease (IBD), obesity, diabetes and cancer, microbiome studies help to identify the pathophysiology of diseases and inspired further studies to understand the link of immune system and therapeutics with human Microbiome (McCarville et al. 2016; Kostic et al. 2012; Durack et al. 2018; Chen et al. 2014). As a therapeutic aspect of microbiome research, a number of studies indicated the faecal microbial transplant (FMT) enhances the antitumor effect in cancer (Vetizou et al. 2015; Routy et al. 2018). Dysbiosis of gut is linked with various pathological disorders, probiotics supplements have restored the balance of microbial community by the production of certain metabolite and could enhance the immune effect (Ritchie and Romanuk 2012). Till now gut-brain axis and gut-lung axis have been studied to explore the influence of gut microbiome on functioning of these major organs (Valles-Colomer et al. 2019; Enaud et al. 2020). Antibiotics are found to modify the dynamics of human microbiome, increase the abundance of antimicrobial resistance genes and impair the healthy microbiota (McInnes et al. 2020). Recently few studies are emerging that explore human Microbiome to control and manage antibiotic resistance (Zhang et al. 2020; Yang et al. 2016). Altogether these studies indicated the importance of microbiome in human physiology, immunity and metabolism. Further development and advances in high-throughput sequencing, omics and other computational resources empower our understanding regarding large data generation, standardisation of protocols, reference genomes and analysis pipelines for human microbiome study.

References

Afgan E et al (2018) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46:W537–W544

Afiahayati KS, Sakakibara Y (2015) *MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning*. DNA Res 22(1):69–77

Altschul SF et al (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Amir A et al (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:2

Asnicar F et al (2015) Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ 3:e1029

Asshauer KP et al (2015) Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics 31(17):2882–2884

Baksi KD, Kuntal BK, Mande SS (2018) TIME': a web application for obtaining insights into microbial ecology using longitudinal Microbiome data. Front Microbiol 9:36

Balvociute M, Huson DH (2017) SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? BMC Genomics 18(Suppl 2):114

Batut B et al (2018) ASaiM: a galaxy-based framework to analyze microbiota data. Gigascience 7 (6):057

Bay L et al (2020) Universal dermal microbiome in human skin. MBio 11(1):02945

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120

Bolyen E et al (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37(8):852–857

Bowers RM et al (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35 (8):725–731

Callahan BJ et al (2016) DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13(7):581–583

Cantor M et al (2015) Elviz - exploration of metagenome assemblies with an interactive visualization tool. BMC Bioinformatics 16:130

Caporaso JG et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335–336

Caporaso JG et al (2011) Moving pictures of the human microbiome. Genome Biol 12(5):R50

Chen Z et al (2014) Incorporation of therapeutically modified bacteria into gut microbiota inhibits obesity. J Clin Invest 124(8):3391–3406

Chen IA et al (2019) IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res 47(D1):D666–D677

Cole JR et al (2014) Ribosomal database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633–D642

Compeau PE, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 29(11):987–991

Costea PI et al (2017) Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol 35(11):1069–1076

Costello EK et al (2009) Bacterial community variation in human body habitats across space and time. Science 326(5960):1694–1697

Das B et al (2018) Analysis of the gut Microbiome of rural and urban healthy Indians living in sea level and high altitude areas. Sci Rep 8(1):10104

DeSantis TZ et al (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72(7):5069–5072

Douglas GM, Beiko RG, Langille MGI (2018) Predicting the functional potential of the Microbiome from marker genes using PICRUSt. Methods Mol Biol 1849:169–177

Durack J et al (2018) Delayed gut microbiota development in high-risk for asthma infants is temporarily modifiable by lactobacillus supplementation. Nat Commun 9(1):707

Eddy SR (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol 4(5):e1000069

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26 (19):2460–2461

El-Gebali S et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47(D1): D427–D432

Enaud R et al (2020) The gut-lung Axis in health and respiratory diseases: a place for inter-organ and inter-kingdom Crosstalks. Front Cell Infect Microbiol 10:9

Eren AM et al (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319

Ewels P et al (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32(19):3047–3048

Field D et al (2008) The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26(5):541–547

Franzosa EA et al (2018) Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 15(11):962–968

Galperin MY et al (2019) Microbial genome analysis: the COG approach. Brief Bioinform 20 (4):1063–1070

Gonzalez A et al (2018) Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 15 (10):796–798

Gregor I et al (2016) PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. Peer J 4:e1603

Grice EA et al (2009) Topographical and temporal diversity of the human skin microbiome. Science 324(5931):1190–1192

Haft DH et al (2013) TIGRFAMs and genome properties in 2013. Nucleic Acids Res 41:D387–D395

Himel Mallick, LJM, Rahnavard A, Ma S, Zhang Y, Nguyen LH, Tickle TL, Weingart G, Ren B, Schwager E, Subramanian A, Lu Y, Waldron L, Paulson JN, Franzosa EA, Bravo HC, Huttenhower C (2021) Multivariable association in population-scale meta-omics studies

Holly M Bik PI (2014) Phinch: an interactive, exploratory data visualization framework for–Omic datasets. In: bioRxiv

Huerta-Cepas J et al (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:309–314

Huey SL et al (2020) Nutrition and the gut microbiota in 10- to 18-month-old children living in urban slums of Mumbai, India. mSphere 5:5

Human Microbiome Project (2012a) Structure, function and diversity of the healthy human microbiome. Nature 486(7402):207–214

Human Microbiome Project (2012b) A framework for human microbiome research. Nature 486 (7402):215–221

Hyatt D et al (2012) Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28(17):2223–2230

Kaminski J et al (2015) High-specificity targeted functional profiling in microbial communities with ShortBRED. PLoS Comput Biol 11(12):e1004557

Kanehisa M et al (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45(D1):D353–D361

Kang DD et al (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165

Kang DD et al (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359

Karlsson F et al (2013) Assessing the human gut microbiota in metabolic diseases. Diabetes 62 (10):3341–3349

Keegan KP, Glass EM, Meyer F (2016) MG-RAST, a metagenomics Service for Analysis of microbial community structure and function. Methods Mol Biol 1399:207–233

Knights D et al (2011) Bayesian community-wide culture-independent microbial source tracking. Nat Methods 8(9):761–763

Kodama Y et al (2012) The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res 40:D54–D56

Kopylova E, Noe L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28(24):3211–3217

Kostic AD et al (2012) Genomic analysis identifies association of fusobacterium with colorectal carcinoma. Genome Res 22(2):292–298

Koutsandreas T et al (2019) ANASTASIA: an automated metagenomic analysis pipeline for novel enzyme discovery exploiting next generation sequencing data. Front Genet 10:469

Laczny CC et al (2015) VizBin—an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome 3(1):1

Li D et al (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31(10):1674–1676

Ma A et al (2019) MetaQUBIC: a computational pipeline for gene-level functional profiling of metagenome and metatranscriptome. Bioinformatics 35(24):5397

McCarville JL, Caminero A, Verdu EF (2016) Novel perspectives on therapeutic modulation of the gut microbiota. Therap Adv Gastroenterol 9(4):580–593

McInnes RS et al (2020) Horizontal transfer of antibiotic resistance genes in the human gut microbiome. Curr Opin Microbiol 53:35–43

McMurdie PJ, Holmes S (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8(4):e61217

McNally CP et al (2018) BURRITO: an interactive multi-omic tool for visualizing taxa-function relationships in Microbiome data. Front Microbiol 9:365

Mehta RS et al (2018) Stability of the human faecal microbiome in a cohort of adult men. Nat Microbiol 3(3):347–355

Meyer F et al (2019) MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. Brief Bioinform 20(4):1151–1159

Mikheenko A, Saveliev V, Gurevich A (2016) MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32(7):1088–1090

Mitchell AL et al (2020) MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res 48 (D1):D570–D578

Morton JT et al (2017) Balance trees reveal microbial niche differentiation. mSystems 2:1

Morton JT et al (2019) Learning representations of microbe-metabolite interactions. Nat Methods 16(12):1306–1314

Namiki T et al (2012) MetaVelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40(20):e155

Nayfach S et al (2019) New insights from uncultivated genomes of the global human gut microbiome. Nature 568(7753):505–510

Nilsson RH et al (2019) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res 47(D1):D259–D264

Nishijima S et al (2016) The gut microbiome of healthy Japanese and its microbial and functional uniqueness. DNA Res 23(2):125–133

Nurk S et al (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Res 27 (5):824–834

Oh M, Zhang L (2020) DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep 10(1):6026

Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a web browser. BMC Bioinformatics 12:385

Ounit R et al (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16:236

Overbeek R et al (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res 42:206–214

Parks DH et al (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25(7):1043–1055

Pasolli E et al (2019) Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176(3):649–662

Peng Y et al (2011) Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27 (13):94–101

Peng Y et al (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28(11):1420–1428

Qian X et al (2020) Gut microbiota in children with juvenile idiopathic arthritis: characteristics, biomarker identification, and usefulness in clinical prediction. BMC Genomics 21(1):286

Qiao Y et al (2018) MetaBinG2: a fast and accurate metagenomic sequence classification system for samples with many unknown organisms. Biol Direct 13(1):15

Qin J et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59–65

Quast C et al (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:590–596

Rajilic-Stojanovic M et al (2012) Long-term monitoring of the human intestinal microbiota composition. Environ Microbiol 10:15

Relman DA, Lipsitch M (2018) Microbiome as a tool and a target in the effort to address antimicrobial resistance. Proc Natl Acad Sci USA 115(51):12902–12910

Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 38(20):e191

Ritchie ML, Romanuk TN (2012) A meta-analysis of probiotic efficacy for gastrointestinal diseases. PLoS One 7(4):e34938

Robertson RC et al (2019) The Human Microbiome and child growth—first 1000 days and beyond. Trends Microbiol 27(2):131–147

Rognes T et al (2016) VSEARCH: a versatile open source tool for metagenomics. Peer J 4:e2584

Routy B et al (2018) Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. Science 359(6371):91–97

Roux S et al (2019) Minimum information about an uncultivated virus genome (MIUViG). Nat Biotechnol 37(1):29–37

Schloss PD (2020) Reintroducing mothur: 10 years later. Appl Environ Microbiol 86(2):e02343

Schloss PD et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75(23):7537–7541

Schoch CL et al (2020) NCBI taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020:062

Segata N et al (2011) Metagenomic biomarker discovery and explanation. Genome Biol 12(6):R60

Segata N et al (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9(8):811–814

Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol 1962:227–245

Shi W et al (2019) gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. Nucleic Acids Res 47(1):637–648

Sinha R et al (2017) Assessment of variation in microbial community amplicon sequencing by the Microbiome quality control (MBQC) project consortium. Nat Biotechnol 35(11):1077–1086

Sommer F et al (2017) The resilience of the intestinal microbiota influences health and disease. Nat Rev Microbiol 15(10):630–638

Song W et al (2019) MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. Microbiome 7(1):36

Stennett CA et al (2020) *A cross-sectional pilot study of birth mode and vaginal microbiota in reproductive-age women*. PLoS One 15(4):0228574

Subramanian S et al (2014) Persistent gut microbiota immaturity in malnourished Bangladeshi children. Nature 510(7505):417–421

Sun J et al (2020a) Role of the oral microbiota in cancer evolution and progression. Cancer Med 9:6306–6321

Sun Y et al (2020b) Population-level configurations of gut mycobiome across six ethnicities in urban and rural China. Gastroenterology 6:31–38

Susic D et al (2020) Microbiome Understanding in Maternity Study (MUMS), an Australian prospective longitudinal cohort study of maternal and infant microbiota: study protocol. BMJ Open 10(9):e040189

Tamames J, Puente-Sanchez F (2018) SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. Front Microbiol 9:3349

Tamburini S et al (2016) The microbiome in early life: implications for health outcomes. Nat Med 22(7):713–722

Thang MWC et al (2019) MetaDEGalaxy: galaxy workflow for differential abundance analysis of 16s metagenomic data. F1000Res 8:726

The Integrative HMP (iHMP) Research Network Consortium (2019) The Integrative Human Microbiome Project. Nature 569(7758):641–648

Tigchelaar EF et al (2015) Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open 5(8): e006772

Treangen TJ et al (2013) MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol 14(1):R2

Truong DT et al (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12(10):902–903

UniProt C (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47(D1): D506–D515

Valles-Colomer M et al (2019) The neuroactive potential of the human gut microbiota in quality of life and depression. Nat Microbiol 4(4):623–632

Vangay P, Hillmann BM, Knights D (2019) Microbiome learning repo (ML repo): a public repository of microbiome regression and classification tasks. Gigascience 8:5

Vetizou M et al (2015) Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. Science 350(6264):1079–1084

Westreich ST et al (2018) SAMSA2: a standalone metatranscriptome analysis pipeline. BMC Bioinformatics 19(1):175

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15(3):46

Wu M, Scott AJ (2012) Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28(7):1033–1034

Wu YW et al (2014) MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2:26

Wu YW, Simmons BA, Singer SW (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32(4):605–607

Yang Z et al (2016) Preliminary analysis showed country-specific gut resistome based on 1,267 feces samples. Gene 581(2):178–182

Yatsunenko T et al (2012) Human gut microbiome viewed across age and geography. Nature 486 (7402):222–227

Yilmaz P et al (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29 (5):415–420

Zhang L et al (2020) Characterization of antibiotic resistance and host-microbiome interactions in the human upper respiratory tract during influenza infection. Microbiome 8(1):39

Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38(12):e132

Zimmermann M et al (2019) Separating host and microbiome contributions to drug pharmacokinetics and toxicity. Science 363(6427):9931

18 Neural Network Analysis

Amit Joshi, Jitendra Sasumana, Nillohit Mitra Ray, and Vikas Kaushik

Abstract

Neural networks play very significant role when it comes to analysis of proteins and nucleic acid sequences. Many of the pattern recognition software are based on neural networks for prediction of biological patterns. Modern sequencing advancement fuels up the collection of data related to DNA, RNA, and protein sequences. The complexity and enormous size of this data require best computational algorithms for analysis and interpretation. This information will assist in developing useful insight for biomolecular structural predictions and prediction of interactions between such molecules. A neural system investigation framework is a succession of computations that attempts to see concealed associations in a lot of data through a technique that imitates the way where the human mind works. In this sense, neural frameworks suggest systems of neurons, artificial in nature. Vectors and matrices based linear algebra and topology designs supported various types of neural architectures. Neural frameworks can conform to advancing info; so the framework makes the best result without hoping to refresh the yield rules. The possibility of neural frameworks, which has its basic establishments in man-made consciousness, is rapidly getting ubiquity in the progression of in silico designing systems. Here, we talk about and sum up the uses of Neural Networks in computational biology, with a specific spotlight on applications in protein and Nucleic acid bioinformatics. We concluded with giving basic insights of neural networks in multiple domains of life sciences like gene prediction, protein structure prediction, epitope prediction, expression, co-expression, protein–protein interaction, and many other domains.

A. Joshi · J. Sasumana · N. M. Ray · V. Kaushik (✉)
Domain of Bioinformatics, School of Bioengineering and Biosciences, Lovely Professional University, Phagwara, Punjab, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_18

Keywords

Protein · DNA · Bioinformatics · Neural networks

18.1 Introduction

Bioinformatics is an amalgamation of biotechnology and computer science, to provide better understanding of biomolecular interactions. From the beginning of Human genome project it was realized that the storing sequencing data will assist in future medicinal developments (Collins et al. 2003). DNA, RNA, Protein sequences of different organism serve as biological data. Enormous amount of biological data requires analysis to generate applicable information like genomic and proteomic interactions. Such bioinformatics analysis assisted not only in finding similarities between sequential stretches of nucleotides or amino acids but also in determining expression levels and control for genomic functional sets. In the past decade, determination of protein based vaccine candidates for different viruses and bacterial organisms become very easy for computational biologists. Neural network is based on cognitive learning of system with statistical probabilistic approach to predict the outcomes in a similar fashion like millions of interconnected functional neurons does (Hopfield 1982). Prediction based modeling of bimolecular structures, and determination of their functionality was found as greater achievement of artificial neural networks in denovo studies. These neural artificial intelligence systems can be utilized for prescient displaying, versatile control and applications where they can be prepared through a dataset. Self-evolving occurs because of experience that exists inside such systems, which can get inferences from a complex and apparently random arrangement of data. Propagation of theories related to neural networks were credited to Alexander Bain (Bain 1873), suggested that interconnections and electrical activity between neuron were responsible for cognitive learning and behavioral actions (Evans 1990). A neural framework is synaptic organization of counterfeit neuronal units that represent an artificial scientific or programming based coded model for information processing. Artificial neural network (ANN) is a flexible structure that changes its structure subject to outside or inside information that travels in the wholesome framework. In this chapter we observe fundamentals of algorithms behind ANN based web servers used in in silico methodologies. Many mathematical modeling tools along with coding developments lead to design fast and effective software and algorithms that evolve with more input data (Biological sequences like DNA, RNA, and proteins). Also observe tools that are deployed in proteomics and genomics for describing biomolecular structure, properties, and functional interactions. Firstly we will introduce biochemical background then summaries the type of neural networks algorithm that were commonly used in recent software's/servers designing for accurate structural and functional predictions. Also details of applications in genomics, transcriptomics, and proteomics are given to develop understanding of neural networks. Machine learning in silico tools based on artificial neural networking are very useful in bioinformatics and provide ease in genomic as well as proteomic analysis. This will resolve big data analysis problems too. Aim of this chapter considered useful, as enormous data should not become problem for investigator rather it will act as boon for life sciences and medical world to develop better and fast regimens for several diseases and this will also interconnect health sector globally.

18.2 Biochemistry and Bioinformatics Background

Proteins and nucleic acids are integral part of each and every cellular entity of all living organisms. These biomolecules participate in almost all functional activities within cells from metabolic reactions to genetic expression of transcriptomes. Many life sustaining processes like cell cycle, apoptosis, enzymatic catalysis, cell signaling, adhesion, and central dogma are always depending on protein–protein and DNA–protein interactions. Proteins and nucleic acids are macromolecular heteropolymers of amino acids and nucleotides, respectively (Giorgini et al. 2020). Different proteins have different amino acids sequences, peptide bond forms between two amino acids due to bonding between amino and carboxyl group of adjacent amino acids (Fig. 18.1b and 18.1c). Similarly in nucleic acid (DNA or RNA) phosphodiester bonds exist between two adjacent nucleotides (Deoxyribose or Ribose sugar, Nitrogenous base (A,T,C,G), and phosphate group) (Fig. 18.1a). Sequencing studies produce enormous data regarding DNA, RNA, and protein sequences. These sequences are stored in databases like NCBI Genbank, DDBJ, and EMBL. Similarly structural information of proteins and ligands interacting to

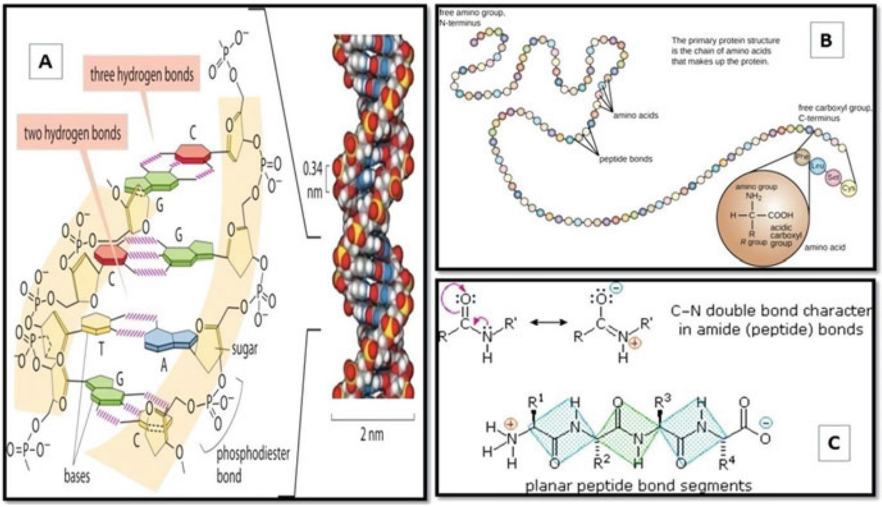

A
three hydrogen bonds
two hydrogen bonds
C
G
G
C
T
A
G
C
sugar
phosphodiester bond
bases
0.34 nm
2 nm
B
free amino group, N-terminus
The primary protein structure is the chain of amino acids that makes up the protein.
amino acids
peptide bonds
free carboxyl group, C-terminus
Phe
Leu
Ser
Cys
amino group
NH2
H—C—COOH
R
R group
acidic carboxyl group
amino acid
C-N double bond character in amide (peptide) bonds
planar peptide bond segments
R1
R2
R3
R4
C

Fig. 18.1 (**a**) DNA structure: revealing phosphodiester bonds and hydrogen bond (**b**) protein chain: primary structure (**c**) peptide bond between amino acids them are also submitted in databases like Pubchem, Chembridge, Maybridge, and Protein Data Bank (RCSB-PDB) by X-ray crystallographic experiments.

Bioinformatics not only provide platform to analyze the structures of protein in primary, secondary, and tertiary forms but also allow user to bring structural changes according to the influence of molecules in its vicinity. This means that protein structures show bonding or interactions with other biomolecules, which can impart inducible changes as in case with induced fit model analogy for enzymatic actions (Morgat et al. 2020). Computational studies created applications for deep neural networks to assist in prediction of protein–protein binding pockets or interaction sites (Zeng et al. 2020). Even protein contacts can also be predicted by metagenomic sequence data and residual neural networks (Wu et al. 2020). Such all recent studies indicate the importance of neural network bioinformatics and biochemistry together to generate a big informative picture of biomolecules functional aspect with accuracy and precision. These studies also suggest that the unsolved structure for known sequences could be easily determined by deploying neural networks architectures in biochemistry and medicine studies.

18.3 Neural Networks and Its Types

Neural networks are part of Artificial intelligence and Machine learning. These networks works like brain neurons, these networks are dependent on weights as we increase loads the networks learn more to predict suitable results or outputs. Deep learning becomes more advanced with the increase in data. Therefore, a neural network also adjusts its performance to greater extant as they grows bigger and deals with enormous flow of information. Neural networks are best in comparison to other machine learning tools that reach a plateau after a point. Activation functions play crucial role in switching on and off artificial nets that connects artificial neural elements. This allows systematic flow of information and deep learning in software based systems (Khan 2020). Each neural element receives multiple inputs and randomized loads and adds them to static bias of each neural element, then directs them to activator function that finally brings output of desired neural element of network (Fig. 18.2a). Activator function can be of linear (simplest), heviside step, and sigmoid function (complex) type. When final neural layer generates output, loss functions are calculated on the basis of inputs and outputs and back propagation conducted to bring alterations in loads that lead to loss minimization (Fig. 18.2b). To determine overall loads adjustment is main or central criteria of neural network architectures (Galushkin 2007).

Neural network architectures can be divided in many subtypes: based on framework, Datum transfer, Counterfeit neurons with weighted-density, multiple-layering and activation functions (Amato et al. 2013).Common types of neural networks are Feed forward network, Multi-layer perceptron, Convolutional neural network, Radial basis neural network, Recurrent neural network, modular networks, etc. (Fig. 18.2c and Fig. 18.3).

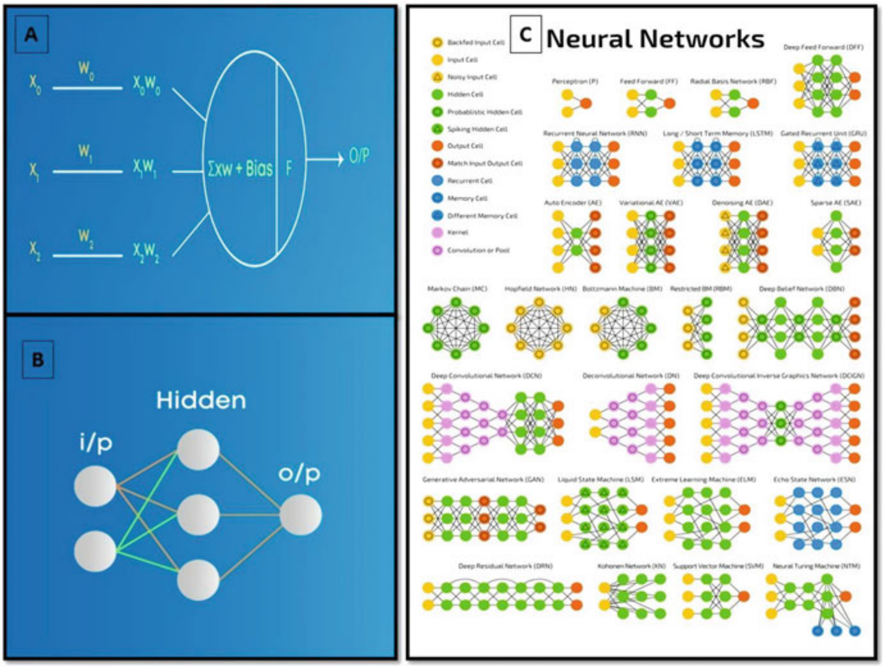


Fig. 18.2 (**a**) Load (numeric) values product with input data in back track to minimize loss, and the relation with activation function "F" to generate output. (**b**) Input layer "i/p" exhibits dimension of input vector, Hidden layer shows intermediary nodes separating input spaces with boundary limits that consider input load sets to synthesize information by activation function. Information as output "o/p" layer shows final information via ANN system architecture. (**c**) Various types of neural network architectures

Feed forward neural network was simplest first ANN to be deployed in bioinformatics. In such system one way flow of information exist (from input to hidden to output). No loops or closed cycles exist in such architecture.

The Feed forward network does not possess backward propagation. These systems have static loads (Shao 2020). Mostly step activation function is used here with 0 to 1 criteria ($f(v) = 1$ iff$v \geq a$;& $f(v) = 0$ iff $v < a$; where $v = \Sigma w_i x_i$, and $a =$ threshold). The neuron is actuated in the event that it is above edge (typically 0) and such counterfeit neuron generate 1 (informative yield). Counterfeit neuron is not enacted on the off chance that it is beneath edge (typically 0) which is considered as -1. They are genuinely easy to keep up and are furnished with to manage information which contains a great deal of commotion. Significant for analysis as simple to design, fast, and also generate good responsiveness to noise. Only disadvantage is that it cannot be deployed in AI-processing tasks because of no deep stratifications and reverse tracking.

Multilayer-supervised model is advancement in Feed forward neural network. Each and Every single node is interconnected. Input and output layers are found in between of multiple hidden Layers (Heidari et al. 2020). It involves forward and

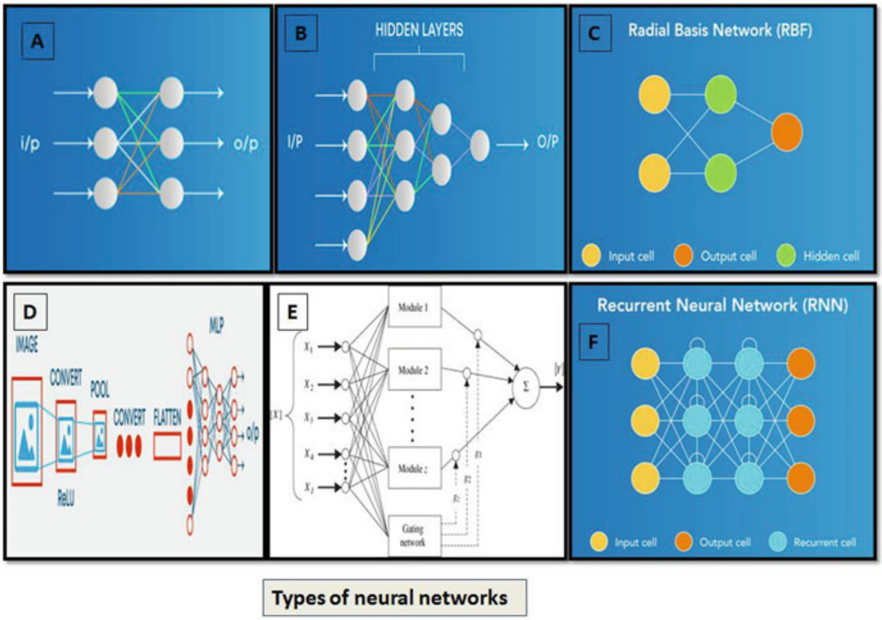


Fig. 18.3 Types of neural networks: (**a**) Feed forward network, (**b**) multi-perceptron network, (**c**) radial basis network, (**d**) convolution network, (**e**) modular network, (**f**) recurrent neural network

backward preparative tracking. Sources of info multiplied with loads and afterward exposed for activation function calculations along with reverse tracking; each informative node shows such alterations, so it can decrease the mislaying in information. Loads are machine taken in values from Neural Networks. Loads (Wi) are competent to self-modify contingent upon the contrast between anticipated yields versus preparing inputs. Nonlinear-initiation activator function is conveyed here, which makes them complex and best for deep learning tasks. Disadvantage includes comparatively slow functionality in huge data analysis.

Convolution neural network contains a 3D course of action of neurons, rather than the standard 2D arrangement. The principal layer is known as a convolution filter with activation mapping of counterfeit neurons. Every neuron in the convolution filter only processes the data from a little piece of the image related data (Chen et al. 2020). Information highlights are taken in clump astute like a channel. The system comprehends the pictures in parts and can figure these activities on various occasions to finish the full picture preparing. Handling includes change picture standards between RGB and dark scaling. Promoting adjustments for CRT screen dots worth assist with identifying corners, such pictures will be easily characterized. Engendering or tracks are following one-direction flow and convolution neural network holds at least one convolution filter accompanied by amalgamation of information and dual-directional when yield of convolution filters transfers information towards associated neural system for ordering the pictures as appeared in the

Fig. 18.3. Channels are utilized to remove certain pieces from picture. In Multilayer-supervised model the sources of info are duplicated with loads and subjected to the activation function. Convolution utilizes nonlinear enactment function followed by softmax. Convolution neural systems show viable outcomes in picture and video acknowledgment, semantic parsing and reword discovery. It is deployed for machine learning analysis. Disadvantage includes complexity in designing and slow functionality.

Radial Basis Function Network comprises an info vector followed by a layer of RBF neurons and a yield layer with one hub for each classification (Zaji et al. 2020). Characterization is performed by estimating the info's similitude to information focuses from the preparation set where every neuron stores a model. This will be one of the models from the preparation set. At the point when another info vector [the n-dimensional vector that you are attempting to classify] should be arranged, every neuron computes the Euclidean separation between the information and its model. Each RBF neuron looks at the info vector to its model and yields a worth running which is a proportion of similitude from 0 to 1. As the info equivalents to the model, the yield of that RBF neuron will be 1 and with the separation develops between the information and model the responses tumbles off exponentially towards 0. The plot created out of neuron's responses tends towards a typical the bell shaped plot. The yield layer comprises of a lot of neurons [one per category]. Its applications are found in power restoration.

Recurrent Neural framework fed back to info or data to provide assistance for anticipating results for each layer. Primary stratified division ordinarily show feed forward architecture accompanied intermittent counterfeit framework strata that holds data (past time-step), so recollected by storage assemblies acting as memory-units (Smyl 2020). Onward tracks executed for such situations. It holds the knowledge relevant for its potential use. On the off chance that the expectation is not right, the learning rate is utilized to roll out little improvements. Consequently, stepwise increment towards making the correct forecast during the back track. Its focal points are Model consecutive information where each example can be thought to be subject to verifiable ones, it is utilized with convolution layers to expand the pixel viability. Significant detriments of such network architecture is Gradient disappearing and detonating issues, preparing repetitive neural frameworks act as troublesome undertaking, hard to info-processing for successive information utilizing rectified-linear-units as initiating set. LSTM (Long short term memory) systems are a kind of RNN that utilizes exceptional units notwithstanding standard units. LSTM units incorporate a “memory cell” that can keep up data in memory for significant stretches of time. A lot of doors is utilized to control when data enters the memory when its yield, and when's it slipped it's mind. There are three types of gates, viz., Input gate (Info door), output gate (yield entryway), and forget gate (overlook entryway). Info door chooses what number of data from the last data set will be kept in memory; the yield entryway manages the measure of information went to the following layer, and overlook entryways control the tearing pace of memory put away. Such architecture lets them learn longer-term dependencies.

A modular neural system has various systems that work autonomously and perform sub-undertakings. The various systems do not generally collaborate with or signal each other during the calculation procedure (Li et al. 2020). They work autonomously towards accomplishing the yield. Therefore, an enormous and complex computational procedure is done essentially quicker by separating it into free segments. The calculation speed increments in light of the fact that the systems are not collaborating with each other but at last associated with one another. It is robust and efficient neural network, but sometimes has moving target problems. Commonly used by stock exchange market for predictions, and biological studies for compression of high level input data, and character recognition.

18.4 Application of Neural Networks

Many in silico tools, servers, and algorithms (Table 18.1) are currently used in both proteomic and genomic analysis. Structural and functional aspect of reacting biomolecules within cellular domains can be easily accessed by neural network algorithms. Neural networks have multiple applications in bioinformatics:

1. Protein and peptide structure prediction, including primary, secondary, and tertiary structures. All related estimations like biochemical properties including Ramachandran plot assessment. Stability investigations. Comparative or homology as well as ab-initio both type of model can easily predicted by deploying artificial neural networks.
2. In modern era fast protein–ligand interaction studies are conducted by using neural networks. Neural networks assists in determining binding pockets for ligand molecules, primarily in computer aided drug discovery.
3. Molecular docking and Molecular simulation studies are also based on neural networks architecture to give precise trajectories for interacting molecules (DNA–Protein as well as Protein–Protein).
4. Genome annotations and alignment of DNA or protein sequences, also uses neural architectures.
5. RNA-Seq or Whole genome Sequence analysis studies are also using neural networks in differential gene expression analysis.
6. In cancer studies, for prediction of pathogenicity of DNA variants.

18.4.1 Prediction of Structure for Proteins

Now a days, dual-direction recurrent neural architectures, PSI-BLAST-derived profiles, and enormous non-redundant guiding sets deployed in tools like PSIPRED (McGuffin et al. 2000) produces two new predictors: (a) SSpro program for secondary structure classification into three categories sheets, helix, and loops and

Table 18.1 List of various modern in silico tools/techniques based on deep learning or neural networks

Tools based on NN	Source	Function
Net MHC server	Lundegaard et al. (2011)	Epitopes selection and prediction from bacterial and viral proteins used in vaccine designing
NeuRiPP	De los Santos (2019)	Identification of genetic clusters to reveal ribosomally synthesized and post-translationally modified proteins
DeepGoPlus	Kulmanov and Hoehndorf (2020)	Protein function prediction
DEEPscreen	Rifaioglu et al. (2020)	Prediction of drug targets
RONN	Yang et al. (2005)	Identification of disordered regions of proteins
RESCUE	Pons and Delsuc (1999)	NMR spectral assignment to proteins
DeepQA	Cao et al. (2016)	Estimation of single protein model
DeepInteract	Patel et al. (2017)	Protein–protein interaction analysis
ProLanGO	Cao et al. (2017)	Protein functionality assessment
DeepDrug3D	Pu et al. (2019)	Drug or ligand binding pocket analysis and identification with in proteins or enzymes
EpiDock	Atanasova et al. (2013)	Molecular docking tool based on MHC class II interactions with epitopes
DeepLNC	Tripathi et al. (2016)	A long non coding RNA elements identification
DeepRibo	Clauwaert et al. (2019)	Gene annotation for prokaryotes based on ribosome profiling signals and binding site patterns
Afann	Tang et al. (2019)	Alignment free genetic sequence comparisons
SECLAF	Szalkai and Grolmusz (2018)	Biological sequence classification
SpliceFinder	Wang et al. (2019)	Prediction of splice sites using convolutional neural network architecture
DeepImpute	Arisdakessian et al. (2019)	Impute single cell RNA-seq data
DanQ	Quang and Xie (2016)	Quantification of DNA functions
RNAsamba	Camargo et al. (2020)	Assessment of translational potential of RNA sequence
REVEL	Ioannidis et al. (2016)	Prediction of pathogenicity of rare missense DNA variants. Assist in cancer biology

(b) SSpro8 program for secondary structure classification into the eight classes produced by the DSSP (dictionary of secondary structure of proteins) program, types include 3/10 helix, alpha helix, pi helix, extended strand in parallel and/or anti-parallel β-sheet conformation, isolated β-bridge, hydrogen bonded turn, bend, and coil. 8-state secondary structure is frequently amassed into 3-state auxiliary structure (Pollastri et al. 2002). Predicting protein structural disorders can be estimated by using feed forward neural networks (Li et al. 1999). Artificial neural networks are also used protein functional determination likely emulsification, and foaming for assisting food industry (Arteaga and Nakai 1993).

18.4.2 Binding Patterns and Epitope Selection: Immuno-Informatics Application

Binding or interaction between receptor and ligand molecules can be easily predicting by deploying neural networks, for example, K_{DEEP} is a fast machine learning tool that uses convolutional neural network architecture for protein to ligand binding (Jiménez et al. 2018). Molecular docking studies with known protein and ligand structure in Pdb format can assist in predicting interaction between their constituents, with proper binding scores, atomic contact energies and RMSD (root mean square deviation) values. Binding pockets within receptor protein is made up of reactive amino acids and ligand amino acids interact with it to exhibit perfect fitting. Even in drug discovery convolutional neural network architecture is mostly deployed to produce perfect results, for example, DeepDrug3D (Pu et al. 2019). Neural network architecture is also used in quality appraisal of protein and ligand interactional domains prediction, for example, FunFOLD-QA (Roche et al. 2012).

One of the studies utilized artificial Neural Network method for developing potential vaccine candidates against mumps virus. This involved a novel concept known as reverse vaccinology in which prediction of peptide epitopes was done which would potentially elicit an immune response in human body by B cells and T cells. Hemagglutinin-neuraminidase (HN) surface glycoproteins are the main antigenic structures present in mumps virus which served as the source of the candidate peptide epitopes. 593 HN glycoprotein sequences were retrieved from NCBI database. Then neural network was used to study binding of these peptide candidates to MHC class I molecules to determine the minimum inhibitory concentration (IC50). Percentile ranks of as low as 0.1 were obtained showing high binding affinity between the candidate epitope and the human MHC class I allele, indicating potential use of the epitope as a peptide vaccine against mumps virus (Babiker et al. 2020).

Prediction of continuous and linear B-cell epitopes and T-cell epitopes for antigens is basis for immunoinformatic analysis to craft rapid vaccination against pathogens (Fig. 18.4); recurrent neural architecture was successfully deployed in such studies (Saha and Raghava 2006). NetMHC server (Lundegaard et al. 2011), NetCTLpan server (Stranzl et al. 2010), and BepiPred (Jespersen et al. 2017) are some of the common tools that are mostly used in predicting epitopes. These servers are based on artificial neural networks and assist user in determining epitopes interacting with MHCI and II HLA alleles. After confirmation with molecular docking as well as molecular dynamic simulation trajectory analysis users can rapidly determine immunogenic properties of humans against deadly viruses like corona viruses (Joshi et al. 2020) and even in the rarest pathogenic bacterium, such as *Tropheryma whipplei* (Joshi and Kaushik 2020).

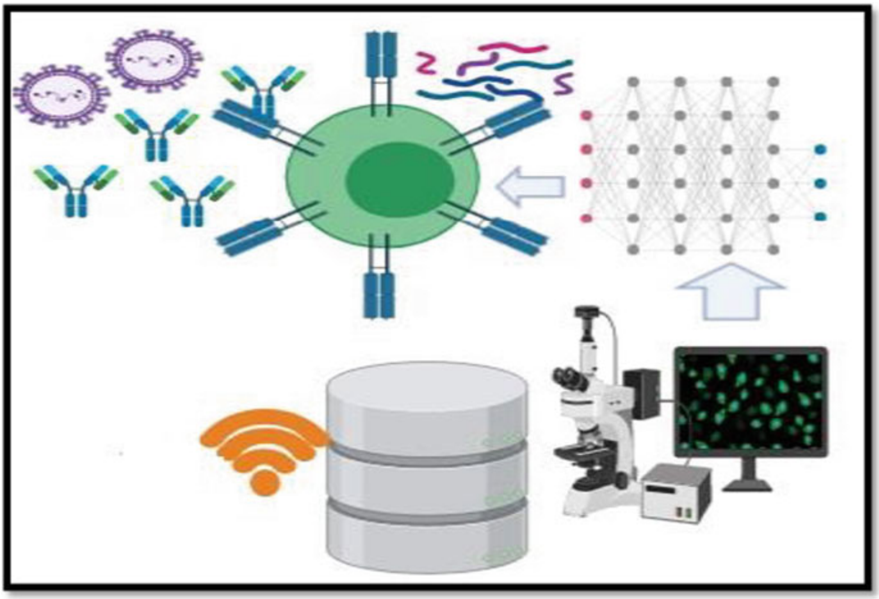


Fig. 18.4 Neural networks in epitope based vaccine crafting for viral pathogens

18.4.3 Role in Genomics and Transcriptomics

Eukaryotic and prokaryotic organisms have complex genomic expression, and to understand the mysteries behind it. Convolutional neural networks assisted users in predicting translational sites, regulatory mechanisms, and splicing domains in genetic elements (DNA or RNA) (Pedersen and Nielsen 1997). Sequential regulatory activity can be predicted across the chromosomes with convolutional neural architecture (Kelley et al. 2018). Deep neural network investigation also opens the gate of opportunities in for gene ontology and annotation (Chicco et al. 2014). Deep neural architecture plays crucial role in modeling RNA structures, and to conduct sequential alignment and comparisons (Wu and McLarty 2012). Modern sequencing studies need analysis of data generated for different organisms, to assist it deep neural networks play very crucial role. In WGS, and RNA-seq analysis, neural network tools like DanQ, RNAsamba tools were used along with Linux based freeware. Genome and transcriptome analysis was always data centric and to make better choices in selecting gene of interest to develop understanding about physiological or biochemical functionality was always primary feature that would lead scientific groups to triumph in the field of medicine.

18.5 Conclusion

Institutional computing facilities were improved lot in the past decade. Amalgamation of neural networks with advanced servers will assist rapid drug discovery, effective error free vaccine crafting, speedy alignments, structural predictions, and physiochemical analysis of biomolecules, etc. Modern world should not starve for better food security, medicinal treatments. To fulfill this broad socialistic view neural networks have intensified power to integrate, to access, and to analyze big data related to agriculture, animal husbandry, medicine, and physiology. Neural networks are constituent of deep learning domain of Artificial intelligence and machine learning; it holds significance in analyzing relationship about the integral features of IoT and bigdata (Mohammadi et al. 2018). Neural networks, as the name suggests it is the network or spider-web of interconnected artificial neurons joining input layer to output layer. Multiple types of neural networks assist users to develop insight about biomolecular structures and functions. Modern fast sequencing techniques generated enormous amount of data related to biological sequences, it was neural networks in bioinformatics who assisted researchers to bring fruitful outcomes in the field of agriculture as well as in medicine. It is ongoing research journey as neural networks are still evolving and linking to upgrading modern computing facilities to show its power of deep learning towards data analysis within the roots of big data and IoT.

References

Amato F, López A, Peña-Méndez EM, Vaňhara P, Hampl A, Havel J (2013) Artificial neural networks in medical diagnosis. J Appl Biomed 11(2):47–58

Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX (2019) DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome Biol 20 (1):1–14

Arteaga GE, Nakai S (1993) Predicting protein functionality with artificial neural networks: foaming and emulsifying properties. J Food Sci 58(5):1152–1156

Atanasova M, Patronov A, Dimitrov I, Flower DR, Doytchinova I (2013) EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. Protein Eng Des Sel 26(10):631–634

Babiker EAA, Almofti YA, Abd-Elrahman KA (2020) Novel T-lymphocytes vaccine candidates against human mumps virus via reverse vaccinology. Eur J Biomed 7(1):45–63

Bain A (1873) Mind and body: the theories of their relation, vol 4. Henry S. King, London

Camargo AP, Sourkov V, Pereira GAG, Carazzolle MF (2020) RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. NAR Genom Bioinform 2(1): lqz024

Cao R, Bhattacharya D, Hou J, Cheng J (2016) DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinform 17(1):495

Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z (2017) ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. Molecules 22(10):1732

Chen Y, Tang L, Yang X, Bilal M, Li Q (2020) Object-based multi-modal convolution neural networks for building extraction using panchromatic and multispectral imagery. Neurocomputing 386:136–146

Chicco D, Sadowski P, Baldi P (2014) Deep autoencoder neural networks for gene ontology annotation predictions. In Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics, pp. 533–540

Clauwaert J, Menschaert G, Waegeman W (2019) DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. Nucleic Acids Res 47(6):e36–e36

Collins FS, Morgan M, Patrinos A (2003) The human genome project: lessons from large-scale biology. Science 300(5617):286–290

de los Santos EL (2019) NeuRiPP: neural network identification of RiPP precursor peptides. Sci Rep 9(1):1–9

Evans RB (1990) William James, "the principles of psychology," and experimental psychology. Am J Psychol 103(4):433–447

Galushkin AI (2007) Neural networks theory. Springer, Berlin

Giorgini E, Biavasco F, Galeazzi R, Gioacchini G, Giovanetti E, Mobbili G et al (2020) Synthesis, structural insights and activity of different classes of biomolecules. In: *The First Outstanding 50 Years of "UniversitàPolitecnicadelle Marche"*. Springer, Cham, pp 463–482

Heidari AA, Faris H, Mirjalili S, Aljarah I, Mafarja M (2020) Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks. In: Nature-inspired optimizers. Springer, Cham, pp 23–46

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79(8):2554–2558

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S et al (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 99(4):877–885

Jespersen MC, Peters B, Nielsen M, Marcatili P (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res 45(W1):W24–W29

Jiménez J, Skalic M, Martinez-Rosell G, De Fabritiis G (2018) K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. J Chem Inf Model 58 (2):287–296

Joshi A, Joshi BC, Mannan MAU, Kaushik V (2020) Epitope based vaccine prediction for SARS-COV-2 by deploying immuno-informatics approach. Inform Med Unlocked 19:100338

Joshi A, Kaushik V (2020) In-Silico proteomic exploratory quest: crafting T-cell epitope vaccine against Whipple's disease. Int J Pept Res Ther 27:169–179

Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res 28 (5):739–750

Khan E (2020) Neural fuzzy based intelligent systems and applications. In: Fusion of neural networks, fuzzy systems and genetic algorithms. CRC Press, Boca Raton, FL, pp 105–140

Kulmanov M, Hoehndorf R (2020) DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36(2):422–429

Li W, Li M, Qiao J, Guo X (2020) A feature clustering-based adaptive modular neural network for nonlinear system modeling. ISA Trans 100:185–197

Li X, Romero P, Rani M, Dunker AK, Obradovic Z (1999) Predicting protein disorder for N-, C-and internal regions. Genome Inform 10:30–40

Lundegaard C, Lund O, Nielsen M (2011) Prediction of epitopes using neural network based methods. J Immunol Methods 374(1–2):26–34

McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16(4):404–405

Mohammadi M, Al-Fuqaha A, Sorour S, Guizani M (2018) Deep learning for IoT big data and streaming analytics: a survey. IEEE Commun Surv Tutorials 20(4):2923–2960

Morgat A, Lombardot T, Coudert E, Axelsen K, Neto TB, Gehant S et al (2020) Enzyme annotation in UniProtKB using Rhea. Bioinformatics 36(6):1896–1901

Patel S, Tripathi R, Kumari V, Varadwaj P (2017) DeepInteract: deep neural network based protein-protein interaction prediction tool. Curr Bioinform 12(6):551–557

Pedersen AG, Nielsen H (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. Proc Inst Conf Intell Syst Mol Biol 5:226–233

Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins Struct Funct Bioinform 47(2):228–235

Pons JL, Delsuc MA (1999) RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins. J Biomol NMR 15(1):15–26

Pu L, Govindaraj RG, Lemoine JM, Wu HC, Brylinski M (2019) DeepDrug3D: classification of ligand-binding pockets in proteins with a convolutional neural network. PLoS Comput Biol 15 (2):e1006718

Quang D, Xie X (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res 44(11):e107–e107

Rifaioglu AS, Nalbat E, Atalay V, Martin MJ, Cetin-Atalay R, Doğan T (2020) DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chem Sci 11(9):2531–2557

Roche DB, Buenavista MT, McGuffin LJ (2012) FunFOLDQA: a quality assessment tool for protein-ligand binding site residue predictions. PLoS One 7(5):e38219

Saha S, Raghava GPS (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins Struct Funct Bioinform 65(1):40–48

Shao C (2020) A quantum model of feed-forward neural networks with unitary learning algorithms. Quantum Inf Process 19(3):102

Smyl S (2020) A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. Int J Forecast 36(1):75–85

Stranzl T, Larsen MV, Lundegaard C, Nielsen M (2010) NetCTLpan: pan-specific MHC class I pathway epitope predictions. Immunogenetics 62(6):357–368

Szalkai B, Grolmusz V (2018) SECLAF: a webserver and deep neural network design tool for hierarchical biological sequence classification. Bioinformatics 34(14):2487–2489

Tang K, Ren J, Sun F (2019) Afann: bias adjustment for alignment-free sequence comparison based on sequencing data using neural network regression. Genome Biol 20(1):1–17

Tripathi R, Patel S, Kumari V, Chakraborty P, Varadwaj PK (2016) DeepLNC, a long non-coding RNA prediction tool using deep neural network. Network Model Anal Health Inform Bioinform 5(1):21

Wang R, Wang Z, Wang J, Li S (2019) SpliceFinder: ab initio prediction of splice sites using convolutional neural network. BMC Bioinform 20(23):652

Wu CH, McLarty JW (2012) Neural networks and genome informatics. Elsevier, Amsterdam

Wu Q, Peng Z, Anishchenko I, Cong Q, Baker D, Yang J (2020) Protein contact prediction using metagenome sequence data and residual neural networks. Bioinformatics 36(1):41–48

Yang ZR, Thomson R, Mcneil P, Esnouf RM (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21 (16):3369–3376

Zaji AH, Bonakdari H, Khameneh HZ, Khodashenas SR (2020) Application of optimized artificial and radial basis neural networks by using modified genetic algorithm on discharge coefficient prediction of modified labyrinth side weir with two and four cycles. Measurement 152:107291

Zeng M, Zhang F, Wu FX, Li Y, Wang J, Li M (2020) Protein–protein interaction site prediction through combining local and global features with deep neural networks. Bioinformatics 36 (4):1114–1120

Role of Bioinformatics in MicroRNA Analysis 19

Indra Mani

Abstract

Bioinformatics emerged as a new interdisciplinary science that provides an excellent way to understand biological sciences. It contains various biological databases such as nucleic acids, proteins, structures, pathways, interactions, etc. In addition, it also a comprehensive source of different softwares and tools. In silico approaches are very much helpful to do curation and annotation of various types of biological data. Due to advancement in DNA sequencing technology such as from Sanger methods to next generation sequencing (NGS), second generation and third generation played a significant role to generate enormous biological data. These biological data are being deposited in the various databases for further analysis and use. A small untranslated RNA molecule like microRNA (miRNA) plays a vital role in the regulation of the different biological processes. This chapter highlights different miRNA databases and mircoRNA prediction tools such as psRNATarget, RNAhybrid, miRscan, miRanda, TargetScan, PicTar, and Diana-MicroT, which are being utilized to mechanistically analysis of miRNA.

Keywords

miRNA · Tools · Databases · Bioinformatics · RNA

I. Mani (✉)
Department of Microbiology, Gargi College, University of Delhi, New Delhi, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_19

19.1 Introduction

As per recent updates, 38,589 entries of miRNAs have been mentioned in the miRBase database (http://www.mirbase.org/release#22.1, October 2018). It is a repository of annotated and published miRNA sequences. The microRNAs (miRNAs) are endogenously expressed small (~22 nucleotides) single-strand RNAs, which binds with the target mRNA and regulates gene expression and able to interfere post-transcriptionally with the protein production of their targets (Bartel 2004; Lewis et al. 2005; Alvarez-Garcia and Miska 2005). First miRNA was discovered in 1993 as lin-4, which suppressed the lin-14 gene expression in *Caenorhabditis elegans* (Lee et al. 1993), and second was identified in same organism as let-7 that suppressed expression of lin-41gene (Reinhart et al. 2000). These miRNAs have worked as same mechanism like both binds to the 3-′-untranslated region of lin-4 and lin 41 genes. Interestingly, more than 1000 miRNAs are encoded by the human genome that may cover approximately 60% of genes of mammalian and are ample in different human cell types (Bartel 2004; Bentwich et al. 2005; Friedman et al. 2009; Hennessy 2017; Narożna et al. 2017; Andrei et al. 2019; Hargreaves et al. 2020).

The typical structural characteristic of miRNAs is its initial transcriptional feature as a long primary transcript (pri-miRNA) that is modified into about 70 nucleotides precursor stem-loop hairpin RNAs (Lee et al. 2004). The pre-miRNAs are entered into a cytoplasm from the nucleus through a nuclear transport receptor exportin-5, where Dicer processes them into mature miRNA (about 22 nucleotides of miRNA each). After that it is included into a miRNA-containing RNA-induced silencing complex (miRISC) (Yi et al. 2003; Cullen 2004; Ambros 2004). It can be used to artificially induced to cleavage of the target, either altering the target or using miRNA sequences that can hybridize to the target (Zeng et al. 2002; Boden et al. 2004). A translational repression or mRNA cleavage is a method, which used by mature miRNA to regulate the gene expression. The use of miRNAs in numerous core cellular pathways as well as in many human diseases further endorses their biological significance (Bartel 2004; Alvarez-Garcia and Miska 2005; Mani et al. 2016).

Due to availability of enormous studies about miRNA, it seems to be it is being involved in numerous pathway (Kontaraki et al. 2014; Huang et al. 2012; Zhu et al. 2011). However, its imbalance could be caused the defective cell functions, and disease occurs. In addition, it also involves in regulation of the various biological processes (Fu et al. 2013; Tüfekci et al. 2014; Vishnoi and Rani 2017; Correia de Sousa et al. 2019). It has been suggested that miRNA could be used in prognosis, diagnosis, and therapeutic (Mani and Vasdev 2018). Remarkably, due to a rapid increase of biological data in the form of sequence, structure, pathway, and interactions, biological sciences have developed data-rich science. There are various databases and tools which are being available to retrieve and analysis of miRNA from different organisms. Following database and tools are discussed in detail.

19.2 microRNA Database

19.2.1 miRBase

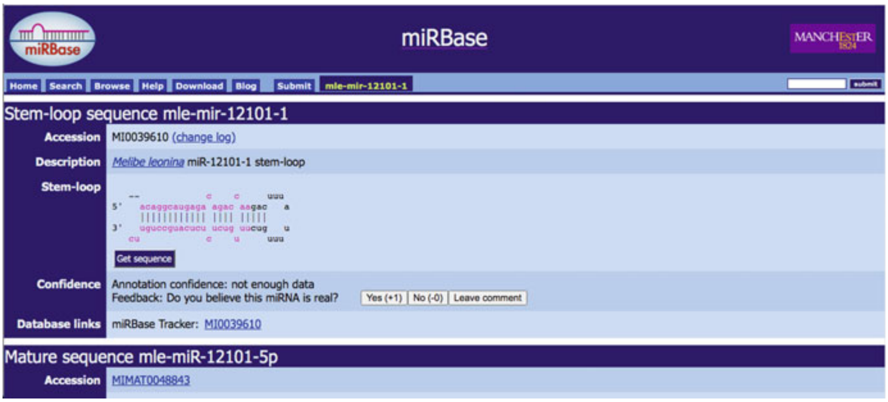
The miRBase database (http://www.mirbase.org/index.shtml) is a collection of annotated and available miRNA sequences. It is a publically online available and searchable database. In this database microRNA has presented in two ways such as miR (predicted hairpin portion of a miRNA transcript) and miR (Sequence and site of the mature miRNA sequence). Using a miRNA database, we can search a particular miRNA with a name and also download the sequence and annotated data (Griffiths-Jones et al. 2006, 2008; Kozomara and Griffiths-Jones 2011, 2014; Kozomara et al. 2019). A searching and browsing of mir-121 using a miRBase database are given in Fig. 19.1.

19.3 Tools for miRNA Target Prediction

There are various tools available for the prediction of miRNA. It is described in detail.

19.3.1 psRNATarget

The psRNATarget prediction server (http://plantgrn.noble.org/psRNATarget/) was developed for the analysis of plant regulatory small RNAs (sRNAs), including microRNAs (miRNAs) and small interfering RNAs (siRNAs) (Dai and Zhao 2011; Dai et al. 2018). The siRNA is generated from double-stranded duplexes of plant regulatory sRNAs while miRNAs generated from the stem-loops of single



Stem-loop sequence mle-mir-12101-1

Accession MI0039610 (change log)

Description *Melibe leonina* miR-12101-1 stem-loop

Stem-loop

```
   --            c    c       uuu
5'   acaggcaugaga agac aagac     a
     |||||||||||| |||| |||||
3'   uguccguacucu ucug uucug     u
   cu            c    u       uuu
```

Confidence Annotation confidence: not enough data
Feedback: Do you believe this miRNA is real?

Database links miRBase Tracker: MI0039610

Mature sequence mle-miR-12101-5p

Accession MIMAT0048843

Fig. 19.1 A miRBase database showing the browsing of mir-121. Users can use the combined search and retrieve the mRNA details. http://www.mirbase.org/index.shtml

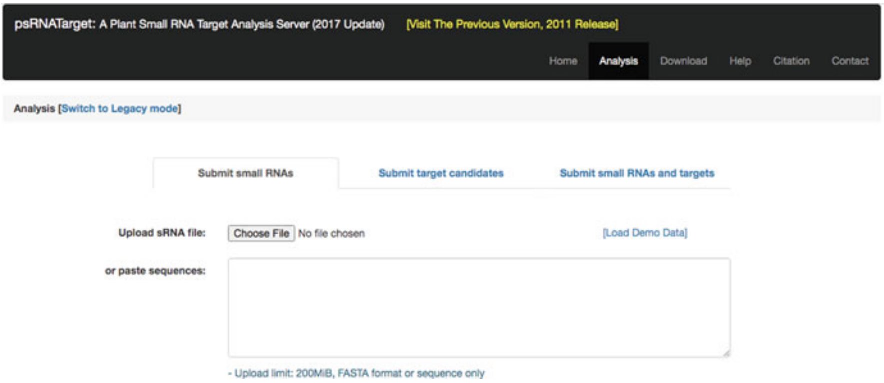
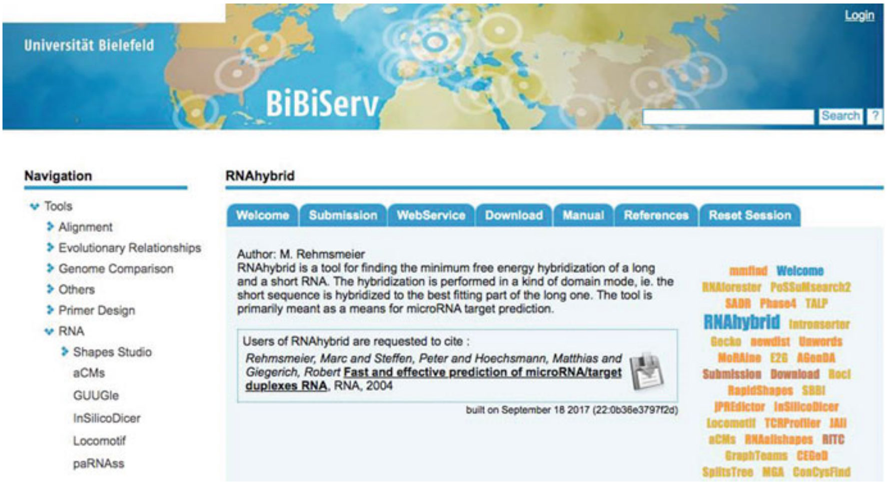

psRNATarget: A Plant Small RNA Target Analysis Server (2017 Update)
[Visit The Previous Version, 2011 Release]
Home
Analysis
Download
Help
Citation
Contact
Analysis [Switch to Legacy mode]
Submit small RNAs
Submit target candidates
Submit small RNAs and targets
Upload sRNA file:
Choose File
No file chosen
[Load Demo Data]
or paste sequences:
- Upload limit: 200MiB, FASTA format or sequence only

Fig. 19.2 Database home page of psRNATarget prediction server for miRNA prediction. http://plantgrn.noble.org/psRNATarget/

stranded precursors of plant sRNAs (Axtell 2013). The psRNATarget includes plant sRNA targets using evaluating complementary equivalent between target mRNA sequence and the sRNA sequence by scoring scheme and considering target site availability. A scoring procedure is adjustable and based on canonical and non-canonical targets (Dai et al. 2018). The psRNATarget has been used to predict 20,815 unigene targets (Ye et al. 2019). In addition, psRNATarget has been utilized for analysis of miRNAs from various plants such as *Hordeum vulgare* (Lv et al. 2012), *Coffea canephora* (Loss-Morais et al. 2014), Tomato (Luan et al. 2014), *Beta vulgaris* (Li et al. 2015), *Chlamydomonas reinhardtii* (Hajieghrari et al. 2016), *Brassica rapa* (Hajieghrari et al. 2017; Zhou et al. 2020), *Arachis hypogaea* (Rajendiran et al. 2019), *Oryza sativa* (Jabbar et al. 2019), and *Passiflora edulis* (Paul et al. 2020). This server (Fig. 19.2) is being utilized to retrieve and analysis of plant miRNAs and siRNAs.

19.3.2 RNA Hybrid

The RNAhybrid is an online tool (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid) specific for miRNA target prediction in plants and mammalians, respectively. It works on the basis of minimum free energy hybridization of a pair of short and long RNA sequences (Rehmsmeier et al. 2004; Krüger and Rehmsmeier 2006; Xia et al. 2009). In addition, RNAhybrid has been utilized for analysis of miRNAs in *Caenorhabditis elegans* (Krüger and Rehmsmeier 2006), *Panax ginseng* (Wang et al. 2019), in the human hepatocellular carcinoma cells (Li et al. 2019), in the heart failure (Fan et al. 2018), in the malignant tumors of the human central nervous system (Sun et al. 2018), in the Burkitt lymphoma (Li et al. 2017), in the lipid metabolism (Vijayaraghavan et al. 2018), in peripheral blood mononuclear cells (PBMCs) infected with EV71 and CA16 (Song et al. 2018), in the *Anopheles sinensis* (Feng et al. 2018), in the mouse during postnatal ovarian development



Author: M. Rehmsmeier
RNAhybrid is a tool for finding the minimum free energy hybridization of a long and a short RNA. The hybridization is performed in a kind of domain mode, ie. the short sequence is hybridized to the best fitting part of the long one. The tool is primarily meant as a means for microRNA target prediction.

Users of RNAhybrid are requested to cite :
Rehmsmeier, Marc and Steffen, Peter and Hoechsmann, Matthias and Giegerich, Robert **Fast and effective prediction of microRNA/target duplexes RNA**, RNA, 2004

built on September 18 2017 (22:0b36e3797f2d)

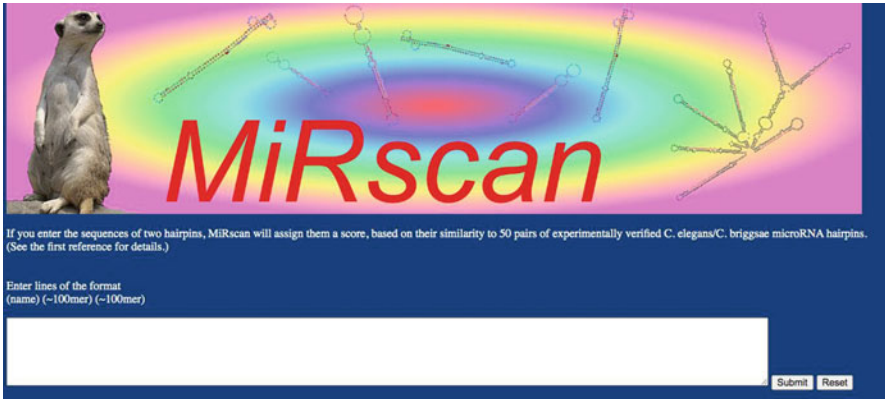
Fig. 19.3 Database home page of RNAhybrid prediction server for miRNA prediction http://bibiserv.techfak.uni-bielefeld.de/rnahybrid

and superovulation (Khan et al. 2015), in the invasion and metastasis of colorectal carcinoma (Yang et al. 2015), and in colon cancer (Xiong et al. 2019). This server (Fig. 19.3) is being utilized to retrieve and analysis of miRNAs and calculate the free energy rapidly and accurately.

19.3.3 MiR Scan

The miRscan is web-available tool (http://hollywood.mit.edu/mirscan/index.html) use for the prediction of specifically miRNAs in hairpins, which are conserved in the two genomes and have the characteristics of identified miRNAs (Lim et al. 2003a). It is assigned a score based on similarity with 50 pairs of microRNA hairpins of *Caenorhbditis elegans/C. briggsae.* Lim et al. (2003b) has used miRscan to analyze the miRNA genes in the model organism *C. elegans.* A study was based on miRscan along with other molecular approaches, which has identified and validated 88 miRNA genes (Lim et al. 2003b). The miRscan has also been used to predict the miRNA in *Drosophila* (Li et al. 2003), human (Lim et al. 2003a), human cytomegalovirus (Grey et al. 2005), and in mouse testis tissues (Yan et al. 2007). This web based server (Fig. 19.4) is being utilized to retrieve and analysis of miRNAs.

In addition, there are various tools that could be used for the prediction of miRNA, such as miRanda (http://www.microrna.org/miranda_new.html) first bioinformatics software, which predicts the miRNA in *Drosophila melanogaster* (Enright et al. 2003), TargetScan (http://www.targetscan.org) tool used for the prediction of miRNA from the mammalian (Lewis et al. 2003), PicTar (http://pictar.bio.nyu.edu) utilized more complex algorithm to predict miRNA in fruit flies, nematodes, and



If you enter the sequences of two hairpins, MiRscan will assign them a score, based on their similarity to 50 pairs of experimentally verified C. elegans/C. briggsae microRNA hairpins. (See the first reference for details.)

Enter lines of the format
(name) (~100mer) (~100mer)

Fig. 19.4 Database home page of miRscan prediction server for miRNA prediction http://hollywood.mit.edu/mirscan/index.html

vertebrates (Krek et al. 2005), and Diana-MicroT program (http://www.diana.pcbi.upenn.edu/cgi-bin/micro_t.cgi) used for the prediction of animal miRNA (Kiriakidou et al. 2004). Furthermore, other tools are also available to prediction of a microRNA from the different organisms.

19.4 Concluding Remarks

The function of miRNA is well established in the regulation of gene expression throughout the posttranscriptional repression. Their upregulation and downregulation are a good molecular marker for prognosis and diagnosis of diseases. In addition, it could be a potential target for various disease treatments. Recently, analysis of miRNA has significantly received consideration, and identification of miRNA is being with combined approaches such as bioinformatics predictions and experimental assays. Moreover, to understand their dynamics in the organism, in silico approach is a very promising tool. Presently, there are numerous in silico prediction tools which are available to analyze their structural property. Further, bioinformatics tools and databases are being helped to increase our understanding about the structure and function of miRNA in different organisms. Moreover, the development of rapid, easy, and high-throughput experimental identification assays would be advantageous to support the bioinformatics predictions of miRNA.

Competing Interests There is no competing interest.

References

Alvarez-Garcia I, Miska EA (2005) MicroRNA functions in animal development and human disease. Development 132(21):4653–4662

Ambros V (2004) The functions of animal microRNAs. Nature 431:350–355

Andrei D, Nagy RA, van Montfoort A, Tietge U, Terpstra M, Kok K, van den Berg A, Hoek A, Kluiver J, Donker R (2019) Differential miRNA expression profiles in cumulus and mural Granulosa cells from human preovulatory follicles. Microrna 8(1):61–67

Axtell MJ (2013) Classification and comparison of small RNAs from plants. Annu Rev Plant Biol 64:137–159

Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116 (2):281–297

Bentwich I, Avniel A, Karov Y et al (2005) Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 37(7):766–770

Boden D, Pusch O, Silbermann R, Lee F, Tucker L, Ramratnam B (2004) Enhanced gene silencing of HIV-1 specific siRNA using microRNA designed hairpins. Nucleic Acids Res 32:1154–1158

Correia de Sousa M, Gjorgjieva M, Dolicka D, Sobolewski C, Foti M (2019) Deciphering miRNAs' action through miRNA editing. Int J Mol Sci 20(24):6249

Cullen BR (2004) Transcription and processing of human microRNA precursors. Mol Cell 16:861–865

Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res 39:W155–W159

Dai X, Zhuang Z, Zhao PX (2018) psRNATarget: a plant small RNA target analysis server (2017 release). Nucleic Acids Res 46:W49–W54

Enright AJ, John B, Gaul U et al (2003) microRNA targets in Drosophila. Genome Biol 5:R1

Fan J, Li H, Nie X, Yin Z, Zhao Y, Zhang X, Yuan S, Li Y, Chen C, Wang DW (2018) MiR-665 aggravates heart failure via suppressing CD34-mediated coronary microvessel angiogenesis. Aging 10(9):2459–2479

Feng X, Wu J, Zhou S, Wang J, Hu W (2018) Characterization and potential role of microRNA in the Chinese dominant malaria mosquito *Anopheles sinensis* (Diptera: Culicidae) throughout four different life stages. Cell Biosci 12(8):29

Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19(1):92–105

Fu G, Brkic J, Hayder H, Peng C (2013) MicroRNAs in human placental development and pregnancy complications. Int J Mol Sci 14:5519–5544

Grey F, Antoniewicz A, Allen E, Saugstad J, McShea A, Carrington JC, Nelson J (2005) Identification and characterization of human cytomegalovirus-encoded microRNAs. J Virol 79 (18):12095–12099

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34: D140-D144

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36: D154-D158

Hajieghrari B, Farrokhi N, Goliaei B, Kavousi K (2016) Identification and characterization of novel miRNAs in *Chlamydomonas reinhardtii* by computational methods. Microrna 5(1):66–77

Hajieghrari B, Farrokhi N, Goliaei B, Kavousi K (2017) Computational identification of microRNAs and their transcript target(s) in field mustard (*Brassica rapa* L.). Iran J Biotechnol 15(1):22–32

Hargreaves BKV, Roberts SE, Derfalvi B, Boudreau JE (2020) Highly efficient serum-free manipulation of miRNA in human NK cells without loss of viability or phenotypic alterations is accomplished with TransIT-TKO. PLoS One 15(4):e0231664

Hennessy E (2017) MiRNA profiling in human induced pluripotent stem cells. Methods Mol Biol 1509:47–56

Huang W, Feng Y, Liang J (2012) MicroRNA-128 regulates Isl-1 via Nkx-2.5/gsh-2 competition during cardiac development. Circulation 126:A11394

Jabbar B, Iqbal MS, Batcho AA, Nasir IA, Rashid B, Husnain T, Henry RJ (2019) Target prediction of candidate miRNAs from Oryza sativa for silencing the RYMV genome. Comput Biol Chem 83:107127

Khan HA, Zhao Y, Wang L, Li Q, Du YA, Dan Y, Huo LJ (2015) Identification of miRNAs during mouse postnatal ovarian development and superovulation. J Ovarian Res 8:44

Kiriakidou M, Nelson PT, Kouranov A et al (2004) A combined computational experimental approach predicts human microRNA targets. Genes Dev 18:1165–1178

Kontaraki JE, Marketou ME, Zacharis EA, Parthenakisand FI, Vardas PE (2014) Differential expression of vascular smooth muscle-modulating microRNAs in human peripheral blood mononuclear cells: novel targets in essential hypertension. J Hum Hypertens 28:510–516

Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. Nucleic Acids Res 47:D155–D162

Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39:D152–D157

Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42:D68–D73

Krek A, Grun D, Poy MN et al (2005) Combinatorial microRNA target predictions. Nat Genet 37:495–500

Kruger J, Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res 34:W451–W454

Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23:4051–4060

Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120:15–20

Lewis BP, Shih IH, Jones-Rhoades MW et al (2003) Prediction of mammalian microRNA targets. Cell 115:787–798

Li JG, Ding Y, Huang YM, Chen WL, Pan LL, Li Y, Chen XL, Chen Y, Wang SY, Wu XN (2017) FAMLF is a target of miR-181b in Burkitt lymphoma. Braz J Med Biol Res 50(6):e5661

Li JL, Cui J, Cheng DY (2015) Computational identification and characterization of conserved miRNAs and their target genes in beet (Beta vulgaris). Genet Mol Res 14(3):9103–9108

Li W, Dong X, He C, Tan G, Li Z, Zhai B, Feng J, Jiang X, Liu C, Jiang H, Sun X (2019) LncRNA SNHG1 contributes to sorafenib resistance by activating the Akt pathway and is positively regulated by miR-21 in hepatocellular carcinoma cells. J Exp Clin Cancer Res 38(1):183

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP (2003a) Vertebrate microRNA genes. Science 299(5612):1540

Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP (2003b) The microRNAs of *Caenorhabditis elegans*. Genes Dev 17(8):991–1008

Loss-Morais G, Ferreira DC, Margis R, Alves-Ferreira M, Corrêa RL (2014) Identification of novel and conserved microRNAs in Coffea canephora and Coffea arabica. Genet Mol Biol 37 (4):671–682

Luan Y, Wang W, Liu P (2014) Identification and functional analysis of novel and conserved microRNAs in tomato. Mol Biol Rep 41(8):5385–5394

Lv S, Nie X, Wang L, Du X, Biradar SS, Jia X, Weining S (2012) Identification and characterization of microRNAs from barley (*Hordeum vulgare* L.) by high-throughput sequencing. Int J Mol Sci 13(3):2973–2984

Mani I, Garg R, Pandey KN (2016) Role of FQQI motif in the internalization, trafficking, and signaling of guanylyl-cyclase/natriuretic peptide receptor-a in cultured murine mesangial cells. Am J Physiol Renal Physiol 310(1):F68–F84

Mani I, Vasdev K (2018) MicroRNA in prognosis, diagnosis and therapy of Cancer. Cell Cell Life Sci J 3:000134

Narożna B, Langwinski W, Jackson C, Lackie P, Holloway JW, Szczepankiewicz A (2017) MicroRNA-328 is involved in wound repair process in human bronchial epithelial cells. Respir Physiol Neurobiol 242:59–65

Paul S, de la Fuente-Jiménez JL, Manriquez CG, Sharma A (2020) Identification, characterization and expression analysis of passion fruit (*Passiflora edulis*) microRNAs. 3 Biotech 10(1):25

Rajendiran A, Vijayakumar S, Pan A (2019) Exploring microRNAs, target mRNAs and their functions in leguminous plant *Arachis hypogaea*. Microrna 8(2):135–146

Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10(10):1507–1517

Reinhart BJ, Slack FJ, Basson M et al (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. Nature 403:901–906

Song J, Jiang X, Hu Y, Li H, Zhang X, Xu J, Li W, Zheng X, Dong S (2018) High-throughput sequencing of putative novel microRNAs in rhesus monkey peripheral blood mononuclear cells following EV71 and CA16 infection. Intervirology 61(3):133–142

Sun G, Wang Y, Zhang J, Lin N, You Y (2018) MiR-15b/HOTAIR/p53 form a regulatory loop that affects the growth of glioma cells. J Cell Biochem 119(6):4540–4547

Tüfekci KU, Meuwissen RL, Genç S (2014) The role of microRNAs in biological processes. Methods Mol Biol 1107:15–31

Vijayaraghavan B, Danabal K, Padmanabhan G, Ramanathan K (2018) Study on regulation of low density lipoprotein cholesterol metabolism using PCSK9 gene silencing: a computational approach. Bioinformation 14(5):248–251

Vishnoi A, Rani S (2017) MiRNA biogenesis and regulation of diseases: an overview. Methods Mol Biol 1509:1–10

Wang Y, Peng M, Chen Y, Wang W, He Z, Yang Z, Lin Z, Gong M, Yin Y, Zeng Y (2019) Analysis of *Panax ginseng* miRNAs and their target prediction based on high-throughput sequencing. Planta Med 85(14–15):1168–1176

Xia W, Cao G, Shao N (2009) Progress in miRNA target prediction and identification. Sci China C Life Sci 52(12):1123–1130

Xiong W, Wang X, Cai X, Xiong W, Liu Y, Li C, Liu Q, Qin J, Li Y (2019) Identification of tRNA-derived fragments in colon cancer by comprehensive small RNA sequencing. Oncol Rep 42 (2):735–744

Yan N, Lu Y, Sun H, Tao D, Zhang S, Liu W, Ma Y (2007) A microarray for microRNA profiling in mouse testis tissues. Reproduction 134(1):73–79

Yang B, Tan Z, Song Y (2015) Study on the molecular regulatory mechanism of MicroRNA-195 in the invasion and metastasis of colorectal carcinoma. Int J Clin Exp Med 8(3):3793–3800

Ye J, Han W, Fan R, Liu M, Li L, Jia X (2019) Integration of transcriptomes, small RNAs, and degradome sequencing to identify putative miRNAs and their targets related to Eu-rubber biosynthesis in *Eucommia ulmoides*. Genes 10(8):623

Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev 17:3011–3016

Zeng Y, Wagner EJ, Cullen BR (2002) Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol Cell 9:1327–1333

Zhou B, Leng J, Ma Y, Fan P, Li Y, Yan H, Xu Q (2020) BrmiR828 targets *BrPAP1*, *BrMYB82*, and *BrTAS4* involved in the light induced anthocyanin biosynthetic pathway in *Brassica rapa*. Int J Mol Sci Jun 21(12):4326

Zhu H, Yang Y, Wang Y, Li J, Schiller PW, Peng T (2011) MicroRNA-195 promotes palmitate-induced apoptosis in cardiomyocytes by down-regulating Sirt1. Cardiovasc Res 92(1):75–84

Bioinformatics for Image Processing 20

Dilip Kumar J. Saini and Dhirendra Siddharth

Abstract

With rising applications in engineering and science, digital image processing is a rapidly evolving sector. Modern digital technology has allowed multidimensional signal manipulation. Digital image processing has a wide range of applications including medical image processing, remote satellite sensing data, acoustic image processing, sonar, radar, and automation. Imaging has become important in fields of clinical practice and medical and laboratory science. Biologists research cells and produce data sets using three-dimensional optical microscope, three-dimensional image visualisation and quantitative analysis could only be carried out with expensive UNIX workstations and custom tools. Today, much of the simulation and analysis can be performed on an inexpensive desktop computer with the necessary hardware and software for the graphics. In these data-intensive problems, the introduction of new image analysis, database, data mining, and simulation strategies to record, evaluate, scan, and manage biological information has been increasingly focused. This recent emerging field of bioinformatics is being referred to as 'bioimage computing'. This chapter discusses the developments made in this field from various perspectives including implementations, main methods, tools, and resources available. The requisite strategies for success in the battle against COVID-19, such as identification of bioimage characteristics, monitoring and segmentation, visualisation, mining, registration, management of image data and annotation, along with a brief description of accessible analytical resources, bioimage databases, and other facilities, are also outlined.

D. K. J. Saini (✉) · D. Siddharth
Department of Computer Science and Engineering, RAMA University, Kanpur, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_20

Keywords

Bioinformatics · Image processing · Image processing recent tools · COVID-19

20.1 Introduction

The deluge of complex biomedical and biological images presents huge obstacles for the image processing community. As a natural extension of the existing biomedical field of image analysis, an emerging modern engineering area is to develop and optimise various image data processing and informatics strategies to handle, capture, scan and compare the biological information of the respective images. This latest field can be known as bioimage informatics. Although due to the quality of information and the high complexity of bioimages, such as the very high cell density (e.g. microglia, neurons, astrocytes), the mechanism of entangled or very fast microtubular growth in a 4-dimensional live cell film (Mathews and Jezzard 2004) makes it extremely difficult to specifically apply current medical imaging techniques to this bioimage computer problem. Multiple colour channels are a single biological image stack and are usually wide. The artefacts of interest like this in an image, such as in the 3D structures in neurons (Wiemer et al. 2003), can have drastic differences in strength and morphology from image to image. It is also not unusual that millions of photographs have to be handled automatically in a high-throughput manner in terms of the number of hours or even days, but not months and even years of manual labour. Both of these difficulties require the emergence of new systems and algorithms for bioimage informatics, mainly from three factors: mining and image processing, visualisation and image database.

There are many researches in bioimage informatics either ongoing or on the past few decades. To address the latest trends in this area, a series of very fruitful workshops have been planned. The aim of this section is to briefly examine the advancement of bioimage informatics from the primary methods, angles of implementation, resource availability and instruments.

The practise of making graphic pictures of the internal systems of body part for medical treatment and diagnostic, along with a direct picture of the role of the internal tissue, is diagnostic imaging. This approach pursues the condition's diagnosis and recovery. This method generates a catalogue of the normal configuration and operation of the organs to make the abnormalities easier to recognise. This technique covers both radiological and organic imaging using scopes, magnetic, thermal imaging, isotope and sonography and electromagnetic energies (X-rays). In order to record details about the function and location of the body, several other devices are used. Compared to those modules that generate images, there are several drawbacks to those techniques. For different diagnostic purposes, billions of images are made annually worldwide. Digital images regularly serve an integral role. The analysis in medical imaging relates to image handling using the machine. In this process involved several processes, including communication, presentation, storage and image retrieval. The image is a feature that means measuring features such as the colour of a visible sight or illumination. There are many benefits of digital images, such as adaptable handling, easy and economical reproduction, immediate quality appraisal, multiple copying with quality reservation, quick storage and communication, and faster and cheaper processing costs. The drawbacks of digital images include the need for quicker processor manipulation, the need for large-capacity memory, the failure to resize with consistency retention and exploitation of copyright.

The use of computers to manipulate digital images is a technique for image processing. This technique has many benefits including connectivity, data management, adaptability and elasticity. This methodology has many sets of synchronous performance instructions for photographs. Multidimensional manipulation of the 2D and 3D Tranform images for different areas such as TV images, humanities, therapeutic applications and environmental enhancement, digital imaging methods have been used. In the time, the editing of photographs became cheap, simple and quicker.

20.2 Medical Imaging Techniques

Medical imaging techniques (MIT) consider the laboratory tests like blood test and specimen tests, it is one of the most popular medical tests. Over the past decade, medical imaging has undergone a revolution with rapid, more accurate and less invasive devices. Medical image techniques can be seen as instruments to learn more about neurobiology and people’s behaviours (Wang et al. 2016). In medical image techniques, the energy source that enter in the body through sensor and detector detects the body part, after that algorithm works on this data which is given by the detector and then displays result.

There are some different tools that can be used to see into the patient, depending on the energy sources. This may be an image of the interior of the patient by sensing the pulse emanating from the body. In this article, important techniques include magnetic resonance imaging, computed tomography, computed thermography and tomography single photon release, optical imaging, radiography X-ray, radionuclide imaging, positron emission tomography (PET), ultrasonography and elastography. Worldwide, 5 billion medical imaging tests were carried out by 2010 (Roobottom et al. 2010).

20.2.1 X-Ray Radiography

It is a medical tool which uses electromagnet ionising radiation, for example, X-rays, to examine body part. With wavelength of about 0.01 and 10 manometers, X-ray is a high-energy radioactive radiation ionising gas capable of penetrating solids. X-rays travel through the body for medical imaging, forming a profile, absorbing or attenuating them at various amounts, depending on the atomic number of the different density and tissues (Spahn 2013). In the X-ray, information is recorded on a sensor that is generated by X-ray. A current is applied through the cathode filament, which heats up and releases electrons via thermionic emission. The electrons were drawn by a spinning metallic anode that supplied the filament wire with an alternating current. The focal point is called as the anode area wherein the X-ray is emitted. The photon energies used vary around 17–150 KeV and a trade-off between the appropriate radiation dosage and the contrast picture possible is the preference for a particular application or tissue.

20.2.1.1 X-Ray Radiography Advantage

1. Non-invasive, painless and quick.
2. Support the planning of medical and surgical treatments.
3. When medical teams treat tumours by injecting catheters into the body.

20.2.1.2 Risks from X-Ray Radiography

1. Ionising radiation raises the risk of cancer in the future.
2. Tissue effects such as cataracts, skin reddening, and hair loss, which occur at relatively high levels of radiation exposure.

20.2.1.3 Applications of X-Ray Radiography

1. Can be used in a number of examinations, including dental, chiropractic, etc.
2. Can be used to demonstrate the activity of organs, such as you can even examine the brain vessels, heart and blood using the colon, stomach and intestine in the body.
3. Projection X-rays, determine the type of a fracture and extent of a fracture, including used to track and imagine physiological changes in the lungs gastrointestinal and intestinal function.
4. Mammogram used for breast tissue diagnosis and screening.
5. Bone densitometry used to measure the mineral content and density of bones.
6. Arthrography that was used to see inside the joint.
7. Hysterosalpingogram used for uterine and fallopian tube examination.

20.2.2 Computed Tomography (CT)

This is a medical method and it combines a computer and cathode ray tube display with X-ray equipment. In this method, it produces images of the part of human body. This method, X-ray film is changed by a sensor that evaluates the X-ray data. There is a spinning frame within the CT scanner with a detector positioned on one side and the X-ray tube on the other side (Xu and Tsui 2014). An X-ray beam is generated as a rotating frame spins the X-ray tube and detector around the body. Any time the detector and X-ray tube perform one full rotation, an image or slice is obtained. The profile is repeated with a two-dimensional version of the slice viewed by the computer. 3-dimensional computed tomography can be acquired by spiral computed tomography (Xu and Tsui 2014) patient anatomy data volume all at one spot. This data collection of volumes will then be reproduced to include three-dimensional representations of complex form. The subsequent 3-dimensional computed tomography images aid in the representation of tumour data in three dimensions. Recently four-dimensional computed tomography has developed in order to solve the difficulties. Four-dimensional computed tomography contains both temporal and spatial data regarding the activity of organs.

20.2.2.1 Computed Tomography Advantages

1. Comprehensive view of veins.
2. Painless, fast and non-invasive
3. Distinguished by slight physical density differences.
4. Should not invasively insert an arterial catheter and a guidewire.
5. Strong resolution of the spaces.

20.2.2.2 Computed Tomography Risks

1. Increases the risk that cancer can grow later in life.
2. The knowledge is not in real time.
3. Cannot spot anomalies in the luminaire.
4. Non-contract-free results (toxicity, allergy).

20.2.2.3 Computed Tomography Medical applications

1. Analysis of certain body parts, for example: wrist, head, elbow, knee, arm, hip, dental, leg, kidney, sinus, neck, elbow, spines.
2. Diagnosis of sickness, trauma and disorder.
3. Planning and instruction of clinical or interventional treatments.
4. Tracking therapy success (treatment for cancer).

20.2.3 Magnetic Resonance Imaging (MRI)

It is a health diagnostic technique for imagery body tissue and monitoring body chemistry using radio and magnetic frequency fields (Caiani et al. 2006). The MRI used to image morphological modifications is based on its ability to detect magnetic spin relaxation times and proton density variations that are typical of the tissue provided in the environment. When we talk about MR scanner, there are mainly three components: a radio frequency system, gradient, central magnet and magnetic field system. The principal magnet that produces a magnetic field is a permanent magnet. In the magnetic field gradient device, there are three orthogonal gradient coils. This coil uses for the signal localisation. The RF system has a transmitter coil capable; it is used for exciting a spin device, producing a spinning magnetic field, and a receiver coil capable of translating magnetisation processing into electrical signals. The MR scanner rebuilds the optical device and measures the signals and signals convert into images. Recently, a novel technique designed for measuring brain movement is called functional magnetic resonance imaging (FMRI) (Ng et al. 2009).

20.2.3.1 MRI Advantages

1. Non-invasive and painless.
2. Without the radiation being ionised.
3. High resolution over space.
4. Personal operator.
5. Easy to blind and flow and speed measuring capability with specialised technique.
6. Should be treated without comparison (allergy to pregnancy).
7. Great contrasting soft tissue.

20.2.3.2 MRI Risks

1. Sensitivity fairly low.
2. Long scanning time and post-processing time.
3. Mass sample volumes may be needed.
4. No details in real time.
5. The intraluminal defects cannot be observed.
6. Could make you feel claustrophobic to others.
7. Young children who cannot sit still will need sedation.
8. Pretty expensive.

20.2.3.3 MRI Medical Applications

1. Review of brain and spinal cord abnormalities.
2. Analysis of cysts, tumours and other abnormalities in any body parts.
3. Examination of joint injuries or abnormalities.
4. Examination of liver and other gastrointestinal diseases.
5. Understanding why women suffer from pelvic pain.
6. Detecting unhealthy body tissue.
7. Projects preparation.
8. To have an overall view of the collateral veins.
9. Offering intra- and extracranial views internationally.

20.2.4 Ultrasonography

It is a diagnostic methodology that provides broadband sound waves of high frequency megahertz. In the ultrasonography, to create medical images, which are reflected to various degrees by the tissue (Ovland 2012). It is located against the body of the patient, close to the problem area. After that transducer emits a flow of sound waves. These sound waves have high frequency. These high frequency waves enter the body. These waves reflect from the patient’s organs. This wave bounces out from the internal heart component by the help of transducer. Different kind of tissues reflect the wave uniquely like signature which is transform in image format for study. When wave enters the body, it is captured by ultrasound machine and transforms into images. These images are live. This continuous image captured in real time will be used to monitor the procedures for biopsy and drainage. Latest Doppler scanner techniques allow blood flow measurement in veins and arteries.

20.2.4.1 Advantages of Ultrasonography

1. Non-invasive and painless.
2. No ionising radiation is used.
3. Evidence in real time.
4. Intra- and extra-luminal irregularities are prone to detect flow changes.
5. Energy to calculate speeds.
6. Possible respiratory control.

20.2.4.2 Risks in Ultrasonography

1. No formal guidelines.
2. Dependent operators.
3. Save money.
4. Blinding is a difficult process.
5. Cannot take a regional view of the veins.
6. Influenced by the state of hydration.

20.2.4.3 Ultrasonography Health applications

1. The monitoring of fatal growth during pregnancy.
2. Imaging several neck and head structures including parathyroid glands and thyroid.
3. Seeing abdominal organs like kidneys, gallbladder, pancreas, spleen, liver, bile ducts, aorta and lower vena cava.
4. Guiding needle injections when inserting local anaesthetic solutions near to the nerves.
5. Echocardiography used to treat ventricles and valves in the heart and function.

20.2.5 Elastography

In medical imaging, it is a non-invasive procedure. In this method, biological tissues are recognised based on their rigidity as opposed to natural tissue (Tyagi and Kumar 2010). The first technique to perform elastography was the biomechanical characteristics and ultrasound elastography of soft tissues is extensively studied in clinical diagnostic applications (Sarvazyan et al. 2011). Through incorporating using MRI and shear waves to visualise their propagation (Asbach et al. 2010). By the use of pulse sequence, MR elastography acts to sensitise the MRI scan. These waves are produced by an electro-mechanical transducer on the body. At about the same frequency are the mechanical excitation and the gradient sensitising wave. This method has features that sense the given parameter on the human body for optimal diagonsis. This elastography is used with optical coherence tomography (Sampson et al. 2013). To create optical coherence elastography practical on human, an annular piezoelectric charging transducer is intended and even a simultaneous image can be obtained from it (Kennedy et al. 2009). Tactile imaging (Hoshi et al. 2010) is a diagnostic imaging method which converts a visual signal into the sense of touch. This method has features that sense the pressure on the body.

20.2.5.1 Benefits from Elastography

1. Non-ionising radiation and non-invasive.
2. To get immediate outcomes.
3. High accuracy calculation methods for 2D time change dependent on strain.
4. To get a detailed map of a standard transmural strain, high frame rate.

20.2.5.2 Elastography Risks

1. By raising the pressure applied, both the elastography images and the elasticity score can affect the elastography, which can contribute to a misdiagnosis.
2. Suffering from medical conditions causing tissue stiffness affected by irregular growths.
3. Resolution too small.

20.2.5.3 Medical Applications for Elastography

1. Detection and evaluation of particularly cirrhosis, liver disease.
2. Soft tissue inquiries.
3. Observing the cardiac muscle's electrical function during different stages of the heart cycle.
4. MR elastography, to analyse changes in the properties of muscle content associated with ageing.

20.2.6 Optical Imaging

Optical imaging is a non-invasive technique which reveals molecular structure and cellular in the living body. Optical imaging is considered an effective method for deep tissue sampling, where light propagates diffusely (Garofalakis et al. 2007). In the optical imaging, tissues morphology and biomolecular process information is extracted. The light will disperse diffusely. Light interaction with different components of the tissue allows the imagining of tissue anomalies (Yodh and Chance 1995). The breast cancer test system is most commonly used in optical imaging.

20.2.6.1 Benefits of Optical Imaging

1. Non-invasive.
2. Radiation anti-ionising.
3. Tumour features can be viewed as the patient lies in a prone position, and the visibility of most breasts is relatively strong.
4. Longitudinal studies can be carried out over a time span.
5. Potential for differentiating soft tissue due to the different scattering or absorption.

20.2.6.2 Risks of Optical Imaging

1. Owing to the diffusive light absorption in the breast tissue, low spatial resolution.
2. Sensitive to accumulation of lipid in breast tissue, water in the blood and blood oxygenation.

20.2.6.3 Medical Devices for Optical Imaging

1. To check haemodynamic.
2. Tumour identification.
3. To include functional brain imaging.
4. Breast cancer scanner.
5. Scanning healthy bones.
6. To check jaws, gums and teeth.

20.2.7 Radionuclide Imaging

It is a medical technology. In this method, radioactive material is used to obtain picture. Radioactive isotopes are given to the patient by injection or mouth in small amount. Human body are absorbing the isotopes and emission happens which is detected by detectors. These detectors detect the radiation in body parts and scanner scans the information and generates image. Three techniques comprise radionuclide imagery; some differences between these techniques are seen in Table 20.1, SPECT (Larsson 2005), PET (Carstensen et al. 2011) and Scintigraphy Planner (Kraft and Havel 2012).

Planner scintigraphy uses other organs to absorb certain radioactive compounds, either for a limited time or indefinitely, after they have been delivered by mouth or injection to a patient. Radioisotopes such as Tc99 m used between 2 and 6 h at the latest after training. The minimum Tc99 m dosage is 20 to 25 millicurie (Nikpoor 2009). It is helpful to hydrate the patient before imagery. Between visualisation and isotope injection, the patient is urged to drink 4–5 glasses of water. Imaging time

Table 20.1 Comparison of radionuclide imaging techniques

	Planar Scintigraphy	SPECT	PET
Origin	At a time one photon emits and this emitted photon moves in random route.	Gamma decay is produced by radioisotopes.	Positron decay is produced by radioisotopes.
Detector	Anger scintillation camera	Rotation of the rage camera to acquire multi-angle projection results.	Special coincidence detector circuitry for concurrently sensing two photons in opposite directions.
Methodology	Method like X-ray but use gamma rays. Only those photon capture which is move on one direction	The photons captured in different directions are identical to X-ray CT.	Capture various different-direction forecasts. Positron decay emits two photons at a time in two opposite directions.

depends on age. The nature of their distribution makes it possible to draw such assumptions regarding the body organ size.

20.2.7.1 Radionuclide Imaging Benefits

1. Provides highly accurate and precise functional details, frequently.
2. Provides an overall overview of the program of concern.
3. Strong contrast unique to the tissues.
4. May test the degree to which cancer has spread, and how well treatment works.

20.2.7.2 Risks with Radionuclide Imaging

1. High cost (production of equipment and isotopes).
2. Special caution appropriate to treat radioactive materials.
3. Many people can feel claustrophobic, which may mean they need sedation.
4. Relatively small spatial resolution.

20.2.7.3 Scientific Radionuclide Imaging Applications

1. Cancer diagnosis (cervical, oesophageal, neck and head, liver, colorectal, lymphoma, melanoma, pancreatic, breast, thyroid, etc.).
2. To assess the therapy's future efficacy.
3. Cardiovascular disease diagnosis
4. Alzheimer's disorder diagnosis, autism, neurological and epilepsy disorders, Parkinson's disease.

20.3 Tools for Image Processing

The important role of processing images is to increase the appearance of an image. In this computer field, there are many image processing software which is used in the healthcare field. Image Processing Tools provide engineering assistance and a wide range of plug-ins, toolkits, image processing features and software analysis for scientists. Most image processing techniques include a two-dimensional treatment of the image signal and use normal signal-processing techniques.

20.3.1 Medical Images

Medical imaging has undergone a significant advance in the modern medical industry. That's it. Technology is critical because it can be used during a real test in which a fee has been paid. Various forms of image analysis have been built up over the years; various medical picture forms adapt to various types of technology. When we study about the medical images then we know that every medical image has its particular benefits and drawbacks.

20.3.2 Medical Imaging and Its Properties

The visualisation of the organs, body or tissues used for medical diagnosis, treatment and disease tracking is medical imaging. The techniques of imaging include the fields of optical imaging, radiology and nuclear medicine. There are a few types of medical images, such as X-rays. In this diagnostic imaging techniques include advance radiation techniques for smart healthcare system.

Magnetic imaging and magnetic resonance imaging (MRI) are also kinds of medical imaging. The image, molecular imaging and CT work with ultrasound and MRI without the radiation being ionised, unlike conventional X-rays. MRIs employ powerful magnets which produce a strong magnetic field that forces protons in the body to align with that field. Imaging techniques, where ionising radiation is not necessary, can be used for certain types of clinical cases. Ultrasound scans hire waves, for example, with low frequency sonority.

20.3.3 Medical Image Processing Tools

The procedure, the method and the practise of medical imaging create visualisation representations of the body's interior aimed at medical practise and health study. Imaging medicine aims at exposing internal mechanisms concealed from the skin and bones, both for the treatment of diseases and diagnosis. Health images sets a basic anatomy and physiology database for Enable the detection of anomalies. Although images of organs and tissues removed for medicinal use are reasonable, such processes are commonly considered to be part of pathology rather than diagnostic image. So, the Chapter is going to be focus primarily on medical image processing equipment.

20.3.4 Medical Images Processing (MIP)

Medical image has its place in the modern medical field submitted a big advance. This technology matters as it may be implemented before an actual operation. On the several kinds of medical imaging have been created years ago, various forms of medical image suit various kinds of images engineering. Could medical images have its own merits and demerits? There are fifteen types of market-driven MIP equipment.

There are many technical resources used for the application of medical image processing. 15 Types of instruments widely used by researchers were introduced to the following section.

20.3.4.1 VTK

It stands for Visualization Toolkit (VTK). This toolkit is accessible for everyone so it is open source framework. VTK is 3D computer graphics software and platform is supported by Kitware, with the community now working to develop the future. This toolkit provides VTK Resources technical guidance and support. In addition, VTK has a robust information visualisation framework, 3D package Widget touch, enables parallel processing, and connects with numerous libraries of GUI toolkits, such as QT (Hanwell et al. 2015).

20.3.4.2 ITK

ITK stands for Insight Segmentation and Registration. This tool provides the image analysis to the developers (Roobottom et al. 2010). ITK is more powerful tool that provides registration algorithm and learning edge segmentation when we study about two and more dimension (Liu et al. 2014). It is an cross-platform framework and it is open source system.

20.3.4.3 FSL

Study produces FSL (FMRIB Software Library) Community, UK, OXFORD and FMRIB. FSL is wide DTI Brain, FMRIand MRI Research Toolkit imageryData (Smith et al. 2004). FSL is a comprehensive library of analysis tools for FMRI, MRI and DTI brain imaging. It provides the important library for algorithm for MRI images and also used for research works.

20.3.4.4 SPM

It stands for Statistical Parametric Mapping. It is used for statistical processes. This package of software is designed by Karl Friston. SPM is used for the brain imaging analysis. It is studying the data sequences like MEG, EEG, PET, etc. SPM helps for analysing brain anomaly or detects the abnormalities in the brain (Sowell et al. 2000).

20.3.4.5 GIMIAS

GIMIAS stands for Graphical Interface for Medical Image Analysis and Simulation. It is most powerful graphical interface, provides the solving simulation problem and also solves the complex biomedical image computing. It has the plug-ins of specific problem and also used for the research work (Larrabide et al. 2009).

20.3.4.6 NiftyReg

It is the most useful image registration software. It is used for the rigid and non-rigid registration. It is open source software developed by Translational Imaging Group (TIG 2014). It gives the more efficient result for medical image compared to other registration software.

20.3.4.7 Elastix

This software helps to solve the image registration software. It has the group of algorithms to solve the problem of registration of image. It is more powerful than other tools like ITK. Compare to other it has the more efficiency like fast configuration and other registration method. This is an open source software and also used for research works (Kerner et al. 2015).

20.3.4.8 ANTs

ANT is very useful for interpretation control and multidisciplinary data visualisation, and can derive information from large datasets (Avants et al. 2011). ANTs stand for Advanced Normalization Tools which is used for visualising multidimensional data and extract data from complex datasets. It is open source data.

20.3.4.9 NiftySeg

This tool has various programs to be used for analysis. It is used for EM based segmentation. This tool is indeed one of the university-developed programs, approved under BSD registration. It is a great thing. A tool involves many picture segmentation or format analysis programmes based on EM (TIG 2014).

20.3.4.10 ITK-Snap

ITK-Snap is a method for the segmentation of structures in 3D medical pictures, Paul Yushkevich produces it. This tool offers semi-automatic segmentation with active use methods of contour, and manual delineation and picture browser (Yushkevich et al. 2006).

20.3.4.11 MITK

It is a development platform which incorporates application structure with the Insight Toolkit Visualization Toolkit (VTK) and Insight Toolkit (ITK). The software is approved in compliance with BSD-Style (Lu et al. 2012). The MITK stands for Medical Imaging Toolkit.

20.3.4.12 NiftyRec

At UCL London, the NiftyRec software project that provides the Tomographic Reconstruction Code was created (Assaf and Alexander 2014). It has several types of package for registration like local and global. Registration of lungs also uses this package. This tool helps to us for free-form deformation algorithm when we use block-matching approach.

20.3.4.13 NiftySim

It is open source finite; high-performance toolkit uses for high graphics processing unit (GPU). This tool also has simulation abilities, developed at London University College, is a nonlinear feature solver, high-performance finite. The GPU-based execution option that allows a solver to greatly outperform market-like packages is a distinctive feature (Johnsen et al. 2015).

20.3.4.14 Camino

Camino is an MRI Processing software toolkit; it is capable of creating production pipelines that contain modules from other systems. Actually, Render Toolkit Maintenance is the imaging community of microstructures at UCL's lead development (Cook et al. 2006).

20.3.4.15 DTI-TK

It stands for Diffusion Tensor Imaging Toolkit. This tool is used for construction of an atlas Tool and spatial normalisation designed for analysing morphometry of white matter using data from DTI. In the year 2011, it published a journal in ImageNeuro. It rated DTI-TK as the cutting-edge method in its category (Keihaninejad et al. 2013).

A review of the fifteen medical photos in this section Table 20.2 tabulates the computing methods. Research is performed on the basis of the guidelines below. The following table provides a comparison of among the 15-MIP tools mentioned.

Fifteen Medical Image Comparisons Tabulated resources for the processing are as above. The requirements relating to comparisons are the most recent versions supported by the tool, Sponsored Device GUI, Medical Imaging Supported, provided tool languages, tool function, type tools, prices and platform type to run in standard to fulfill the requirement of consumers. As for the GUI tools, all 15 types of tools provided user interface power, so the existence of user-friendly and easy GUI view for beginners. There are fifteen different medical image-processing software modalities to help in smart healthcare. For example, VTK supports only the 3D Medical imaging, in which SPM can support 5 forms of PET, MRI, CT-Scan, EEG and fMRI medical imaging. MITK, GIMIAS and Elastix meanwhile endorse all forms of medical imaging.

Furthermore, each instrument has distinct characteristics, such as, Camino and Elastix supported both functions of their unit, while NiftyReg supporting only viewing and segmentation. In which there are different programing languages used to create the App Tool. On the other hand, C # is Camino, ITK-Snap, DTI-TK Tools SPM and FSL, NiftySeg and MITK Language C Utilities. GIMIAS, VTK and Elastix are the JAVA tools for C++. The remaining instruments, such as NiftyReg, NiftyReg, NiftySim and NiftyRec Equipment for pythons, are DTI-TK, ITK and Camino.

Based on the user experience and survey, for medical photos, Elastix would like to suggest tool that offers the most optimal image processing. This helps customers to download the new technical update of the tool and the GUI programme. It also encourages all forms of imaging, such as MRI, CT-scan, radiography and ultrasound. This promotes a broad variety of functions while image processing is performed and this procedure can be run as a separate device or paired with IDE software, for example, Visual studio, NetBeans, MATLAB. Furthermore, version on their official website tutorials for beginners is also offered. Elastix is an open access programme, with the latest available. It also has a big alternative for importing MIP services. Elastix is complete impact exporting. It supports Bitmap, PNG, Dicom, TIFF and JPEG variants of picture to import data.

It is for consumers who are looking for a full featured interface, The Elastix for MIP tool is recommended for research. During while users that either want to explore segmentation or segmentation, visualisation application, Camino should be carried out. Camino is similar to Elastix, which aims to cure both kinds of medication. Imaging and even offering all defined functionality, but it is possible to add only windows. Provide the installer with another operating system, while Camino

Table 20.2 The image processing tools comparison

Image processing tool	VTK	ITK	FSL	SPM	GIMIAS	NIFTYREG	Elastix	ANTS	NifySeg	ITK-Snap	MITK	NiftyRec	NiftySim	Camino	DTI-TK
Latest supported version	6.1	4.0	5.0	12	1.5	3.1	4.7	2.1	3.1	3.2	03	1.6.9	2.0	2.0	3.0
Date of last published	2014	2014	2014	2014	2013	2013	2014	2014	2014	2014	2014	2014	2014	2013	2011
EEG				y			y				y				Y
PET				y	y		y				y				Y
Mammogram					y	y	y		y		y			y	
fMRI			y	y	y		y	y			y			y	
Ultrasound					y	y	y		y		y		y		Y
CT-Scan				y	y		y				y	y	y		
X-ray					y	y	y		y		y				Y
MRI		y	y	y	y	y	y	y	y		y	y	y	y	Y
System Interface	y	y	y	y	y	y	y	y	y	y	y	y	y	y	Y

Image processing tool		VTK	ITK	FSL	SPM	GIMIAS	NIFTYREG	Elastix
Functions	3D Images	y				y		y
	Visualisation	y				y	y	y
	Reconstruction	y		y	y	y		y
	Generic	y	y	y	y	y		y
	Registration		y		y	y	y	y
	Diffusion	y			y	y		y
	Simulation	y		y		y		y
	Segmentation		y		y	y		y
System Language	VB.Net							
	JAVA	y	y					
	C				y			
	Python	y					y	
	C#			y				
	PHP							
	C++	y				y		y
Paid								
Standalone tool		y	y	y		y	y	
Open Source		y	y	y	y	y	y	y
Framework					y			
Plugin/Integration		y		y				
Platform	Mac OS X	y	y		y		y	y
	Windows	y	y	y	y	y	y	y
	Linux	y	y	y	y	y	y	y

Image processing tool		ANTS	NifySeg	ITK-Snap	MITK	NiftyRec	NiftySim	Camino	DTI-TK
Functions	3D Images		y	y	y	y			y
	Visualisation			y	y	y	y	y	
	Reconstruction	y		y		y	y		y
	Generic	y	y					y	
	Registration	y	y		y			y	y
	Diffusion				y	y		y	y
	Simulation			y		y	y		y
	Segmentation		y		y	y	y	y	
System Language	VB.Net	y							
	JAVA							y	y
	C				y	y			
	Python		y			y	y		
	C#			y					y
	PHP						y		
	C++								
Paid									
Standalone tool		y		y	y	y	y	y	
Open Source		y	y	y	y	y	y	y	y
Framework									
Plugin/Integration			y				y		y
Platform	Mac OS X	y	y	y	y	y		y	y
	Windows	y		y	y	y	y	y	y
	Linux	y	y	y	y	y	y	y	y

does not strong practical tool foundation. Some of the software are available on MIP for non-commercial use, free of charge is labelled as free from the Upper, bench. Free downloads can be found at online platform for the respective resources. In the case of MIP tools which are not listed as free, this means the consumer has to buy a license for the use of a given MIP tool.

20.4 Conclusion

This chapter provides several techniques of medical imaging and discusses that how digital image processing is useful in bioinformatics technology. We also discuss advantages, disadvantages, benefits and accuracy of these techniques. Many bioinformatics technologies and tools are used in images. It also describes some useful toolkits for custom solutions to be created. The development of medical imaging technology has provided a large amount of data. There are several types of medical image processing technique that has different constraints. When we study about MIP tools, there are only 15 MIP tools used. These MIP tools play a very important role in bioinformatics techniques. In future, these techniques and tools are improving accuracy for better result and detecting other diseases.

References

Asbach P, Klatt D, Schlosser B, Biermer M, Muche M, Rieger A, Loddenkemper C et al (2010) Viscoelasticity-based staging of hepatic fibrosis with multifrequency MR elastography. Radiology 257(1):80–86

Assaf Y, Alexander DC (2014) Advanced methods to study white matter microstructure. In: Cohen-Adad J, Wheeler-Kingshott CAM (eds) Quantitative MRI of the spinal cord. Academic Press, London, pp 156–163

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54(3):2033–2044

Caiani EG, Toledo E, MacEneaney P, Bardo D, Cerutti S, Lang RM, MorAvi V (2006) Automated interpretation of regional left ventricular wall motion from cardiac magnetic resonance images. J Cardiovasc Magn Reson 8:427–433

Carstensen MH, Al-Harbi M, Urbain JL et al (2011) SPECT/CT imaging of the lumbar spine in chronic low back pain: a case report. Chiropr Man Therap 19:2. https://doi.org/10.1186/2045-709X-19-2

Cook PA, Bai Y, Nedjati-Gilani SKKS, Seunarine KK, Hall MG, Parker GJ, Alexander DC (2006) Camino: open-source diffusion-MRI reconstruction and processing. In: 14th scientific meeting of the international society for magnetic resonance in medicine (vol. 2759). Seattle WA, USA

Garofalakis A, Zacharakis G, Meyer H, Economou E, Mamalaki C, Papamatheakis J, Kioussis D, Ntziachristos V, Ripollcts J (2007) Three-dimensional in vivo imaging of green fluorescent protein–expressing T cells in mice with noncontact fluorescence molecular tomography. Mol Imaging 6(2):96–107

Hanwell MD, Martin KM, Chaudhary A, Avila LS (2015) The visualization toolkit (VTK): rewriting the rendering code for modern graphics cards. SoftwareX 1:9–12

Hoshi T, Takahashi M, Iwamoto T, Shinoda H (2010) Noncontact tactile display based on radiation pressure of airborne ultrasound. IEEE Trans Haptics 3(3):155–165

Johnsen SF, Taylor ZA, Clarkson MJ, Hipwell J, Modat M, Eiben B, Ourselin S (2015) NiftySim: a GPU-based nonlinear finite element package for simulation of soft tissue biomechanics. Int J Comp Assisted Radiol Surg 10(7):1077–1095

Keihaninejad S, Zhang H, Ryan NS, Malone IB, Modat M, Cardoso MJ, Ourselin S (2013) An unbiased longitudinal analysis framework for tracking white matter changes using diffusion tensor imaging with application to Alzheimer's disease. NeuroImage 72:153–163

Kennedy BF, Hillman TR, McLaughlin RA, Quirk BC, Sampson DD (2009) In vivo dynamic optical coherence elastography using a ring actuator. Optical Exp 17(24):21762–21772

Kerner GS, Fischer A, Koole MJ, Pruim J, Groen HJ (2015) Evaluation of elastixbased propagated align algorithm for VOI-and voxel-based analysis of longitudinal 18FFDG PET/CT data from patients with non-small cell lung cancer (NSCLC). EJNMMI Res 5(1):15

Kraft O, Havel M (2012) Sentinel lymph node identification in breast cancer - comparison of planar scintigraphy and SPECT/CT. Open Nucl Med J 4:5–13

Larrabide I, Omedas P, Martelli Y, Planes X, Nieber M, Moya JA, Bijnens BH (2009) GIMIAS: an open source framework for efficient development of research tools and clinical prototypes. In: Functional imaging and modeling of the heart. Springer, Berlin, pp 417–426

Larsson A (2005) Corrections for improved quantitative accuracy in SPECT and planar scintigraphic imaging. Print & Media, Sweden

Liu Y, Kot A, Drakopoulos F, Yao C, Fedorov A, Enquobahrie A, Clatz O, Chrisochoides NP (2014) An ITK implementation of a physics-based non-rigid registration method for brain deformation in image-guided neurosurgery. Front Neuroinform 8:33

Lu T, Liang P, Wu WB, Xue J, Lei CL, Li YY, Liu FY (2012) Integration of the image-guided surgery toolkit (IGSTK) into the medical imaging interaction toolkit (MITK). J Digital Imaging 25(6):729–737

Mathews P, Jezzard P (2004) Functional magnetic resonance imaging. J Neurol Neurosurg and Psychiatry 75(1):6–12

Ng B, Abugharbieh R, Huang X, McKeown MJ (2009) Spatial characterization of fMRI activation maps using invariant 3-D moment descriptors. IEEE Trans Med Imaging 28(2):261–268

Nikpoor N (2009) Scintigraphy of the musculoskeletal system. In: Weissman BN (ed) Imaging of arthritis and metabolic bone disease. W.B. Saunders, Philadelphia, pp 17–33

Ovland R (2012) Coherent plane-wave compounding in medical ultrasound imaging, Master thesis. Norwegian University of Science and Technology

Roobottom CA, Mitchell G, Hughes GM (2010) Radiation-reduction strategies in cardiac computed tomographic angiography. Clin Radiol 65(11):859–867

Sampson D, Kennedy K, McLaughlin R, Kennedy B (2013) Optical elastography probes mechanical properties of tissue at high resolution. Biomedical Optics & Medical Imaging, SPIE

Sarvazyan A, Hall TJ, Urban MW, Fatemi M, Aglyamov SR, Garra BS (2011) An overview of elastography–an emerging branch of medical imaging. Curr Med Imaging Rev 7:255–282

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Niazy RK (2004) Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23(Suppl. 1):S208–S219. External Resources Pubmed/Medline (NLM) CrossRef (DOI)

Sowell ER, Levitt J, Thompson PM, Holmes CJ, Blanton RE, Kornsand DS, Toga AW (2000) Brain abnormalities in early-onset schizophrenia spectrum disorder observed with statistical parametric mapping of structural magnetic resonance images. Am J Psychiatry 157 (9):1475–1484

Spahn M (2013) X-ray detectors in medical imaging. Nucl Instrum Methods Phys Res Sect A 731:57–6311

TIG (2014) The TIG, Image processing tool. Retrieved from http://cmictig.cs.ucl.ac.uk/wiki/index.php/Main_Page

Tyagi S, Kumar S (2010) Clinical applications of elastography: an overview. Int J Pharma Bio Sci 1 (3)

Wang L, Alpert KI, Calhoun VD, Cobia DJ, Keator DB, King MD, Kogan A, Landis D, Tallis M, Turner MD, Potkin SG, Turner JA, Ambite JL (2016) SchizConnect: mediating neuro-imaging databases on schizophrenia and related disorders for large-scale integration. Neuroimage 124:1155–1167

Wiemer J, Schubert F, Granzow M et al (2003) Informatics united: exemplary studies combining medical informatics, neuroinformatics and bioinformatics. Methods Inf Med 42(2):126–133

Xu J, Tsui BMW (2014) Quantifying the importance of the statistical assumption in statistical X-ray CT image reconstruction. IEEE Trans Med Imaging 33(1):61–73

Yodh AG, Chance B (1995) Spectroscopy and Imaging with diffusing light. Phys Today 48 (3):34–40

Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage 31(3):1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015

21 Artificial Intelligence in Bioinformatics

Hari Om Sharan

Abstract

Artificial intelligence and Bioinformatics have a solid link and Artificial Intelligence gradually expanded attention in bioinformatics research. AI has become common for the researchers to deploy the readymade systems to categorize and data mining. In current scenerio, there are numerous intelligent systems exists. Bioinformatics combines the biology and informataion system (intellegent system). Artificial Intelligence can be used to examine procedure and classify the biological data in short time. Various Artificial Intelligence algorithms have been developed and used in bioinformatics analyses. This chapter summarizes the applications of Artificial Intelligence that deployed in bioinformatics.

Keywords

Artificial intelligence · Bioinformatics · Intelligent bioinformatics · Intelligent DNA sequencing · Genetic algorithms · AI tools in bioinformatics

21.1 Introduction

The study of life is called Biology—It is one of the most interesting aspects of science. A microlevel study of biology/science is sequencing of DNA and RNA strands, protein classification, and the analysis of gene expression on DNA microarrays through Artificial Intelligence.

The combination of biology and computational intelligence is called Bioinformatics. It combines the data science and biology which uses the application of

H. O. Sharan (✉)
Department of Computer Science and Engineering, Faculty of Engineering and Technology, Rama University, Kanpur, Uttar Pradesh, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_21

Machine learning and artificial intelligence for real and important objective (Hanif et al. 2019a).

We use the applications of computational intelligence in bioinformatics, and artificial intelligence and data science discover the biological systems and methods.

Artificial intelligence is used in bioinformatics for prediction with the growth and the data at molecular level, machine learning, and deep learning to predict the sequence of DNA and RNA strands (Ezziane 2006).

Bioinformatics is one of the major contributors of the current innovations in artificial intelligence.

Through machine learning we can develop the better understanding of biological data based on large datasets. By the use of machine learning applications we can predict and also detect the pattern based on big data sets.

We can solve various biological problems by the implementation of the mathematical/statistical models, algorithms and computational intelligence (Narayanan et al. 2002). A machine can work intelligently by the use of artificial intelligence.

There are numerous problems exists in bioinformatics which required a new concept or intelligent technology for being addressed to exploit biological data. Artificial intelligence is a branch of computer science and its approaches excel to deal the problems, pattern recognition, and prediction. And there is a lot of scope to predict and recognize the pattern of bioinformatics problems for the applications of artificial intelligence (Hanif et al. 2019b). Regression analysis (linear regression and logistic regression) and various AI algorithms in bioinformatics increase the capacity to solve the biological problems.

21.2 Overview of Artificial Intelligence

Artificial Intelligence is the subdivision of computer science that focuses on the development of machine intelligence, unlike the natural intelligence shown by humans, for example, problem solving, pattern recognition, learning, prediction and planning, etc.

By the use of Artificial Intelligence we can build smart machines which can learn from past experience, regulate the new inputs, and can work like humans.

The principal of Artificial Intelligence is based on the human intelligence in terms of learning and working. Generally people think about robots when we talk about artificial intelligence, but it is more than this. Artificial Intelligence simulates the intelligence in machines like humans for learning and problem solving (Altman 2001; https://www.educba.com/importance-of-artificial-intelligence/). The specific objective of artificial intelligence includes the more complex problem solving, learning and prediction, etc.

21.2.1 Classification of Artificial Intelligence

Artificial Intelligence can be classified into three categories: narrow artificial intelligence, general artificial intelligence, and artificial super intelligence.

Narrow Artificial Intelligence Narrow artificial intelligence is also known as Weak Artificial Intelligence. It is designed to complete a specific task (singular task) such as games, personal assistance system, and Google search (https://medium.com/optima-ai/how-ai-is-shaping-the-future-of-bioinformatics-f4aa17bce5a6).

General Artificial Intelligence General Artificial Intelligence is also known as strong artificial intelligence. And it is designed to perform the tasks on par the human capabilities and more complex problems without human intervention, for example, self-driving car (https://medium.com/optima-ai/how-ai-is-shaping-the-future-of-bioinformatics-f4aa17bce5a6).

Artificial Super intelligence Artificial Super intelligence is the system which is more capable than the human capabilities; artificial super intelligence is also known as hypothetical artificial intelligence (Agrawal and Srikant 1994). In this type of system machines become self-aware. And we can say that artificial super intelligence is the future of artificial intelligence (https://medium.com/optima-ai/how-ai-is-shaping-the-future-of-bioinformatics-f4aa17bce5a6).

21.2.2 Importance of Artificial Intelligence

The main goal of artificial intelligence is to make the system intelligent that are able to learning, prediction, accepting, and executing the tasks like to humans or beyond the capability of humans (https://www.educba.com/importance-of-artificial-intelligence/). Here we are going to discuss the few importance of artificial intelligence:

- Artificial intelligence automates the repetitive learning like robotic automation (Douzono et al. 1998).
- Artificial intelligence enhances the intelligence into the existing systems/products (Douzono et al. 1998).
- Artificial intelligence analyzes the data by the use of artificial neural network.
- By the use of deep learning or deep neural network we can achieve the accuracy in medical images (Burge and Karlin 1997).
- By the use of self-learning algorithms we can train the data for future use.
- Medical sciences

21.2.3 Limitations of Artificial Intelligence

By the use of artificial intelligence we can change every industry, but in limits. Artificial intelligence learns from the past experience and this is the basic limitation of artificial intelligence. There is no further method by which we can incorporate the information. Therefore we can say if input data is incorrect than it will affect the output data or may give the wrong results. And any extra feature of prediction can be incorporated separately (Cannata et al. 2008).

Artificial intelligence systems are trained enough to complete the defined task. And artificial intelligence systems cannot perform the different task for which they are nor designed, for example, the system that plays the cricket cannot solve the Sudoku (Cannata et al. 2008; Hassanien et al. 2008).

Therefore we can say that the artificial intelligence systems are very specifically designed for singular task, they focused to perform the single task.

21.3 Application of Artificial Intelligence (AI)

In early days technology was used only for automation and to minimize the use of papers for keeping record, but now a day's artificial intelligence is not only a theory it has many practical applications (Hassanien et al. 2008; https://blog.adext.com/applications-of-artificial-intelligence/). Here we are discussing few vital applications of artificial intelligence:

1. AI-Driven Chat bots
2. AI in e-Commerce
3. AI in Human Resource Management
4. AI in Healthcare
5. AI in Cyber security
6. AI in Supply Chain Management
7. AI in Modernized Industrial Engineering
8. AI in Retail Management
9. AI in Analysis of Image Clarification
10. AI in Precision
11. AI in Virtual Health Assistance
12. AI in Computer Based Coding
13. AI in Banking and Financial Assistance
14. AI in Air Transport
15. AI in Gaming and Entertainment
16. AI in Digital Media
17. AI in Agricultural Prediction
18. AI in Drug Discovery

21.4 Working of Artificial Intelligence

Artificial Intelligence works on the large data sets with speed, regression, and algorithms. It permits to data sets to be trained automatically from feature extraction on the data sets (Keedwell et al. 2002). Artificial Intelligence is the wide area of which includes several tools with the help of following sub-fields:

1. Machine Learning
2. Deep Learning
3. Artificial Neural Network
4. Natural Language Processing
5. Cognitive Processing
6. Computer Vision
7. Internet of Things (IoT)
8. Graphical Processing Unit

21.5 Overview of Bioinformatics

Bioinformatics is the combination of biological science and computer science, which focuses on the analysis of biological data through software or information technology tools. In other words we can say that bioinformatics is a developing stream of biological science which includes biological studies through information technology, it develops the process through algorithms for understanding the biological data or analyzing the biological studies (Kohonen 1982).

Bioinformatics is an application of computer science to analyze the biological data. Bioinformatics is also known as interdisciplinary research which includes biological data, computer science, and statistics (Azuaje 2001).

21.5.1 Challenges in Bioinformatics

Challenges in bioinformatics vary based on the scope, some of the challenges seen by biologist and some of the challenges seen by computer scientist (Liang et al. 1998). Here we discuss some of the challenges which were seen by both scientists, which are following:

1. Ability to forecast the record where and when will happen in a genome.
2. Ability to forecast the pattern of any primary record.
3. Ability to forecast cellular reaction to external provocation.
4. Determining operative protein-DNA, protein-RNA and protein-protein recognition codes (Michalewicz 1996)
5. Precise structure forecasting
6. Coherent design of small molecule inhibitors of proteins
7. Understanding to grow of new protein.

8. Understanding the occurring of speciation of molecular details (Golub et al. 1999).
9. Understanding to describe the continuous development of gene ontologism-systematic ways of any gene or protein.
10. Complete genome-genome assessments.
11. Quick analysis of polymorphic genetic variations.
12. Structural determination of large macromolecular assemblies (Lipman et al. 1989; Rost and Sander 1994).
13. Quick structural clustering of proteins.
14. Forecasting of unknown molecular structures.
15. Dynamic function and membrane structure through Computer simulation.
16. Simulation of genetic networks/algorithms.

21.5.2 Bioinformatics Applications

The general use of Bioinformatics is to abstract the information from biological data or biological studies through software. Bioinformatics is used in numerous fields such as molecular medicine, modern medical research, and investigation of genomics, drug discovery and development, prediction of protein structure, prediction of gene therapy, microbial applications, and many more (D'Haeseleer et al. 1999).

There are some software tools are available to manage the database, and retrieval of data or knowledge discovery and we can analyze the useful biological/molecular data. It is also having many research applications (Rumelhart and McClelland 1986).

21.6 Usage of Artificial Intelligence in Bioinformatics

The study of biological data is known as bioinformatics. We study the analysis of biological sequence and molecular structure in bioinformatics also includes the modeling of biological system. Application of artificial intelligence in bioinformatics comprises the clinical research through the matching of biological sequencing, protein structuring, gene therapy, etc. (Ryu and Sung-Bae 2002). With the help of analysis done through artificial intelligence we can design and develop the drugs and also analyze the complex systems.

- By the use of artificial intelligence methods, numerous complex problems and biological systems can be solved. Now a day's artificial intelligence is popular concept in bioinformatics research and computational molecular biology (Shepherd 1999).
- Prediction and analysis of gene or protein structure.
- Analysis of pattern recognition.
- Knowledge extraction from biological data.
- Prediction of design and development of drug.
- Biological data record keeping through AI algorithms.

- AI methods used for analyzing the biological data.
- AI methods used for analysis of DNA and RNA sequencing.
- AI methods can give the solution for complex systems determine the transformation of biological data in to data science.
- Meta knowledge in biological domain.

The success of AI in bioinformatics has widely used algorithms and methodologies including neural networks, probabilistic approach, decision trees, cellular automata, hybrid methods, and genetic algorithms to solve numerous biological problems (Narayanan et al. 2002).

21.6.1 Genetic Algorithm

A genetic algorithm is a heuristic-based search algorithm inspired by the natural evolution theory of Charles Darwin. The genetic algorithm follows the idea of natural selection to select the fittest survival and the significant output. The genetic algorithm is designed to perform the tasks in five major phases individuals known as population and each individual have genes. The combinations of defined set of genes form a chromosome. The fitness function determines the efficiency of an individual and calculates a score for each individual and the calculated score determines the selection of an individual leads to the selection of the fittest individual based on the fitness score for the progeny. The parents are selected for the reproduction through this process. The crossover is considered as the significant step for each pair of parents for the random selection of the genes results to the generation of offspring. There is a probability that the strings of the genes can flip for the generation of new offspring at mutation phase. The algorithm terminates by the repeated formation of the same generations, considered as the final product. Genetic Algorithm is utilized to effectively enhance the multiple sequence alignment (Su et al. 2002). This approach uses a population of alignments to generate a fitness score based on the matching and mismatching of the columns. Half of the suitable alignments are copied to the next generation leads to the crossover points to choose a cut for the random selection point in the first alignment sequence and another cut is made for the second alignment to adjust the first sequence. If we will consider as one parent is spliced to add the gap leading to the splicing, than another parent to add the gap to ensure the alignment consistency (Alga and Tomassini 2002). Genetic algorithm predicts efficient alignments results as compared to other alignment algorithms (Haupt 2007).

21.6.2 Use of Artificial Intelligence in DNA Sequencing

We can use the concept of artificial intelligence and machine learning in the area of molecular biology. When the artificial intelligence introduced in this area, many algorithms are designed and deployed to the analysis of different data sets. Nature of artificial intelligence said that it will be useful in practical rather than theory. It is a regular practice for most of the researchers that they compare the new approaches with the older one, to analyze the effectiveness and efficiency on the defined data sets (Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology). If we study about the molecular biology, then we have to discuss about DNA sequencing, it is an important assignment in the molecular biology. In this reference DNA chips are the best alternative and very important method for DNA sequencing.

21.7 Conclusion and Future Direction

Algorithms of artificial intelligence play an important role in Bioinformatics to streamline the complex systems to perform the multidisciplinary analysis within one frame. The existing techniques of artificial intelligence which are used to abstract the knowledge of discovery and pattern recognition from the complex biological data. A combination of biologist and computer scientist concludes which software tools are useful to solve the complex biological problems.

Future of interdisciplinary research on analysis of biological data or we can say that future of bioinformatics is driven by artificial intelligence that will save time, efforts, and accelerate the biological research. Artificial intelligence researchers search the prospect to deploy the AI algorithms in the new domain and upgrade the methods.

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. BMC Bioinf 3(35):12–16

Alga E, Tomassini M (2002) Parallelism and evolutionary algorithms. IEEE Trans Evol Comput 6:443œ462

Altman R (2001) Challenges for intelligent systems in biology. IEEE Intell Syst 2(32):14–18

Azuaje F (2001) An unsupervised neural network approach for discovery of gene expression patterns in B-cell lymphoma. Online J Bioinf 1:26–41

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Cannata N, Schröder M, Marangoni R, Romano PA (2008) Semantic web for bioinformatics: goals, tools, systems, applications. BMC Bioinf 9(4):1

D'Haeseleer P, Liang S, Somogyi R (1999) Gene expression analysis and modelling. Proceedings of Pacific Symposium on Biocomputing, Hawaii, (PSB99). Available from www.cgl.ucsf.edu/psb/psb99/genetutorial.pdf

Douzono H, Hara S, Noguchi Y (1998) An application of genetic algorithm to DNA sequencing by oligonucleotide hybridization. Proceedings of the IEEE international joint symposia on intelligence and systems Rockville Maryland, USA 5(34):92–98

Ezziane Z (2006) Applications of artificial intelligence in bioinformatics: a review. Exp Syst Appl 30:2–10. www.elsevier.com/locate/eswa

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–536

Hanif W et al (2019a) Artificial intelligence in bioinformatics ISSN 2410-955X– an international biannually. J Rev Article 5(2):114–119

Hanif W, Afzal MA, Ansar S, Saleem M, Ikram A, Afzal S, Khan SAF, Larra SA, Noor H (2019b) Artificial intelligence in bioinformatics. Biomed Lett 5(2):114–119

Hassanien AE, Milanova MG, Smolinski TG, Abraham A (2008) Computational intelligence in solving bioinformatics problems: reviews, perspectives, and challenges. Springer, Berlin 2 (34):3–47

Haupt R (2007) Antenna design with a mixed integer genetic algorithm. IEEE Trans Antennas Propag 55(3):577œ582

Keedwell E, Narayanan A, Savic DA (2002) Modelling gene regulatory data using artificial neural networks. Proceedings of the International Joint Conference on Neural Networks (IJCNN'02), Honolulu, Hawaii, pp 183–188

Kohonen T (1982) A simple paradigm for the self-organized formation of structured feature maps. In Amari S, Arbib M (eds) Competition and cooperation in neural nets, Lecture notes in biomathematics. Springer

Liang S, Fuhrman S, Somogyi R (1998) REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput 3:18–29

Lipman DJ, Altschul SF, Kececioglu JD (1989) A tool for multiple sequence alignment. Proc Natl Acad Sci 86(12):4412–4415

Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology

Michalewicz Z (1996) Genetic algorithms + data structures = evolution programs. Springer, Berlin 1(32):3

Narayanan A, Keedwell EC, Olsson B (2002) Artificial intelligence techniques for bioinformatics. Appl Bioinf 1:191–222

Rost B, Sander C (1994) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232:584–599

Rumelhart DE, McClelland JL. and the PDP Research Group (1986) Parallel distributed processing: volume 1 foundations. The Massachusetts Institute of Technology

Ryu J, Sung-Bae C (2002) Gene expression classification using optimal feature/classifier ensemble with negative correlation. Proceedings of the International Joint Conference on Neural Networks (IJCNN'02), Honolulu, Hawaii, pp 198–203

Shepherd A (1999) Protein secondary structure prediction with neural networks: A tutorial. Available from http://www.biochem.ucl.ac.uk/~shepherd/sspred_tutorial/ss-index.html

Su T, Basu M, Toure A (2002) Multi-domain gating network for classification of cancer cells using gene expression data. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN'02), Honolulu, Hawaii, pp 286–289

Big Data Analysis in Bioinformatics

22

Anugrah Srivastava and Advait Naik

Abstract

Biology is the science of nature's life. It regards living things in one cell or separate cells (e.g. creatures, plants, and micro-organisms). Natural sciences include numerous areas, including the study and characterization of creatures by subatomic components in cells. It also shows how the collaboration between the ecosystems is determined by species formed in cells. To enhance their exhibition at the characterized activities, sciences can be covered by various sub-topics such as drugs, organic chemistry, and brain study. Natural structures consist of various animals to work together in executing such orders. These systems will possibly boost key developments in human well-being and environmental management. For example, software engineering, bioinformatics, and materials science may provide different logical instructions on how natural structures can change under a variety of conditions after some time. These systems have qualified answers for the most natural and medical treatment frameworks. Through using broad knowledge analysis and datasets, the exhibition of the scientific frameworks in the plant, bioinformatics, and medical services can be enhanced. While there still exists a contrast to the seriousness of the word in large quantities of information, we understand the data of a huge volume and a wide variety of items, which are constantly refreshed and placed in many outlets, as well as extraordinary developments in the skill, movement, handling, and investigation of this information. This chapter emphasizes on applications of

A. Srivastava (✉)
Computer Science Engineering Department, Bennett University, Greater Noida, Uttar Pradesh, India

A. Naik
Department of Computer Engineering, Vivekanand Education Society's Institute of Technology, Mumbai, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_22

big data tools and techniques in bioinformatics. we also addressed big data issues and challenges in the field of bioinformatics.

Keywords

Natural science · Organic chemistry · Medical treatment · Software engineering · Bioinformatics · Brain study

22.1 Introduction

The systematic study "Hype Cycle for Emerging Technologies, 2015" of Gartner Inc. (Gartner 2015), where the investigation of consumer fervour, production, as well as gain over 2000 new mechanical systems are provided in the graphical framework, the idea of enormous data as a separate invention of the spectrum and handling of immense informational indicators has disappeared. The organization company has clarified its decision, acknowledging the idea of "enormous information" involves countless developments that were used successfully, which are parts of other mainstream areas and drifts and have become regular working devices. To date, relevant information must be isolated from the above primary task of dealing with such information. The greatest wins have been gained in businesses by working closely with the consumer and by being able to take advantage of a correct evaluation and prediction of potential buyers in the same way. This refers primarily to banks, broadcasting, retail, electricity, and utilities. Currently, we address the professional use of a lot of knowledge by companies in their ability and preparation business cycles and their capacity to assist businesses. Huge information resources allow associations to track assets proficiently, to predict opportunities that influence their operations, and to decide on conclusion more quickly. Computer science responded to gradual public activity changes by developing new logical tests, including online inquiries and business information examinations. The Network Intelligence is a territory where creative work reviews the work and realistic findings of human sensitivity usage (information presentation, arrangement, and association of information disclosure, data mining, the use of canny professionals) and progressive data (remote organization, e-mail) is being carried out in these exploration regions. Business insight (BI) incorporates mechanical instruments to assist corporate heads, business managers, and opposing customers in making business rules-based choices for assortment, handling, and analysis of business data. Business research involves a vast variety of instruments, applications, and implementations which permits associations to gather information from internal structures. A high value is given here to machine learning, methods for finding rules and relations in vast quantities of data, information mining, advanced methods for the interpretation and self-examination of information, choice of emotionally supporting networks and man-made reasoning, the arrangement of the recognition of normal dialects, and so on. In 2008 we can recall that at first it was mainly concerned with a circle of logic and in a huge measure with bioinformatics, recalling at the same period the beginning of the term “big data”. B. Hesper and P. Hogeweg in 1970 were the first to use the term “bioinformatics” in an article defined as “the study of information processes in biotic systems” written in Holland (Hogeweg 2011). The developers considered the administration of data in various systems, for instance, the collection of data during the time spent growth, the transmission of data via DNA to intercellular and intracellular cycles, the comprehension of data at different life levels as a characterizing life property. Bioinformatics today is a science that builds on the use of PC techniques to explore a variety of genomic knowledge. A major part of the progress of bioinformatics has been the rapid improvement in PC innovation and computational information handling techniques and the development of new broadcast communications advances.

This significant source of knowledge for scientists and medicine policymakers, the openness of the highest organic material—the human genome—made accessible to scientists worldwide and empowered bioinformatics as an aggregate science in which the achievements of particular groups would be made available to researchers promptly. With the Internet of Things approach today a large number of sensors rapidly gather information. CCTV and news outlets, for example, are gushing staff continuously with knowledge from social, portable, and various applications. Such data also need to be treated gradually and the effect is only worthwhile if the time needed for the preparation is limited. For instance, different sensors are utilized to screen basic frames and actual conditions. Sensor knowledge consists of a complex and unrestricted dataset that can be treated continuously to make control framework choices. A variety of information still exists, but typically a predetermined information system is forced to manage and break down information. As Excel tables and social knowledge bases, structured information is typically available. As far as vast information is concerned, different kinds of information are presently used, treated, and examined. Such information designs include content, texts (SMS), messages, tweets, blogs, site information, blog information, GPS data, images, sound, video, sensor data, reports, and social datasets. Therefore, huge information also involves a mixture of organized, semi-organized, and unstructured information on different arrangements that have to be created and broken down. Despite the difficulties, a large-scale inquiry can maybe inspect a lot of information to discover overviewed examples, connections, and useful expertise in various fields, such as consumer analysis, advocating promise, proposal systems, online media review and response, extortion, natural and man-made prevention. For example, the use of a large-scale information survey in agriculture areas is used to broaden crop respect to resolve food safety issues. Additionally, large-scale information analysis is used to investigate therapies and disease remedies such as malignancy. Research in bioinformatics is regarded as a field with large, extended, and complex datasets.

Bioinformatics is an interdisciplinary region that primarily involves nuclear science, software engineering, arithmetic, and insight. To understand and organize data on organic particles and to mention derivations from objective facts it mainly manages to demonstrate natural cycles in the subatomic stage. Bioinformatics, for example, focuses on genomics, proteomics, transcriptomics, metabolomics, and glycolic, in the statistical investigations of datasets. Today, the essential extension of natural knowledge that poses stockpiling and planning difficulties is the use of high-performance cutting-edge sequencing. Genome introduces to the entire arrangement of qualities or cell genetic material (DNA) of a creature in its first step (from 2014 to 2017) to meet some significant problems in the information science field and to promote data-driven disclosure (National Institutes of Health 2018). The H. Winkler (botanist) in 1920 suggested “genome” for the assignment of chromosomes for the age of an ever-growing number of rational terms, which finished with “-ome” (Winkler 1920). Until then there were definitions of the biome (the arrangement of living creatures) and the rhizome (root framework), but now there are numerous “-omes” (Baker 2013) among the researchers. The Greek addition “-ome” is a large number of these words, usually meaning “having the idea of” simultaneous progress in PC limitations and further advances in obtaining knowledge in various science controls found by genomes guided to the creation of wide orders named “-omics” in bioinformatics, which dissects the entire living being in their auxiliary relationship (DNA, RNA, protein, metabolites, and so on) The genomics, metagenomics, transcriptomics, proteomics products, metabolomics, interactomics, and various bioinformatic zones examine genomes, metagenomes, transcriptomes, proteomes, metabolomes, and other papers (Ohashi et al. 2015). Every discipline of bioinformatics has its reading items and own knowledge gains. However, they all produce enormous measurements of information in different configurations and levels that should be interpreted, arranged, appreciated, and pictured to expand current information and reinforce information. Over the long term, genomics is studying the structure, function, development, preparation, and modification of the genome of the life-form. The most detailed knowledge in bioinformatics is found in DNA groupings. DNA consists of nucleotide particles. Guanine (G), adenine (A), cytosine (C), and thymine (T) are the codes for the information found in DNA. The need for these foundations is what the hereditary code decides on. DNA sequencing is the way to establish the exact application of bases A, G, C, and T within a DNA strand. There can be several million bases for an ordinary bacterial genome. There are approximately 3.2 billion human bases in the human genome and about 200 gigabytes in magnitude of a solitary sequenced human genome (Robison 2014a). The major human genome was sequenced fully in June 2000 and about 228,000 human genomes were sequenced from 2014 onwards (Rosenberg 2017). Late, more than 500,000 human genomes were sequenced at Illumina, the greatest producer of DNA sequencers (Herper 2017). At the end of the day, natural information is still more easily sequenced. The Cancer Genome Atlas (Li et al. 2013) and the DNA Element Encyclopaedia (The ENCODE Project Consortium 2012) are instances of two large datasets. Science data storage stores size 40 petabytes (EMBL-European Bioinformatics Institute 2014a) are available at the European Bioinformatical Institute (EBI). As the information collected from different sources is routinely used, it is heterogeneous since it is positioned in different arrangements. Additionally, organic and clinical knowledge (e.g. clinical imagery in medical services) is generated increasingly and fast. Another organic broad knowledge characteristic is that it is scattered topographically (Kashyap et al. 2015).

The knowledge survey that gathers a plethora of information from natural and biomedical information like genetic preparation for the grouping of DNA lead to the creation of a prediction of the human well-being and infection that promote disease relief and the development of human well-being and lives. This is a set of huge information issues and institutions, like the national health institutes (NIH), understand the importance to tackle the huge information problems that have been found in managing and researching organic information. In 2012, NIH sent large-scale data to knowledge to allow creative biomedical work of inventive methodology and devices to boost the usefulness of biomedical large-scale information in the territory of huge information science. Massive quantities of details are used. The calculation of the knowledge generated every day in the advanced world today is enormous. For example, almost 500 million daily photographs are registered, around 56 million images on social media, and about 200 billion daily messages are sent. Current knowledge collections in petabytes are projected and exabyte datasets are soon usual. The simple calculation of information that needs to be dissected and dismissed is an important issue with huge information, but there are also problems in improving the other two features, namely the speed and the range of information. For example, customary databases contain very static and restricted stock information, deals, and customer information. Handling of information like this is not delayed as the approaching information stream rate is slower than the preparation period and, despite any preparatory delay, the preparatory results are usually still useful. We are presenting key ideas in research on large-scale knowledge, including both algorithms for “machine learning” and “unsupervised” and “supervised” cases. Here we are talking about the developments in current bioinformatics found by the creation of high-performance sequencing phases that contributed to the expansion of research and science skills and lead to the wonder of big data in science. The former is a field of logical developments that explores the impact of the modern World Wide Web objects, authorities, and frameworks and employs the use of computerized reasoning and data innovation (IT), and the latter is its space which tends towards dynamic issues. The need for more progress and strategies for ability organization, board, review, and interpretation of large data is validated. Current bioinformatics faces a wide range of techniques for translating and introducing the data, the concurrent existence of various programming instruments and information designs, and more than the problem of managing colossal amounts of heterogeneous information. New knowledge base management systems, rather than social frameworks, can help to address the problem of supplying huge information and to set a worthy timeframe for search results. Latest programming developments, like standard computer programming and visual writing programs, and aimed at addressing the problem of the numerous genomic information designs and at addressing them.

22.2 Big Data and Bioinformatics

Big data has influenced bioinformatics extremely well in recent years. The area of exploration is tremendous and complex. Scientists from around the world have attempted a few things by splitting the application and instruments into the area of bioinformatics. These methods can be used to manage vast volumes of information using multiple and dispersed progress in registration. This survey paper discusses a few uses of huge knowledge and gives us a diagram of its current and lets us consider the openings of future research.

Bioinformatics science and the limited effort of the information age are moving us into an era of "big data" represented by voluminous and gradual datasets and complex information-examination techniques. The uncovering of the DNA architecture took years of the joint work of several exploration groups from different nations after the sequencing of the human genome. Present-day developments take into account the sequence of the whole genome in a matter of days. Accessibility of vast information gives rise to extraordinary possibilities, but it also poses challenges in information mining and investigation. The AI techniques utilized in bioinformatics are equivalent and iterative. The methods can be utilized to tackle with large-scale information using sufficient and equitable developments in registration. Generally, massive information apparatuses conduct group-mode calculations and streamlined unavailable for iterative handling and large information dependency between working. Equal, incremental, and multi-see AI calculations have been proposed over the years. In addition, diagram-based systems and large in-memory information resources are established to reduce input/output costs and advance iterative handling. Normal large information systems are still deficient. Similarly, suitable devices are not available for some major bioinformatics problems, such as the rapid development of co-articulation and administrative organizations and the remarkable module ID, the identification of buildings through the development of protein–protein cooperation information, rapid analysis of vast DNA, RNA, and protein-related information, and rapid questioning of incremental and heterogeneous processes. During this data age, information is being generated by a wide range of sources, such as sensors that are embedded in MRI scanners, video recognition cameras, and other than individuals and staff. Considering the yearly revolution of the digital age, in the advanced world—information we generate every year—44 zettabytes or 44 trillion gigabytes will be produced continuously in 2020, the computerized universe being many times the size in 2013 (Turner et al. 2014b) (Turner et al. 2014a). With the digitization of our entire devices and the subsequent invention, the rapid transition to the data era has been completed. Since time immemorial development has been digitized, such as basic communication and video cameras.

Elite advancements are utilized in the logical examination, for instance, fast information capture devices and exceptionally high goal satellite information recording are used for the logical analysis. Aside from digitizing administrations and efforts, a different trend came into being late in the day in time for arranging all the items that have been produced around us, including signals, home appliances, cars, and force metres. Devices talk to one another to exchange information gathered from various sensors, to make knowledgeable operational choices without anyone. The company is referred to as the stuff web (IoT) (Gershenfeld et al. 2004). However, new trends are on the pinnacle of notoriety, depending on the concept of huge knowledge. The Internet of Things (IoT) is one such model. In any event, it should be remembered that not everything that we construct is of importance to be studied directly or carefully. It is just a piece that is helpful when classified as target-rich knowledge in the computerized universe. More objective than the information itself, metadata is more objective. As indicated by Turner et al. (2014b) (Turner et al. 2014a), the general information on IT in 2014 was about all the objective rich information; in any event, more than 20% of IoT information will be on the lake rich in objective information constantly in 2020. The hugely high quality of this knowledge is large because it installs true circumstances like, for example, natural shifts, digital assaults, purchasing floats, and pestilences approaching and because they are increasingly being created and exchanged. The knowledge is then commonly used for complex and clever control in large measure. Because of this broad availability of data and the improvement in the registration of elite data, a comprehensive information analysis has taken place to conduct constant and accurate illustrative surveys on monstrous information calculation to prepare clever and informed choices.

Bioinformatics research is rapidly filled with a volume of knowledge. Wide sources of information are currently not limited to molecular materials research or logs and lists of web-crawlers. With digitalization, all considerations, and accessibility of high-performance devices at lower prices, the amount of knowledge is growing everywhere and bioinformatics research is remembered. A single human genome is bigger in size by around 200 gigabytes, for instance (Robison 2014a, b). This pattern in the increasing volume of knowledge is helped also by a decrease figuring costs and the creation of major developments in the field of science. Currently, scientists are not using traditional laboratories to find a new biomarker but rely on vast, reliable genome data made available by numerous research meetings. For example, the robotized genome sequencer is getting lower and compelling advances in capturing bio-information and offers huge bioinformatics information to this new era. In the course of the years, the scale of knowledge in bioinformatics has grown considerably. In 2014, in comparison to the 18 petabytes (EMBL-European Bioinformatics Institute 2014a) (EMBL-European Bioinformatics Institute, 2014b) the European Bioinformatics Institute (EBI), science knowledge store-holders, had approximately 40 petabytes of information on quality, protein, and small particles. Their entire storage capacity increases every year. The Hinxton server farm group, with 17,000 centres and 74 terabytes of RAM, has been launched by the European Bioinformatics Institute to manage the information of their employees. EBI is not the principal association in the gigantic bio-information store, particularly critically. Its registration power is expanded consistently. Numerous organizations, including the National Centre of Biotechnology Knowledge (NCBI), the USA, and the National Institute of Genetics, Japan, layout and produce a range of organic databases and disseminate them around the world. For the more precise investigation, accessibility of high information volumes is useful particularly in a deeply moving exploration field such as bioinformatics. In any case, the great challenges of information here vary significantly from other essential issues of information, for instance, molecular materials information gathered from CERN or satellite high target information from the open information archive of NRSC/ISRO2. Initially, the knowledge in bioinformatics is deeply heterogeneous. Many questions of bioinformatics testing involve various heterogeneous and autonomous bases of knowledge for assumption and approval. Numerous unregulated connections often generate bioinformatics information and thus their origins speak to similar types of information in different systems. Second, monster and filling data in bioinformatics are all topographically appropriated around the world as far as calculation and number of occasions. Although a section of this data can be transmitted through the Internet, because of its scale, expense, safety, and other moral problems, the remainder is not adaptable (Marx 2013). This occasionally allows the analysis to be conducted at a distance and the results to be shared. Having regard to volume, pace, and assortment, but also geologically suitable details, huge issues in bioinformatics can be identified. Distributed computing innovations have been used, with a lot of accomplishments, to deal with these difficulties of high knowledge in bioinformatics. The best approach is to utilize the cloud both for storage and measurement purposes (Marx 2013). Truth is that this approach helps to take care of the huge knowledge challenges of monstrous, evolving, and distantly circulating information that is forced by bioinformatics science. A wide-scale genome study on a variety of cloud-based PCs can be used for Gaea. Bina Technologies, Stanford University, and UC Berkeley turn off a cloud-based genome examination system for an equipment piece, the Bina box, to allow the preparation of genome data and a cloud-based study section about pretreated data. Furthermore, for their successful cloud-section sharing, the Bina box reduces genome information size. This arrangement professes to substantially increase the efficiency of the genome analysis beyond customary methods (Rojahn 2012).

Big data examination examines immense, unstructured, and rapid difficulties with knowledge. A part of the unbelievable scientific methods is applied to gigantic knowledge. The majority of organizations, associations, and governments today produce various kinds of extraordinary and varied knowledge about nature. Associations typically use valuable data and serious benefits accumulated by massive knowledge initiatives. One of the major problems is to efficiently and quickly extract important data from such sources. In order to enhance market competence and render common purposes, comprehensive details can be accepted by the review instruments. Recently, test instruments have been applied to provide enormous information on the volume, speed, and range. Note that they are not expensive because some are open source available. Big data analysis is one of the exam systems that are most commonly used and involves equipment and open-source programming. It acquires gigantic measures of information to disperse them in modest circles just as it offers numerous detailed assets to effectively split information. All the above-listed technologies and devices must use the synchronization of the information sources within and outside. They are fundamental parts of the main methodology of information (Zakir et al. 2015). This section applies to a range of huge information research fields, including specifics of large information scans, research trends of extensive information scans, designs for extensive information scanning, major information advancement, and cloud-based information screening administrations.

22.3 Big Data Problems and Bioinformatics

In the field of bioinformatics, there are many other big data issues that are yet to be addressed. In view of the recent biotechnology broad data boom, many of these issues must be tackled as a matter of urgency, as discussed above. We divide into seven categories the issue of big data analytics in bioinformatics. The following is addressed.

22.3.1 Gene–Gene Network Analysis

Gene regulatory networks (GRNs) modifies a number of odd conditions, such as malignant development. The development of high-performance sequencing technique, system scientists are prepared to build gigabytes of knowledge. As a rule, the creation of such a huge amount of knowledge is not feasible. The reconciliation of enormous different GRNs from different sources helps to recreate the GRN brought together. Recreation of GRNs locally and subsequently by joining them through the Cloud Foundation can allow system scholars to investigate the diseased organization more easily. In addition, genomic medication can be believed. Despite the fact that a large number of GRN induction instruments exist, their overall consistency is obscure due to the lack of a large-scale approval. Issues are required to locate the best deduction component to identify anomalies in organizations and to organize objective drug-ability proteins and to use fast, accurate, and adaptable models. The quality co-articulation network analysis assesses the linkages between the various quality organizations obtained from the quality articulation investigation. The distinctive co-articulation surveys identify progressions in the long term or separate phases of the disease triggered by quality buildings. This helps to describe the relationship between buildings of quality and interest characteristics. In order to track genotypical similarity, quality buildings from various species may also be clustered. The co-articulation study of a quality network is a highly complex and iterative topic with a wide range of information investigation structures.

22.3.2 Microarray Data Analysis

The number and size of microarrays are quickly extending, fundamentally because of diminished expenses and a boundless utilization of trials with microarray. To catch the movements in articulation esteems over the long run or over different phases of an ailment, microarray tests were likewise led in quality example time spaces. For quick co-articulation and administrative organizations using huge volume microarrays, large information innovations are significant. As quality articulation information are gotten at different phases of an infection over the long haul, qualities influenced by the sickness can be recognized and the biomarkers for the illness can be distinguished. The expansion of time to the third measurement computationally makes the investigation much more muddled than the customary quality examination.

22.3.3 Pathway Analysis

Pathway research covers phenotypes of interest for genetic products, gene function prediction, recognition of biomarkers and characteristics, and patient and sample classification. Genetic, metabolic, and proteomic data is rapidly developing and big data technologies are needed to interpret these data in large quantities.

22.3.4 PPI Data Analysis

Complexes and variations of protein–protein interactions hinder high data levels for numerous diseases. PPI networks are researched in various life sciences fields with the development of large volumes of data. The volume, variance, and speed of data are a real big problem in PPI complex analysis. A structured and scalable architecture is required to quickly and accurately produce, validate, and classify PPI complexes.

22.3.5 Evolutionary Research

Molecular biological technological advancements have recently become a popular source for broad data generation. Numerous microbial projects such as entire microarrays, genome, and metabolomics have generated huge quantities of data. This wealth of knowledge is a valuable forum for analysing and archiving bioinformatics. A major big data complication in bioinformatics is research of functional adaptation and patterns of advances through microbial research through the study of primitive species.

22.3.6 Disease Network Analysis

The disease networks are formulated and continue to expand and new networks are introduced in their own format from various sources. The multi-target correlations between diseases in heterogeneous networks are helpful in learning disease-to-network ties. Traditional network analytics techniques will be inadequate beyond unstructured and heterogeneous data, without sacrificing the standard of information. The connections between heterogeneous disease networks need large-scale data technologies. Complex molecular networks describe causal or predictive genes or associated disease technique. Further datasets that could not be studied before can be analysed by researchers with the capacity to process these data rapidly. While large data collections can be analysed using current technology, data integration techniques are still inefficient. The study of many, heterogeneous databases of omics involves optimal integration methods. Moreover, modern high-performance approaches gather custom phenotypes from a large number of people. To recognize and visualize complicated patterns in data for the purposes of genesis analysis and diagnosis of disease, large machine learning tools are needed. Some bioinformatics issues were present before the big data era, they have increased considerably in complexity and efficiency since big data evolved. The existence of vast amounts of data made other issues possible. In each case, sophisticated technologies for big data analysis are urgently needed to address these major problems.

22.3.7 Sequence Analysis

As a successors of microarray technology, RNA sequencing technology is presented. The result of this successor is its precise and quantitative measurements of gene expression. Additional information requiring substantial master-learning models includes the RNA sequence data. Big data technology is utilized to display mutations, all-specific expressions, and exogenous RNA (e.g. viruses) (Fig. 22.1).

22.4 Big Data Analytics Techniques

The most commonly used methods for descriptive and predictive analytics on big data are supervised, unsupervised, and hybrid machine learning approaches. In addition, in big data analytics, different methods from mathematics are implemented. The big data volume problem can be reduced by the decrease of dimensionality. For dimensionality reduction, linear mapping techniques, like principal component analysis (PCA) and singular value decomposition, similarly nonlinear mapping techniques, like Sammon's mapping, kernel main component analysis, and Laplacian eigenmaps, are commonly utilized. Mathematical optimization is a powerful method utilized in big data analytics. Subfields of optimization are generally

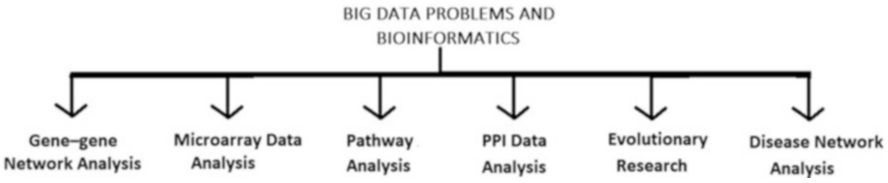


Fig. 22.1 Big data problems and bioinformatics

used in machine learning problems, like constraint satisfaction programming, dynamic programming, and heuristics and metaheuristics. Multi-objective and multimodal optimization approaches are other essential optimization methods, like Pareto optimization (Pareto 1964) and evolutionary algorithms (Fogel 2006), respectively. Statistics are used as an equivalent to machine learning and vary from the algorithm of the data model. The two areas have jointly subsumed concepts. In machine learning issues statistic principles, like expectation maximization and PCA, are commonly adopted. Macro-learning methods are likewise employed in applied statistics, like possibly roughly accurate learning. These two methods were, however, used extensively to analyse big data. Big data processing is similar to data mining. Because of the large amount of information, big data mining is complex as compared to conventional data mining. Extending current data-mining algorithms to large datasets is the standard procedure, executing specimen of big data and merging the specimen output. The clusters category includes both CLARA (clustering large applications) (Kaufman and Rousseeuw 1990) and BIRCH (balanced iterative reducing using cluster hierarchies) (Zhang et al. 1996). This is a classification of clustering algorithms. Researchers have stressed that data extraction algorithms are decreasing in their machine complexity. For example, discrimination in specimen regression reduces time and space complexity dramatically by simplifying discriminatory examination into a group of regularized minor square problems (Cai et al. 2008). Likewise, the spatial complexity of O(n^2) to O(n) nonlinear discriminant analyses is reduced by Shi et al. (2008) in order to minimize the computing and storage problem for large datasets.

Nevertheless, for a real-time study on big datasets, the time and space complexity of machine learning and statistical processes is complex. Because of their scalability, efficiency, and reliability, distributed and parallel computing technologies have proved in recent years to be the primary solution to large-scale computer problems. Therefore, attempts have been made with distributed computing to conduct big data analytics under strict efficiency and reliability limits. The literature, therefore, suggested distributed algorithms for data analytics. The collection of scattered has evolved as a modern data analytics model. It must be noted that to be efficient, the nodes must carry out calculations independently, without intermediate data with peer nodes being constantly exchanged. The distributed algorithms on classification education, association rule mining, and clustering are discussed in Park and Kargupta (2002). Rana et al. proposed to build distributed data mining applications a component-based framework known as PaDDMAS (Rana et al. 2000). Similar machine learning systems such as MLbase are proposed, for example (Kraska et al. 2013). Furthermore, cloud computing infrastructure-based frameworks, like the distributed GraphLab Architecture (Low et al. 2012), emphasize accuracy and fault tolerance in distributed analytics, have been put forward for distributed machine training. The market research on large commercial applications has been the key driving force behind big data analytics. Cluster and grid computing have been in use for a long time, they are uniquely built for specific applications and require huge costs and experience. Consequently, big data analytics technologies did not dramatically develop in that era. Research into big data analytics has became widely accessible on the cloud computing infrastructure and distributed processing systems like MapReduce (Dean and Ghemawat 2005) and their open-source executions. Iterative graphical treatment systems have also been put forward to address large-scale functional computer problems. At Google, Pregel (Malewicz et al. 2010), the patented graphic processing architecture addresses distributed handling of huge real-life diagrams. Apache Giraph offers extraordinary features, like edge-oriented entry and out-of-core computation, as an open-source counterpart of Pregel. In addition, the rising amount of data has led to a growing demand for big data analysis. Differentiated systems technologies like HDFS (Shvachko et al. 2010) and QFS (Ovsiannikov et al. 2013) and NoSQL databases are widely utilized in big data research in recent years for unstructured information, including mongoDB and CouchDB. For big data analytics, machine learning libraries were created. Apache Mahout (Owen et al. 2011), which includes deployments of various machine learning techniques including classification unit, clustering, and recommendation systems that are scalable to larger datasets, is the most important machine learning library for big data analysis. MLlib is a related library that offers Big Data Machine Learning, a MapReduce version that can be used to measure big data easily and on iterative terms. However, many important learning methods in machines are still missing and further contributions from the community are needed.

22.5 Big Data Analytics and Architectures

With multiple architectures, big data analytics systems were proposed. The three major architectures classified large-data solutions with each having its own advantages, limits and the suitability of algorithm depends on the design and specifications of the algorithm. The following is discussed.

22.5.1 MapReduce Architecture

MapReduce was originally created by Google (Dean and Ghemawat 2008) as an architecture parallel to data. Apache Hadoop is a commonly utilized MapReduce open-source programme. A daemon MapReduce runs all the time on the nodes. The design and control of the entire execution of this problem is performed by one master node. The remaining nodes are called working nodes and execute real data calculations. In addition, the master node divides the files, allocates files to worker nodes, and places them as pairs in the global memory. The figure demonstrates MapReduce's fundamental architecture, in which the nodes of the worker are indicated by Ni's. In rounds, MapReduce operates with two part each, specifically the map and the phases reduction. In both maps a node can be used and phases reduced. The three states of input, measurement, and output are in each step. There are two consecutive phases that have one synchronizing barrier. The local node memory is cleared and written into the global memory during synchronization. The master node will read/write on the global memory and interact continuously with the

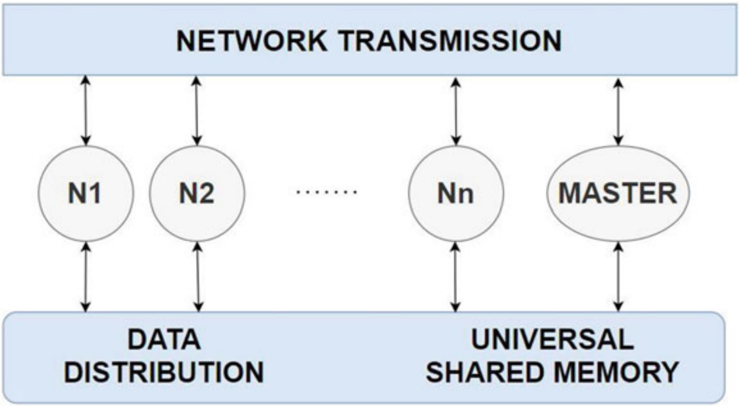


Fig. 22.2 MapReduce architecture

different nodes. However, only throughout syncing can the worker nodes read/write in the global memory. In Fig. 22.2, the thick and thin arrows have been separated. The problem is distributed between working nodes during the map process and limited results provided from the working nodes are saved in the global memory. The limited results are incorporated to produce the final result that is deposited in the global storage during the reduction process. The processes are replicated again when the intermediate findings are to be processed further. When the size of the data is big and the problem is embarrassedly parallel, the MapReturn architecture works well. By replacing the device (done by the failed node) for the process in a different node, the architecture provides defect tolerance. The design however has limitations on problems with high data dependency. Furthermore, iterative calculation cannot be used by the architecture and with high input/output overhead is inefficient. Research are performed to minimize and enhance the efficiency of the MapReduce architecture. Instead of writing on the distributed memory after each process, Twister (Ekanayake et al. 2010) optimizes iterative computations of the MapReduce architecture. However, because of in-memory processing Twister has problems with fault tolerance. To acknowledge the processing of memory and failure tolerance by reconstruction of an incorrect partition in the event of node failure, Apache Spark extends Hadoop to a resilient distributed database (RDD) (Zaharia et al. 2012).

22.5.2 Fault-Tolerant Graph Architecture

MapReduce inhibits high data dependencies are not very articulate if computer dependencies are complex between data and statistical methods. MapReduce is also not their best architecture. Supporting fault tolerance is needed to efficiently process the difficult and iterative problem. Fault scalability is also critical, as it permits insecure networks such as the Internet to be used. In order to do this, Low et al. (2014) first introduced a fault-tolerant graph-based architecture called the

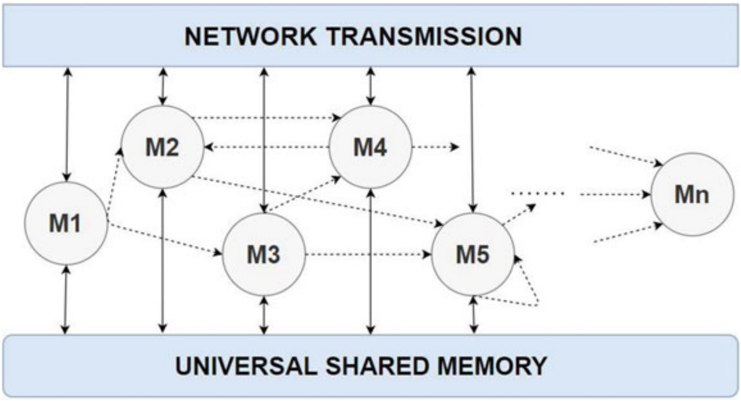


Fig. 22.3 Architecture for graph with global shared memory

GraphLab, and then similar architecture was adopted by several other major data solutions. The algorithm is heterogeneously split into nodes, with each one doing certain unique tasks. The model data is split into two sections: (1) a shared (distributed) global memory and (2) a computer node graph. The machine nodes are denoted by Mi's and the dotted arrows indicate the node dependencies and the network communication. Like MapReduce, the calculation is performed synchronously in execution intervals. With the input data the shared database is started. The node reads the shared database at the beginning of each loop and then allows computation using its own data and the data of its neighbour. The result is then combined and then returned for use in the next execution cycle in the globally shared database. When one cycle fails, one node is republished and one cycle loses the dependent nodes. Even if the efficiency is decreased by a loop, failure tolerance is assured. It is replaced when a node fails on a permanent basis. The architecture offers expression for complex data dependence and iteration problems. The architecture requires high disc input/output and hence is not executed optimizely. There is no further proposal to expand the use of RDD to promote memory processing and failure tolerance. In addition to GraphLab, Pregel and Giraph are other major graphic Big Data solutions. Graphical packages, for example, GraphX and the Hama graph package called Angrapa, are also built for the MapReduce architecture (Fig. 22.3).

22.5.3 Streaming Graph Architecture

The above-mentioned graphical architecture enables scalable distributed computing, complex operating data dependence, fault tolerance, and effective iterative processing. However, it is not effective for streaming data due to its high disc reading/writing overhead. While packages for analysis on MapReduce architecture stream data, for example, Spark Streaming, transfers stream data internally to bats for executing. Stream applications need high-bandwidth in-memory processing.

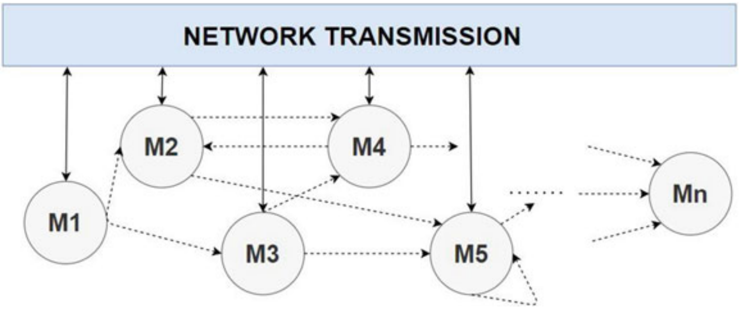


Fig. 22.4 Architecture for graph-based asynchronous processing

This issue is best tackled via the well-known Message Passing Interface (MPI) (Turner et al. 2014b). At application level, MPI has an API similar to MapReduce, and MPI can be integrated with almost all MapReduce programmes. Figure 22.4 shows the graph architecture for distributed processing in large-scale applications, for large bandwidth and iterative applications with large data dependence. The architecture is accompanied by rising calculated velocity and reliability of the network and increased bandwidth. Between this and the previous architecture, there are three significant differences. In this architecture, instead of global shared memory the nodes exchange data directly through peer-to-peer communication. Secondly, asynchronous operations are carried out. Only during their merger activities the various data flows are synchronized. Finally, data does not need to be stored in disk in this architecture. As every day memories get cheaper, large-volume data memory processing is feasible, thus increasing the total performance considerably. This architecture's key drawback is lack of tolerance of faults. If one of these nodes does not operate, then the process must restart from the start. As a consequence, in insecure networks such as the Internet, this architecture is unsuitable. In essence, this creates problems with scalability. However, if there is a stable network and there is a high data reliance on the algorithm, this architecture will provide more efficiency than the others. This architecture can be executed using MPI for processing big data on standalone clusters.

22.6 Big Data and Machine Learning

Big data is utilized in advanced organic sciences and regularly utilized for advanced knowledge analysis. Through organized measurable investigation (for example, the building of measurable inspection and trial conditions) it is not imaginable to analyse all the currently accessible details. The measure of accessible information overwhelms the capacity of a person to perform, significantly less decipher, the after effects of every single imaginable test currently, machine learning procedures begin to fill the void. Machine learning is a part of computerized reasoning (AI) and relies on the possibility of system models and meetings benefiting from them through the planning of detailed inputs without unambiguous programming. Thus, machine learning may promote knowledge analysis because it mechanizes the structure of the systematic model. Machine learning today is commonly used by numerous organizations, for example, cars, hereditarily characteristics to significantly enhance knowledge of the human genome, medical care, budgetary authorities, electricity, diversion method, and web-based method. Machine learning is now being extended to several different businesses. Information science and machine learning are developing important for business separation and endurance now and then. Machine learning performs a crucial role in resolving various issues in the bioinformatics industry, such as calculations for quality discovery and genome articulation, GWAS, and genomic preference. Machine learning is commonly used in bioinformatics, as enormous measurements of the subatomic science are currently being made available (Bhaskar et al. 2006) and the deep-informed nature of many issues in bioinformatics makes it illogic, if possible, to generate physically such calculations which will impeccably illuminate them. In recognition of designs, machine learning calculations have proven extremely convincing and are being implemented with exceptional success in bioinformatics applications. Machine learning calculations fall into two expansive classes: unsupervised and supervised algorithms. Both groups are better suited for specific inquiries and both will be required to accurately concentrate "big data" on biomedical squeezing issues. Machine learning techniques may be implemented to carry out viable integrative research, on which enormous scientific knowledge is assured. For genome-wide science, operational expenses of the information age are no big concern. The plant science network looks for innovative answer to knowledge problems involving a suitable fair measurement method, a shrewd information mining review, and plans for huge datasets. Four key methods of machine learning are as follows.

22.6.1 Supervised Learning

Supervised machine learning comprises a solution that uses a bunch of data collection, containing the sources of information and yields (marked with a correct yield), to provide a precise system resolution when new information is entered. The machine learning task then deduces a capacity to draw a contribution to the output that depends on the knowledge and returns from the information gathering model. By comparing its true output and its correct output, the calculation learns to discover errors and thus iteratively refine the model until a good executive level is reached. The most functional machine learning utilizes the technique of supervised learning. The point where the yield variable is a class is an arrangement problem, i.e. a certain number of qualities can be predicted. For instance, given the arrangement of information includes, the indicator capacity ought to anticipate either a tumour is favourable or harmful. The binary classification and multi-class classification are two kinds of arrangement to the problem. Binary classification is the position where yield maybe one or two possible qualities, normally 1 or 0, but in multi-class one of the three classes can be grouped, e.g. when the kind of disease can be expected.

Machine learning grouping issues estimates include option trees, strategic rebound, guileless habitats, near-K neighbour, irregular backwoods, and straight SVC (support vector classifier). A recurrence problem is a point where the predicted rates variable, for instance, temperature and weight, has a genuine or consistent value. Calculations of recurrence normal include recurrency, recurring trees (e.g. Random Forest), and support vector regression (SVR). The least difficult basic direct relapse model attempts to locate a measurable connection among two nonstop factors by adhering to a meaningful boundary that best fits the information. In (Segal et al. 2003), a variety of microarray regression approaches analysis like the vector support machine (SVM) have been presented.

22.6.2 Unsupervised Learning

Datasets with names are utilized in supervised machine learning, while unmarked datasets are utilized for unsupervised machine learning. With unsupervised machine learning, the structure is necessary for analysing the authentic information to discover parallels, examples, and links in the information to research about links in the information. Supervised machine learning is ideal for information with little information, such as "what examples exist in the quality articulation of malignant growths?" for example to address addresses. The two renowned tasks are the clustering of data and dimensionality reduction of data. Clustering is the way to find correlations in unlabelled knowledge that can be aggregated into a category. There are various types of grouping strategies available in which each philosophy follows an alternative concept or collection of rules to define the comparability level between information centres. The clustering of genes in expression data is the most typical application of clustering in bioinformatics (Larrañaga et al. 2006). The most popular use of bioinformatics bundling is the combination of qualities in articulation details. A few DNA microarray tests typically allow an estimate of the articulation levels of huge qualities. Bunching can be used to combine attributes in any example into a category with a comparative articulation level. The K-implies grouping and progressive grouping of the two most commonly used bunching computations in machine learning. K-implies bunching, a form of partial grouping calculation, follows the centroid model all the more clearly. It is an iterative grouping calculation by which a centre of the rags depends on the proximity of the details. The K-implies that the information is computed into K bunches, in which every group has a group site called the centroid. First of all, the focus of the K bunch is precariously set and the information to the closest community is poured out. The K-group fixates are checked again depending on whether the information in the bundles is inscribed. The proximity of the information to the latest server farm is examined and the cycle will be repeated until information does not change group involvement. Another known clustering technique is part-type but the model-based calculation is expectation–maximization (EM), also known as soft clustering.

22.6.3 Reinforcement Learning

Reinforcement learning is a method for measuring behaviour without the data collection, i.e. by experimenting to decide which activities are the best prize. Reinforcement learning comprises essential parts such as: (1) specialist (learning agent); (2) climate (specialist interfaces with climate); and (3) moves (specialists can make moves). A climate expert profits by consulting with him and accepting awards for activities. The professional will decide how he can achieve his objective by taking the best action to maximize compensation within a specific period. Such learning is objective or conducted based. Roundabout companies have a beginning and an end point (terminal state), but there are no end-state instructions, i.e. the expert stays until it unmistakably ends. Fortification is also used for mechanical and gaming technology. The Monte Carlo and temporal difference learning methods are two popular techniques for reinforcement learning. In bioinformatics, reinforcement learning is used to resolve the problem of the fragment assembly (Bocicor et al. 2011), the bi-dimensional problem of protein folding (Czibula et al. 2017), RNA reverse folding (Kohvaei 2015), and the 2D-HP protein folding problem (Dogan and Ölmez 2015).

22.6.4 Deep Learning and Neural Networks

Deep learning is an algorithm that aims to simulate the human brains' ability to learn. In these lines, the structure of the brain of neural organizations is activated by a profound learning design. Artificial neural networks (ANN) are used to provide in-depth learning to break up detailed knowledge and resolve complex problems using the immense personal power that is now available (e.g. example managing GPUs). Neural organizations have long been around but state-of-the-art ANNs are "profound"—a traditional neural organization consisting regularly of a couple of concealed layers. The use of ANN to produce and prepare models efficiently over time is conceivable. The algorithms learn with a deep learning model and determine all alone whether or not a projection is accurate. Automated driving, speech translation, and automated detection of cancer cells are deep learning applications.

22.7 Big Data Analytics Challenges and Issues

In recent years research into bioinformatics has quickly become a big data concern. Big data are distributed and gradual and have not only volume, velocity, and variety. This is why conventional data analysis can be done easily and reliably because of these broad data properties. In managing massive data analytics, machine learning techniques may be beneficial as they have developed in the computer science market with goals such as accuracy and effectiveness. This segment summarizes some of the difficulties and problems of big data analytics research using machine learning.

22.7.1 Big Data Analytics and Challenges

Big data are not appropriate for the methods used for analysing and visualizing classic databases. The number, speed, variety, distribution, and incrementality of such data present challenges in the conventional data analysis methods. Data generation volume and data transfer speed are increasingly increasing. Napatech, a high-speed network accelerator manufacturer reports a growth rate of 23% annually through 2018 for all network results. Recently, there is an exponential increase in hand-holder devices and their associated sensors. The rate of information production and circulation is growing alongside the rise in data volumes. The average connection speed for the mobile network in 2014 was 1683 kbps, according to the Cisco report (Cisco 2015), which will hit around 4.0 Mbps in 2019. With high data speed, real-time analysis of big data becomes harder. While batch mode analytics with distributed and parallel computing techniques are scalable to high data speed, the moderate input/output procedures critically influence the analytical efficiency. Currently, input/output rate is way behind machine rate, which serves as the boundary parameter. In addition, the data produced continuously are highly heterogeneous. In a set of specified schemes, conventional databases are organized. After the extraction – transformation – loading operations, data stores and upgrade data. Big data architecture constantly captures information from heterogeneous sources in high-speed and varied ways, the organized database, like data warehouse, which malfunction to dynamic storage and retrieval at the same moment. Due to various problems, conventional techniques for data analysis, like machine learning and statistical analysis, are not successful for large data as originally developed. The issue of machine learning powered analytics must therefore be explored from a large-scale data perspective. Another huge problem for big data analysis is data protection, especially in bioinformatics and in the healthcare sector. Data sources may use anonymity or publish only partial information to protect sensitive information. Imperfect or anonymous data collection can be complicated and counterproductive.

22.7.2 Big Data Analytics and Issues

Big data analytics involve massive handling of poly-structured, structured, and semi-structured data. Efficient time analysis introduces an extra time-bound computing requirement. Unstructured data may use machine learning techniques for identifying patterns and relationships. However, conventional analysis of big data has some efficiency problems mentioned below.

1. It still does not have a coordinated engineering for analysis in big data which endures blames and is equipped for handling huge, fluctuated information in clusters and in a continuous stream progressively.
2. The essential way to deal with and control the immense volume of enormous information is disseminated processing. Most machine learning, data mining,

and factual examination strategies, notwithstanding, were not at first proposed for appropriated count. Despite the fact that appropriated algorithm in writing was proposed, basically scholarly exploration, with absence of vigorous execution given the different MapReduce framework. (Choudhury et al. 2002; Raftery et al. 2005; Wright and Yang 2004).

3. There is no standardized data format for a big data store. Big data analysis requires instead to process heterogeneous data obtained by different types of sensors. Smart algorithms are therefore needed from disparate data to find a clear sense. This makes analytics more complex.
4. Unorganized, half-organized, and poly-organized information face extra issues, for example, confusion of information and redundancy. Due to their heterogeneity and huge volume, prepreparing information is costly. As far as existence intricacy, conventional information examination methods which attempt to deal with conflicting, uproarious information are exorbitant.
5. Big data analysis includes mining informational collections at different deliberation levels. Be that as it may, permitting researcher to decipher information at various degrees of deliberations clarifies the interest of semantics organic information. This incredibly builds the extent of investigative techniques.
6. A critical examination issue is implicit request that (1) co-articulation and administrative organizations for huge and different human and other miniature cluster datasets can be created all the more rapidly and (2) normal and particular highlights are similar and broke down more rapidly.
7. It is an extreme assignment to make a cost productive, adaptable, versatile design that empowers huge scope information investigation to inquiry heterogeneous natural information sources on any sickness organization.
8. A further exploration challenge is the advancement of an incorporated framework for the quicker investigation of voluminous and shifted quality articulation information assortments over the GST locale to perceive joint and exceptional patterns that help infection determination. The structure ought to likewise permit utilizes articulation, semantic, geography, and succession likenesses inside and remotely to approve the mining results.
9. Inference of large-scale diseased-compared GRN in the TF-target prediction of both networks accompanied by priority care in patients on the basis of a topological study of both.
10. A multidimensional perspective on an organization commonly unique regarding terms. Dynamic GRN portrayal is a major information investigation challenge.
11. Most of the strategies for derivation are insatiable in nature and computer-cost. They sometimes fall short for more extensive organizations best. The appropriated figuring model utilizing MapReduce and Hadoop could be investigated as an option without risking precision of the induction results.
12. The incorporation of enormous numerous GRNs from different sources assists with reestablishing the bound together GRN. The dramatic improvement of innovation for superior sequencing permits framework scientists to gather information gigabytes. It is regularly hard to move such enormous documents.

Neighbourhood GRN reclamation lastly cloud framework incorporation will help framework scholars better assess a confused organization.

22.8 Big Data Analytics Tools and Bioinformatics

For microarray information investigation, huge number of programming instruments (for example, caCORRECT) have been customized to examine different microarray information. Every one of these techniques is not in any case, used to deal with huge scope information. For the huge measured articulation information assortment, techniques are worked for the investigation of quality organizations (for example, FastGCN). A profoundly tedious technique is a PPI complex that looks to find some issue. The autonomous usage for PPIs with managed and unattended discovering frameworks requires days or even a long time to characterize the complexities of a broad dataset. To advance PPI complex discovering purposes, thus PPI information examination instruments (for example, NeMo) ought to be utilized. The Hadoop MapReduce system is created apparatuses for succession investigation (for example, the BioPig) (Nordberg et al. 2013) to deal with information measures on enormous arrangement information. To put it plainly, a few techniques (for example, GO-Elite) for pathway examination uphold the pathway investigation reason (Kashyap et al. 2015).

22.9 Conclusion

The chapter portrayed different definitions, portrayals, and encounters on biological big data examination. The measurement of natural information using HPC dependent multi-centre models has already been completed. Such a base can be very expensive and cannot be accessed easily. Furthermore, enormous bioinformatics knowledge measures are available via cutting-edge sequencing development. In bioinformatics applications, detailed knowledge examination and proper registration, i.e. distributed computing, are increasingly obtained and a community of figures are used to plan and break down information. Machine learning procedures are the great breakthrough that can reform bioinformatics applications. The written approach to bioinformatics questions is broadly suggested for machine learning. In this chapter, various methodologies were implemented for machine learning calculations. Deep learning is also used to fix more stunning problems in bioinformatics. It is common practise for the use of deep information of bioinformatics to significantly enhance the comprehension of the human genome and support to discover a solution to different diseases. Bioinformatics is essentially a confusing area of research. For every dataset and every organization no single computational approach would be optimal. A good information exam here undoubtedly calls for an imaginable combination of different information analysis techniques. In order to speak about appropriate data regarding a major bioinformatics knowledge investigation, forms, questions, and apparatus were identified. The chapter also discusses the continuing inundating volume as well as estimation in bioinformatic information stores. This rapid development of information will be continued in the future with the appearance of new high-performance and modest information capture instruments. Bioinformatics is voluminous, heterogeneous, incremental, and geologically adapted across the globe. The huge knowledge research strategies are therefore important in order to take care of the bioinformatics problems. Problems, sources of knowledge, and forms of information are complex in nature in bioinformatics.

A detailed information analysis, simple, open-minded wide-ranging and reliable, suitable and advanced for iterative and complex calculations, is not fully responded to with the current enormous information structures. The notable MapReduce architecture for acceptable registration is performed in a bunch mode with broad circle overhead read/compose. Again, the prototypes based on the map for streaming applications fail to adjust to non-critical failure. Coordinated large-scale knowledge inquiry technology that meets the requirements of bioinformatics issues is a significant need. However, large volumes of data present additional problems with the usual learning strategies in terms of speed, spectrum, and constant knowledge. Conventional learning techniques typically implement iterative management and dynamic dependency on knowledge between tasks. The traditional machine learning techniques cannot therefore be used to rapidly prepare gigantic knowledge, for example, MapReduce, using enormous data stages. The paper also addresses customary methods of machine learning, its shortcomings and its attempts in the years to extend them to include knowledge on complex biological problems, such as incremental, equal, and multi-site clustered techniques. There is still not enough information on the latest instruments for certain bioinformatics problems. Similarly, huge information devices in Hadoop or the cloud also fail to support other big bioinformatics problems, for example, the PPI network analysis or research into disease networks. Given the large volume of data in bioinformatics and the subsequent openings for study, massive knowledge research in bioinformatics should be acceptable in view of enormous progress in information production and viable exercises, such as machine learning.

References

Baker M (2013) The ‘Oms puzzle. Nature 494:416–419

Bhaskar H, Hoyle DC, Singh S (2006) Intelligent technologies in medicine and bioinformatics. Comput Biol Med 36:1104

Bocicor M-I, Czibula G, Czibula I-G (2011) A reinforcement learning approach for solving the fragment assembly problem. In: 2011 13th international symposium on symbolic and numeric algorithms for scientific computing, September 26–29, 2011. IEEE, Timisoara, Romania

Cai D, He X, Han J (2008) Srda: an efficient algorithm for large-scale discriminant analysis. IEEE Trans Knowl Data Eng 20(1):1–12

Choudhury A, Nair PB, Keane AJ et al (2002) A data parallel approach for large-scale Gaussian process modeling. In: SDM. SIAM, New Delhi, pp 95–111

Cisco (2015) Cisco visual networking index: global mobile data traffic forecast update, 2014–2019. Cisco Public Information, San Franisco

Czibula G, Bocicor M-I, Czibula I-G (2017) A reinforcement learning model for solving the folding problem. Int J Comput Appl Technol 2017:171–182

Dean J, Ghemawat S (2005) MapReduce: simplified data processing on large clusters. In: OSDIn'04, pp 137–150

Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51(1):107–113

Dogan B, Ölmez T (2015) A novel state space representation for the solution of 2D-HP protein folding problem using reinforcement learning methods. Appl Soft Comput 26:213–223

Ekanayake J, Li H, Zhang B, Gunarathne T, Bae SH, Qiu J, Fox G (2010) Twister: a runtime for iterative MapReduce. In: Proceedings of the 19th ACM international symposium on high performance distributed computing. ACM, New York, pp 810–818

EMBL-European Bioinformatics Institute (2014a) EMBL-EBI Annual Scientific Report 2013

EMBL-European Bioinformatics Institute (2014b) EMBL-EBI annual scientific report 2013

Fogel DB (2006) Evolutionary computation: toward a new philosophy of machine intelligence, vol 1. Wiley, New York

Gartner (2015) What's new in Gartner's hype cycle for emerging technologies. Accessed 17 Feb 2017

Gershenfeld N, Krikorian R, Cohen D (2004) The internet of things. Sci Am 291(4):76

Herper M (2017) Illumina promises to sequence human genome for $100 but not quite yet. Forbes, Jersey City, NJ

Hogeweg P (2011) The roots of bioinformatics in theoretical biology. PLOS Comput Biol 7(3): e1002021

Kashyap H, Ahmed HA, Hoque N, Roy S, Bhattacharyya DK (2015) Big data analytics in bioinformatics: a machine learning perspective, CoRR, arXiv:1506.05101

Kaufman L, Rousseeuw PJ (1990) Finding groups in data. An introduction to cluster analysis, Wiley series in probability and mathematical statistics. Applied probability and statistics, vol 1. Wiley, New York

Kohvaei P, (2015) Reinforcement learning techniques in RNA inverse folding, Master's Thesis, Albert-Ludwigs Universität Freiburg

Kraska T, Talwalkar A, Duchi JC, Griffith R, Franklin MJ, Jordan MI (2013) Mlbase: a distributed machine-learning system. In: CIDR

Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, Robles V (2006) Machine learning in bioinformatics. Brief Bioinform 7 (1):86–112

Li J, Lu Y, Akbani R, Zhenlin J, Roebuck PL, Liu W, Yang J-Y, Broom BM, Verhaak RGW, Kane DW, Wakefield C, Weinstein JN, Mills GB, Liang H (2013) TCPA: a resource for cancer functional proteomics data. Nat Methods 10:1046–1047

Low Y, Bickson D, Gonzalez J, Guestrin C, Kyrola A, Hellerstein JM (2012) Distributed graphlab: a framework for machine learning and data mining in the cloud. Proc VLDB Endow 5 (8):716–727

Low Y, Gonzalez JE, Kyrola A, Bickson D, Guestrin CE, Hellerstein J (2014) Graphlab: a new framework for parallel machine learning. arXiv:1408.2041 (preprint)

Malewicz G, Austern MH, Bik AJ, Dehnert JC, Horn I, Leiser N, Czajkowski G (2010) Pregel: a system for large-scale graph processing. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data. ACM, New York, pp 135–146

Marx V (2013) Biology: the big challenges of big data. Nature 498(7453):255–260

National Institutes of Health (2018) Big data to knowledge phase I & II, June 2018. Accessed June 27 2018

Nordberg H, Bhatia K, Wang K, Wang Z (2013) BioPig: a Hadoop based analytic toolkit for large-scale sequence data. Bioinformatics 29(23):3014–3019

Ohashi H, Hesegawa M, Wakimoto K, Miyamoto-Sato E (2015) Next-generation technologies for multiomics approaches including interactome sequencing. BioMed Res Int 2015:104209

Ovsiannikov M, Rus S, Reeves D, Sutter P, Rao S, Kelly J (2013) The quantcast file system. Proc VLDB Endow 6(11):1092–1101

Owen S, Anil R, Dunning T, Friedman E (2011) Mahout in action. Manning, Shelter Island, NY

Pareto V (1964) Cours d'e'conomie politique. Droz, Gene've

Park BH, Kargupta H (2002) Distributed data mining: algorithms, systems, and applications. In: Data mining handbook. Springer, Berlin, pp 341–358

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev 133(5):1155–1174

Rana O, Walker D, Li M, Lynden S, Ward M (2000) Paddmas: parallel and distributed data mining application suite. In: Parallel and distributed processing symposium, 2000. IPDPS 2000. Proceedings. 14th International. IEEE, New York, pp 387–392

Robison RJ (2014a) How big is the human genome? In: Precision medicine. Elsevier, Amsterdam

Robison RJ (2014b) How big is the human genome? In: Precision medicine. Elsevier, Amsterdam

Rojahn SY (2012) Breaking the genome bottleneck. MIT Technology Review.

Rosenberg E (2017) The human genome, Ch. 11. In: It's in your DNA. From discovery to structure, function and role in evolution, cancer and aging. Academic Press, Cambridge, MA, pp 97–98

Segal MR, Dahlquist KD, Conklin BR (2003) Regression approaches for microarray data analysis. J Comput Biol 10(6):961–980

Shi W, Guo YF, Jin C, Xue X (2008) An improved generalized discriminant analysis for large-scale data set. In: Machine learning and applications, 2008. ICMLA'08. Seventh international conference. IEEE, New York, pp 769–772

Shvachko K, Kuang H, Radia S, Chansler R (2010) The Hadoop distributed file system. In: Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium. IEEE, New York, pp 1–10

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Turner V, Gantz J, Reinsel D, Minton S (2014a) The digital universe of opportunities: rich data and the increasing value of the internet of things. International data corporation, white paper, IDC_1672

Turner V, Gantz J, Reinsel D, Minton S (2014b) The digital universe of opportunities: rich data and the increasing value of the internet of things. International data corporation, white paper, IDC_1672

Winkler H (1920) Verbreitung und Ursache der Parthenogenesis im Pflanzen - und Tierreiche. Verlag Fischer, Jena

Wright R, Yang Z (2004) Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 713–718

Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin MJ, Shenker S, Stoica I (2012) Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: In: proceedings of the 9th USENIX conference on networked systems design and implementation. USENIX Association, Berkeley, CA, p 2

Zakir J, Seymour T, Berg K (2015) Big data analytics. IIS 16:81–90

Zhang T, Ramakrishnan R, Livny M (1996) Birch: an efficient data clustering method for very large databases. In: ACM SIGMOD record, vol 25. ACM, New York, pp 103–114

Soft Computing in Bioinformatics

23

Vivek Srivastava

Abstract

In this chapter, we explored the soft computing based techniques for bioinformatics. Necessity of soft computing techniques and their compatibility for solving wide spectrum of bioinformatics related problems is reviewed. Basics of soft computing techniques are discussed and their relevancy in solving many bioinformatics based problems is also elaborated. Actual experimental results on two real world bioinformatics data demonstrated the efficacy of soft computing techniques over conventional one for biological data problems.

Keywords

Soft computing · Data · Bioinformatics · Sequences · Artificial neural network

23.1 Introduction

Human brains are enigmatic field of research that has been always being a remarkable area for researchers from long past. Biological neurons in human brains are responsible for transmission of information from one end to another. The crucial characteristics of human intelligence are the ability of recognition and classification of patterns. Artificial neurons imitated from biological neurons perform human like capability of intelligence for recognition and learning. Artificial neural network composed of such neurons offers learning ability which shows similarity of artificial intelligence with the human intelligence. Artificial intelligence demonstrates intelligent behaviours of machines similar to the human beings. Artificial intelligence is a grand field of research that primarily includes searching methods, knowledge

V. Srivastava (✉)
Rajkiya Engineering College Kannauj, Kannauj, Uttar Pradesh, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_23

representation and machine learning. However, conventional rule based artificial intelligence is not to provide proper solutions for various real world applications because it is incompetent to deal with huge amount of data. This gives rise to the non-conventional computation models for such applications. Soft computing is one of the research fields that emphasize on non-conventional computing models.

Soft computing actually obtained from artificial intelligence techniques that are centric to the natural way of problem solving. It basically includes neural network, fuzzy logic, evolutionary computation, support vector machines, swarm optimization, memetic computing, ant colony optimization and their synergism thereof. Soft computing techniques are vigorous especially in vague problems and provide efficient solutions of those problems also having uncertainty. Therefore, these methods are most significant for various real world problems and provide solutions effectively where traditional techniques are difficult to apply. Further, it has been investigated in various recent researches (Bhattacharjee et al. 2010; Ozbey et al. 2006; Yuan et al. 1995) that combination of two or more techniques is more efficient as compared to single technique. Basically, the techniques neural networks, fuzzy and evolutionary are correlated rather than competitive. A proper synergism among these techniques can yield efficient computing model and improved performance system for various real world problems.

On the other hand, Bioinformatics refers to the development and applications of methods and techniques for exploration and expansion of medical, behavioural and biological data. Soft computing techniques are successfully applied for knowledge derivation from biological data in most of the domains of bioinformatics. Problems are categorized into six different domains: proteomics, genomics, text mining, microarrays, systems biology and evolution (Larranaga et al. 2006).

23.2 Necessity of Soft Computing in Bioinformatics

Bioinformatics is a discipline that includes both biology and information technology. In general, it refers basic strategies to organize, store, achieve, analysis and visualize biological data. Hence, from various literature surveys, it is well established that soft computing techniques are more suitable for several bioinformatics related problems like clustering, classification, selection of gene and image processing.

There are any many tasks in bioinformatics that can be addressed and solved with the help of soft computing. Drug design, gene/promoter identification, exploration of biological data like alignment of gene sequences, genomics, proteomics, protein structure and DNA Prediction, protein folding, genetic analysis of gene expression data, etc. are the main problems associated with bioinformatics. In previous researches, statistical analysis tools like regression and estimation were used in bioinformatics. Therefore, in bioinformatics, soft computing techniques can be utilized in dealing complex, huge and inherently uncertain biological data that can provide robust and computationally efficient solution to problems. For example, fuzzy system can provide a natural framework for analysing biological data as fuzzy systems incorporate natural way of overlapping of memberships. In many bioinformatics based problems, there is a need of efficient searching and optimizing the solution. Soft computing techniques are able to provide robust, fast and close approximate solutions for these problems. In order to explore huge and multi-model solutions, soft computing techniques like genetic algorithm, ant colony optimization and particle swarm intelligence can lead to provide powerful searching and analysis methods.

Soft computing techniques are adaptive, i.e. easily adapted to a changing environment. Hence they can be more useful in field of molecular biology as in research related to molecular biology there is always an update in data. Models associated with soft computing techniques are not necessarily required to be redesigned as there is change in the environment. Further, there are several problems that contain multiple conflicting objectives, in such scenario, multiple objective optimization algorithms are more appropriate.

Genomics is the crucial field in bioinformatics that addresses genetics applications, explore the functional working and architecture of genomes, recombinant DNA and DNA sequencing techniques. In order to acquire useful information, genomics data requires pre-processing. One can extract the location and structure of the genes from genome sequences. In order to predict gene function and RNA secondary structure, sequence information can be utilized (Carter et al. 2001; Mathe et al. 2002). Proteomics is another field of exploration for which soft computing techniques can be feasible for approximation of protein structure. Proteins are complex macromolecules consisting of thousands of bounds and atoms. In order to solve human illness problem (Aerts et al. 2004), genomic and proteomic data analysis is important interface for understanding the crucial facts. There is wide application of genomic and proteomic technologies that involves large amount of complex data (Lancashire et al. 2009). In such scenarios, neural networks can be useful for diagnosis of illness problems due to its ability of dealing with large and complex data. Microarray domain refers to the computational application in biology that consists of complex experimental data. Complex experimental data further involves two issues. The first issue is data pre-processing and second is data analysis that depends on what we search for. Common techniques that are well applied in microarray data are neural networks, genetic algorithms. Systems biology consists of biological components like molecules, cells, organisms. Soft computing techniques are also appropriate for system biology. Modelling of the life processes is challenging task that occurs inside the cell. In such problems, soft computing techniques are very helpful for designing biological networks (Bower and Bolouri 2004). Text mining domain is another interesting field that deals with large amount of data. This domain is applied in location estimation for cells, functional annotation and protein interaction exploration (Krallinger et al. 2005).

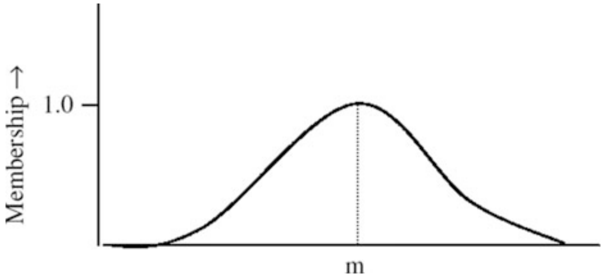
23.3 Soft Computing Techniques and Applicability in Bioinformatics

Soft computing is initially coined by (Zadeh 1994). First time, (Bezdek 1992) called synergism of fuzzy logic, artificial neural networks and genetic algorithms as soft computing. According to the wide acceptance, the most prominent parts of soft computing are fuzzy logic, neural networks, evolutionary algorithms, swarm intelligence, ant colony optimization, probabilistic reasoning, belief networks and artificial life. In this chapter, main focus is given to the fuzzy systems and fuzzy clustering, neural networks and support vector machines, evolutionary algorithms, hybrid intelligent system, ant colony optimization and particle swarm optimization. These approaches are presented in Sects. 23.3.1, 23.3.2, 23.3.3, 23.3.4, 23.3.5, and 23.3.6, respectively.

23.3.1 Fuzzy Systems and Fuzzy Clustering

Fuzzy sets are sets whose elements have different degree of membership which may be defined by membership function. Membership function decides the membership of an element in a set. Basically, membership refers to the degree of belongingness of an element to a set. An element can belong to different sets by different degree of belongingness. Fuzzy set provides flexibility in decision boundaries. An example of fuzzy membership function is shown in Fig. 23.1. The degree of fuzziness is denoted by fuzzifier (m) which can achieve maximum value up to 1. In present work, we focus on fuzzy clustering and its synergism with evolutionary computation. Fuzzy clustering employs a membership function which represents the fuzzy partition of input data into clusters. A pictorial representation of fuzzy clustering is shown in Fig. 23.2. This figure shows the three clusters with overlapping boundaries. It has been shown that a single point can belong to two or more clusters with different degree of belongingness. Introduction of evolutionary algorithms in fuzzy clustering provides optimized results in fuzzy classification (Hruschka et al. 2009). It is well established that fuzzy clustering provides comparatively better ability in solving many problems including classification problems (Mingoti and Lima 2006) over the conventional ones. Various versions of fuzzy clustering have been well applied for problems with noise (Yang et al. 2011). Hence, combining fuzzy c-means clustering

Fig. 23.1 Degree of belongingness: membership function



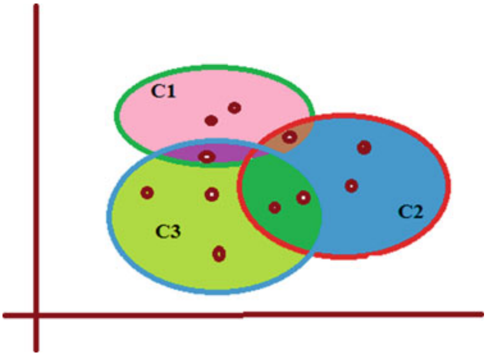


Fig. 23.2 A pictorial representation of Fuzzy clustering

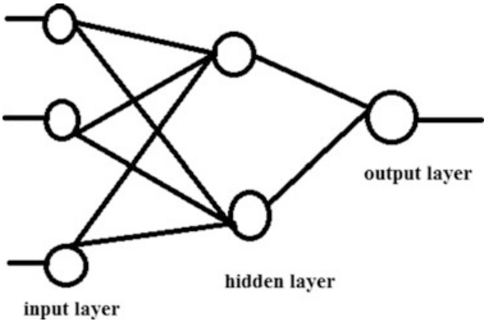


Fig. 23.3 The conventional neural network: multi-layer perceptron

with evolutionary computation provides better performance than fuzzy clustering (Hruschka et al. 2009; Fazendeiro and de Oliveira 2007).

In many researches, wide applicability of fuzzy clustering in bioinformatics is demonstrated. In (Dembélé and Kastner 2003), authors used fuzzy clustering for partitioning of microarray data into clusters. They also observed that fuzzy clustering is more suitable than conventional k-mean algorithm as it not provides overlapping shape of clusters. In other research (Maji and Paul 2017), authors applied rough fuzzy c mean for identification of co-expressed microRNAs, grouping functionally similar genes from microarray data and segmentation of brain magnetic resonance images using standard validity indices.

23.3.2 Neural Networks and Support Vector Machines

Artificial neural network is analogous to the biological neural network for machine intelligence. The artificial neuron model was firstly proposed by Warren MacCulloch (McCulloch and Pitts 1943). Two basic functions of this model are as follows: summation part which aggregates the input with weights and second part produces output by applying activation functions on aggregated information. A general neural network structure is shown in Fig. 23.3. It is basically a three layer architecture which contains input, hidden and output layer. Each layer may have different number of neurons. The selection of proper number of neurons in each layer is problem specific which greatly influences the system performance.

Another supervised method used for classification is Support Vector Machine (SVM). It categorizes by designing an N-dimensional hyper plane which separates the data into two different categories optimally. Neural networks are very similar to SVMs. SVM using a sigmoid feature of the kernel equates more or less to the artificial neural perceptron network of two layers. SVM is an alternative training technique for a function of the radial base, a polynomial and a multi-layer perceptron cluster where network weights are detected by resolving linear, limited, quadratic programming problems instead of by resolving an unconstrained, non-convex problem as in conventional neural network.

Neural artificial network identity has been developed for robustness in ill-defined classes and noisy patterns. This is because of the strong ability of artificial neural networks to generalize, connect and understand. Latest proposals (Gaxiola and Melin 2010; Barbosa et al. 2009) suggest and successfully apply various forms of neural networks for different applications. Some of the major variants are the radial base neural network, multi-layer perceptron, vector support machines, neural modular network and neural networks of higher order (Tripathi and Kalra 2011b; Tripathi and Kalra 2010a). These neural network variants are also used extensively in higher dimensions for many machine learning problems (Tripathi and Kalra 2011a; Tripathi and Kalra 2011c; Tripathi and Kalra 2010b). Further, it has been demonstrated in few works (Bhattacharjee et al. 2010; Lu et al. 2007) that combination of neural network with fuzzy performs better than the neural network only.

Neural networks are well applied in various bioinformatics based problems. In (Zamani and Kremer 2013), potential applicability of artificial neural network has been demonstrated in wide spectrum of computational biology based problems. On the other hand, in (Tang et al. 2019), authors showed deep learning employed successfully in bioinformatics problems. In DNA sequence analysis (Leondes 2003), artificial neural network proved its efficacy over existing methodologies. In (Tampuu et al. 2019), for identifying viral genomes in human samples, deep learning is applied on raw DNA sequences. In (Huang et al. 2018), support vector machines are also used for cancer genomics.

23.3.3 Evolutionary Computation

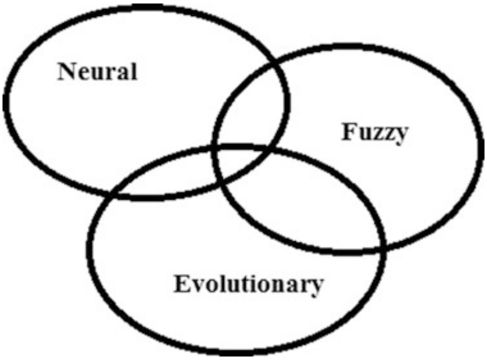
Evolutionary computation is based on the evolutionary process of biological species by natural selection. The fundamental concept behind the evolutionary algorithm is to find the fittest population among given populations based on natural selection, i.e. survival of the fittest. For this, a quality measure is introduced for assessment of fitness and candidate solutions are generated by maximizing quality measure. This quality measure works as abstract fitness measure. Evolutionary algorithms have gained importance in recent past due to its efficient problem solving capability primarily in field of searching, optimization and learning. The main function of the evolutionary algorithms is the generation of population and selection of best from different off springs generated. The various kinds of evolutionary algorithms are proposed (Hruschka et al. 2009; Deb 2001) and successfully implemented for solving clustering problems in the literature. Evolutionary algorithms are suitable for optimization of various aspects of clustering such as partitioning quality and number of clusters. It has been analysed by various researches (Hruschka et al. 2009; Zio and Baraldi 2005; Handl and Knowles 2007) that incorporation of evolutionary algorithms with clustering provides better clustering solutions that are also suitable for bioinformatics problems.

Evolutionary computation generally involves genetic algorithms, evolutionary strategies and evolutionary programming. All three components are rigorously involved in solving various kinds of bioinformatics problems (Fogel and Corne 2003). In (Chiesa et al. 2020), Genetic Algorithm is used for the identification of a Robust Subset of features in high-dimensional datasets.

23.3.4 Hybrid Intelligent System

In the last few decades (Ozbey et al. 2006; Zio and Baraldi 2005), hybridization of techniques of soft computing has been promising and successful. The creation of the computationally improved system will result from synergy of progressive calculations, fuzzy logic and a neural network. The pictorial representation of hybrid soft computing is shown in Fig. 23.4. The three circles with overlapping represent these three techniques. The intersection region between two circles represents synergism of two techniques that indicates evolution of evolutionary fuzzy, neural-fuzzy and evolutionary-neural system. The intersection region among all designates the hybridization of three techniques which evolves the development of evolutionary-neuro fuzzy system. In recent researches, in order to effectively resolve the different real world concerns, hybrid computer intelligence approaches have been used (Sio-Iong 2009; Su 2011; Zhang et al. 2002). The techniques also refer to applications of biometrics (Bhattacharjee et al. 2010; Zhang et al. 2002; Haddadnia et al. 2003). Various literatures (Bhattacharjee et al. 2010; Ozbey et al. 2006; Su 2011; Zhang et al. 2002) identify the strength and efficacy of such techniques. The

Fig. 23.4 Hybrid intelligent system



key explanation why hybrid computer intelligence techniques are outperformed by individual techniques is because they are complementary and not competitive.

The researches establish the significance and efficacy of hybrid computational intelligence techniques over individual technique. In techniques (Ozbey et al. 2006; Sio-Iong 2009; Saha et al. 2009; Wang et al. 2012), various combinations of fuzzy clustering with neural networks are developed. Method described in (Ozbey et al. 2006) employs fuzzy self-organizing layer for single-neural network preclassification. In (Saha et al. 2009), the authors merged the fuzzy c-medoid clustering algorithm with the categorical data neural network classifier. In (Sio-Iong 2009), the author developed a hybrid intelligent algorithm for the non-parametric regression model based on neural network regression and fuzzy c-means clustering. In (Wang et al. 2012), an incremental learning method based on the probabilistic neural network and adjustable fuzzy clustering was developed by the writers. These researches have demonstrated the wide applicability of hybrid computational intelligence techniques for various engineering applications. In most of the researches, the synergisms of any two aforementioned computational intelligence-based techniques are being designed and deployed for efficient solutions. However, a comparatively more efficient system can be designed by synergistic integration of all three techniques.

23.3.4.1 Learning Algorithm of Hybrid Intelligent Model

In this section, we will elaborate learning algorithm for hybrid intelligent model (HI-model) consisting of evolutionary clustering and neural networking. The HI model employs back propagation algorithm with momentum which is based on the principle of gradient learning theory. It is stated below:

Input:

A set contains training tuples (Training Set)

Associated desired output values (d)

Q-H-Z architecture of each associated neural network with total C neural networks

N = Number of training patterns

```
Method:
Random Initialization
{
Weights (between input and hidden layer): wih, where $1 \leq i \leq Q$, $1 \leq h \leq H$
Weights (between hidden and output layer): whj, where $1 \leq h \leq H$, $1 \leq j \leq Z$
Biases: $\theta_h$
}
Set learning rate (η), error tolerance (τ) and momentum (α) as desired.
For each associated NN {
Repeat {
Compute signals on forward pass as follows:
```

$$z_h^k = \sum_{i=1}^{Q} x_i^k w_{ih}^k + \theta_h^k \tag{23.1}$$

$$\delta(z_h^k) = \frac{1}{1 + \exp(-z_h^k)} \tag{23.2}$$

$$y_j^k = \sum_{h=1}^{H} w_{hj}^k.\delta(z_h^k) + \theta_j^k \tag{23.3}$$

$$\delta(y_j^k) = \frac{1}{1 + \exp(-y_j^k)} \tag{23.4}$$

Compute errors/deltas at output neurons:

$$error_j^k = (d_j^k - \delta(y_j^k))\delta'(y_j^k) \tag{23.5}$$

$$\Delta w_{hj}^k = \eta.error_j^k.\delta(y_j^k) \tag{23.6}$$

Compute errors/deltas at hidden neurons:

$$error_h^k = (\sum_{j=1}^{Z} error_j^k.w_{hj}^k)\delta'(z_h^k) \tag{23.7}$$

$$\Delta w_{ih}^k = \eta.error_h^k.x_i^k \tag{23.8}$$

Update weights as follows:

$$w_{ih}^{k+1} = w_{ih}^k + \Delta w_{ih}^k + \alpha.\Delta w_{ih}^{k-1} \tag{23.9}$$

$$w_{hj}^{k+1} = w_{hj}^k + \Delta w_{hj}^k + \alpha.\Delta w_{hj}^{k-1} \tag{23.10}$$

Update biases as follows:

$$\Delta\theta_h^k = \eta.error_h^k \tag{23.11}$$

$$\theta_h^{k+1} = \theta_h^k + \Delta\theta_h^k + \alpha.\Delta\theta_h^{k-1} \tag{23.12}$$

Compute mean square error on entire training patterns:

$$Er_k = \frac{1}{2}\sum_{j=1}^{Z} (d_j^k - \delta(y_j^k))^2 \tag{23.13}$$

$$Er_{av} = \frac{1}{N}(\sum_{k=1}^{N} Er_k) \tag{23.14}$$

```
} until ($Er_{av} < \tau$)
}
```

Output:

Trained associated neural network.

Trained neural network is further responsible for classification of biological data. Such HI-model is more prominent for biological data learning as they are trained in natural way.

23.3.5 Ant Colony Optimization (ACO).

Ant Colony Optimization (ACO) is a general search method based on population for the solution of difficult complex problems that is inspired by the pheromone trail laying behaviour of real ant colonies. To solve the problem, an ant can be regarded as a simple computing agent that builds, for example, a solution. States are regarded as partial problem solutions. The identity of small subsets of extremely predictive and biologically relevant genes in bioinformatics is very tedious. Ant Colony Algorithm (ACA) is an algorithm that contains previous information and enables the efficiency of a sample space search to allow the algorithm to identify small subsets of very important and biological genes without the need for an extensive preselection of features when applied to multiple high-dimensional datasets. Thus, it is obvious that ant colony optimization and ant colony algorithms both are very useful for bioinformatics.

Various recent researches demonstrate that ant colony optimization is also very applicable in various bioinformatics problems. In (Kleinkauf et al. 2015), ACO meta-heuristics is employed for RNA and superior performance is obtained in biological datasets. For the 2D and 3D hydrophobic polar protein folding problem, ACO algorithm is applied (Shmygelska and Hoos 2005). In (Do Duc et al. 2018), authors successfully designed an efficient Ant Colony Optimization algorithm for protein structure prediction. In other research (Wu 2020), ACO is also useful for DNA sequence alignment.

23.3.6 Particle Swarm Optimization

The PSO lists social behaviours, the strategies used for locating roosting sites, food sources or other suitable habitat in bird flocking or fishing schools. PSO is originally developed by Kennedy and Eberhart (Kennedy and Eberhart 1995). In the search space of the given problem, the PSO algorithm simultaneously implements several candidate solutions. Each solution for each candidate is obtained by optimizing the objective function and evaluating the fitness of each algorithm. The PSO algorithm initially randomly chooses candidate solutions in the space of search. Each candidate solution can be taken as a “flying” particle in the fitness scene when seeking the maximum or minimum objective function.

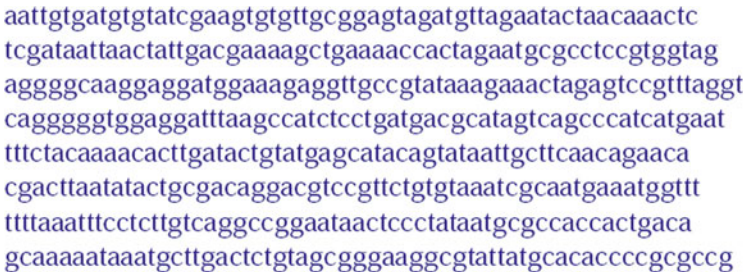
In (Das et al. 2008), there is wide demonstration of applicability of swarm intelligence in bioinformatics. In other research, author combined particle swarm optimization and simulated annealing for solving protein multiple sequence alignment problem (Chaabane 2018). PSO and gene scoring strategy is employed for hybrid gene selection in (Han et al. 2019).

23.4 Case Study

In this section, we discussed experimental analysis of soft computing techniques on two bioinformatics problem, i.e. Promoter gene sequence DNA problem and primate splice junction gene sequence problem. It has been demonstrated that soft computing techniques perform better than the conventional techniques. Moreover, hybridization of two or more technique is producing more prominent results rather than single technique.

23.4.1 Promoter Gene Sequence (DNA) Problem

Promoter gene sequence problem (Harley and Reynolds 1987; Towell et al. 1990) contains 106 instances with 59 attributes in each. Basically, promoters initiate the process of gene expression. The problem is to predict the member/non-member of class of sequences with biological promoter activity. The dataset contains non-numeric domain of attributes. The attributes are one of the ‘a’, ‘g’, ‘t’ and ‘c’ (a = Adenine, b = Guanine, t = Thymine and c = Cytosine). The snapshot of DNA gene sequence is shown in Fig. 23.5. The class distribution in this dataset is 50% for each, i.e. 53 instances for positive class and 53 instances belong to negative class. In



```
aattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaatactaacaaactc
tcgataattaactattgacgaaaagctgaaaaccactagaatgcgcctccgtggtag
aggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgtttaggt
caggggggtggaggatttaagccatctcctgatgacgcatagtcagcccatcatgaat
tttctacaaaacacttgatactgtatgagcatacagtataattgcttcaacagaaca
cgacttaatatactgcgacaggacgtccgttctgtgtaaatcgcaatgaaatggttt
ttttaaatttcctcttgtcaggccggaataactccctataatgcgccaccactgaca
gcaaaaataaatgcttgactctgtagcgggaaggcgtattatgcacaccccgcgccg
```

Fig. 23.5 Promoter DNA gene sequences

Table 23.1 Comparative analysis for promoter gene problem

Soft computing methods	Accuracy (%)	
	Training set	Test set
Fuzzy c mean clustering (FCM)	56.23	26.75
Evolutionary fuzzy clustering	73.89	53.43
Neural network (NN)	100	92.67
EFCMD-FMNN	100	95.81
Hybrid intelligent model	**100**	**96.22**

Bold values refer to the best results obtained by hybrid intelligent model as compared to other existing techniques.

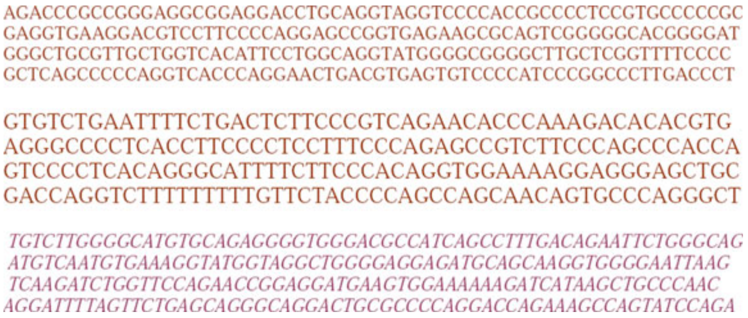
this work, we have selected 50% data for training set and rest 50% data for test set. Table 23.1 summarizes the results obtained for promoter gene dataset. It has been seen that NN yields far better results than Fuzzy clustering. The considered MLP consists of two NNs of size 59-18-1 for each. It yields maximum accuracy 92.67% with 2 × 1099 learning parameters on 20,000 learning cycles. HI-model again yields best accuracy of 96.22% just at C = 3 with 3 × 498 learning parameters in associated NN structure of 59-8-2(3), on 4000 average learning cycles.

It has been observed that on selecting more than two members in a cluster, the performance decreases. The performance also degrades when selecting more or less than three clusters. The maximum accuracy is attained at C = 3 and MCM = 2. On comparing with other exiting work, we found that maximum 93.33% accuracy is achieved by decision tree technique reported in Noordewier et al. (1991). Hence, it has been strongly investigated that soft computing methods are well applied and yield promising results for this bioinformatics based data.

23.4.2 Primate Splice-Junction Gene Sequence Problem

The primate dataset contains a total of 3190 occurrences with 62 attributes in each. Splice junctions are considered as the points on DNA sequence. During the process of protein creation in higher entity, superfluous DNA is removed in such junctions (Noordewier et al. 1991). The main problem is to find the boundaries between exons (the part of DNA sequence retained after splicing) and introns (the part of DNA sequence that are spliced out) in given DNA gene sequence. Hence, the problem contains three classes. First is intron–exon (IE) boundary which is sometimes called donors. Second is exon–intron (EI) boundary which is sometimes called acceptors. Third and last class belongs to neither donors nor acceptors (Neither). The class distribution is as follows: 767 instances for (IE), 768 instances for (EI) and 1655 instances for (Neither) class. The snapshot of this gene sequence is shown in Fig. 23.6. The gene sequences shown in smaller font are the examples of acceptor class while gene sequence shown in larger font implies donor class. The italicized font shows the example of neither class.

For splice gene data, we have selected training and testing data as division made in Bower and Bolouri (2004) for making a fair comparison. Therefore, training set



AGACCCGCCGGGAGGCGGAGGACCTGCAGGTAGGTCCCCACCGCCCCTCCGTGCCCCCGC
GAGGTGAAGGACGTCCTTCCCCAGGAGCCGGTGAGAAGCGCAGTCGGGGGCACGGGGAT
GGGCTGCGTTGCTGGTCACATTCCTGGCAGGTATGGGGCGGGGCTTGCTCGGTTTTCCCC
GCTCAGCCCCCAGGTCACCCAGGAACTGACGTGAGTGTCCCCATCCCGGCCCTTGACCCT

GTGTCTGAATTTTCTGACTCTTCCCGTCAGAACACCCAAAGACACACGTG
AGGGCCCCTCACCTTCCCCTCCTTTCCCAGAGCCGTCTTCCCAGCCCACCA
GTCCCCTCACAGGGCATTTTCTTCCCACAGGTGGAAAAGGAGGGAGCTGC
GACCAGGTCTTTTTTTTGTTCTACCCCAGCCAGCAACAGTGCCCAGGGCT

TGTCTTGGGGCATGTGCAGAGGGGTGGACGCCATCAGCCTTTGACAGAATTCTGGGCAG
ATGTCAATGTGAAAGGTATGGTAGGCTGGGGAGGAGATGCAGCAAGGTGGGGAATTAAG
TCAAGATCTGGTTCCAGAACCGGAGGATGAAGTGGAAAAAAGATCATAAGCTGCCCAAC
AGGATTTTAGTTCTGAGCAGGGCAGGACTGCGCCCCAGGACCAGAAAGCCAGTATCCAGA

Fig. 23.6 Splice-junction gene sequences showing acceptor, donor and neither classes

Table 23.2 Comparative analysis for splice gene problem

	Accuracy (%)	
Soft computing methods	Training set	Test set
Fuzzy c mean clustering (FCM)	33.86	21.95
Evolutionary fuzzy clustering	56.23	51.92
Neural network (NN)	95.60	90.12
SVM	92	92
Hybrid intelligent model	**99.08**	**96.60**

Bold values refer to the best results obtained by hybrid intelligent model as compared to other existing techniques.

contains 2000 instances (500 for acceptor, 500 for donor and 1000 for none class). Test set contains rest 1190 instances (267 for acceptor, 268 for donor and 655 for none class). The comparative analysis for splice gene sequence is shown in Table 23.2. The considered MLP consists of three NNs of each size 62-20-1. It achieves accuracy up to 90.12% with 3 × 661 learning parameters on 30,000 average learning cycles. HI-model obtains maximum accuracy of 95.6% with only 3 × 531 learning parameters in associated NN of structure 62-8-3(3) on average 6000 learning cycles.

On comparison with techniques reported in (Noordewier et al. 1991), we found that decision tree technique yields 93.6% accuracy and SVM obtains 92% accuracy. However, assoDNA method in (Noordewier et al. 1991) yields 96.1% accuracy. It is clear that HI-model yields better accuracy than these techniques. Hence, it has been again proved that HI-model emerged as best computing model over other mentioned techniques even for highly complex gene sequence prediction.

23.5 Conclusions

Bioinformatics is a vast field of research with wide applicability of soft computing techniques. The main characteristics of soft computing based models and techniques reported here include the use of synergistic integration of soft computing techniques, dealing with uncertainty, overlapping boundaries, easily adaptive and efficient classification for bioinformatics. Theoretical and experimental justifications pose the evidence of soft computing techniques for bioinformatics correctly. The crucial aspect of outperformance of soft computing techniques over statistical based techniques is the compatibility of characteristics and tuning of aforementioned techniques with demand of biological data used for bioinformatics.

References

Aerts S, Van Loo P, Moreau Y et al (2004) A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. Bioinformatics 20(12):1974–1976

Barbosa F, Devito K, Filho W (2009) Using a neural network for supporting radiographic diagnosis of dental cares. Int J Appl Artif Intell 23(9):872–882

Bezdek JC (1992) On the relationship between neural networks, pattern recognition and intelligence. Int J Approx Reason 6:85–107

Bhattacharjee D, Basu DK, Nasipuri M, Kundu M (2010) Human face recognition using fuzzy multilayer perceptron. Soft Comput 14:559–570

Bower JM, Bolouri H (eds) (2004) Computational modeling of genetic and biochemical networks. Cambridge, MA, MIT Press

Carter RJ, Dubchak I, Holbrook SR (2001) A computational approach to identify genes for functional RNAs in genomic sequence. Nucleic Acids Res 29(19):3928–3938

Chaabane L (2018) A hybrid solver for protein multiple sequence alignment problem. J Bioinforma Comput Biol 14(4):185–200

Chiesa M, Maioli G, Colombo GI et al (2020) GARS: genetic algorithm for the identification of a robust subset of features in high-dimensional datasets. BMC Bioinform 21:54

Das S, Abraham A, Konar A (2008) Swarm intelligence algorithms in bioinformatics. In: Kelemen A, Abraham A, Chen Y (eds) Computational intelligence in bioinformatics. Studies in computational intelligence, vol 94. Springer, Berlin

Deb K (2001) Multi-objective optimization using evolutionary algorithms. Wiley, New York

Dembélé D, Kastner P (2003) Fuzzy C-means method for clustering microarray data. Bioinformatics 19(8):973–980

Do Duc D, Thai Dinh P, Thi Ngoc Anh V, Linh-Trung N (2018) An efficient ant Colony optimization algorithm for protein structure prediction. 12th international symposium on medical information and communication technology (ISMICT), Sydney 1–6

Fazendeiro P, de Oliveira JV (2007) A semantic driven evolutionary fuzzy clustering algorithm. In: Proceedings of IEEE International Fuzzy Systems Conference. IEEE, New York, pp 1–6

Fogel GB, Corne DW (2003) Evolutionary computation in bioinformatics. Morgan Kaufmann, Burlington, CA

Gaxiola F, Melin P (2010) Modular neural networks for person recognition using segmentation and the iris biometric measurement with image pre-processing. In: International joint conference on neural networks (IJCNN). IEEE, New York, pp 2765–2771

Haddadnia J, Faez K, Ahmadi M (2003) A fuzzy hybrid learning algorithm for radial basis function neural network with application in human face recognition. Pattern Recogn 36(5):1187–1202

Han F, Tang D, Sun Y, Cheng Z, Jiang J, Li Q (2019) A hybrid gene selection method based on gene scoring strategy and improved particle swarm optimization. BMC Bioinform 20(Supp 8):1–13

Handl J, Knowles J (2007) An evolutionary approach to multiobjective clustering. IEEE Trans Evol Comput 11(1):56–76

Harley C, Reynolds R (1987) Analysis of E. Coli promoter sequences. Nucleic Acids Res 15:2343–2361

Hruschka ER, Campello RJGB, Freitas AA, de Carvelho ACPLF (2009) A survey of evolutionary algorithms for clustering. IEEE Trans Syst Man Cybern Part C 39(2):133–155

Huang S, Cai N, Pacheco P, Narandes S, Wang Y, Xu W (2018) Applications of support vector machine (SVM) learning in Cancer genomics. Cancer Genom Proteom 15(1):41–51

Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: Proceedings of IEEE international conference on neural networks, Piscataway, NJ, vol 4. IEEE, New York, pp 1942–1948

Kleinkauf R, Mann M, Backofen R (2015) antaRNA: ant colony-based RNA sequence design. Bioinformatics 31(19):3114–3121

Krallinger M, Erhardt RA, Valencia A (2005) Text-mining approaches in molecular biology and biomedicine. Drug Discov Today 10(6):439–445

Lancashire LJ, Lemetre C, Ball GR (2009) An introduction to artificial neural networks in bioinformatics- application to complex microarray and mass spectrometry datasets in cancer studies. Brief Bioinform 10(3):315–329

Larranaga P, Calvo B, Santana R et al (2006) Machine learnsing in bioinformatics. Brief Bioinform 7(1):86–112

Leondes CT (2003) An application of artificial neural networks to DNA sequence analysis. In: Leondes CT (ed) Computational methods in biophysics, biomaterials, biotechnology and medical systems. Springer, Boston, MA

Lu J, Yuan X, Yahagi T (2007) A method of face recognition based on fuzzy c- means clustering and associated sub-NNs. IEEE trans. Neural Netw 18(1):152–160

Maji P, Paul S (2017) Fundamentals of rough-fuzzy clustering and its application in bioinformatics. In: Pattern recognition and big data. World Scientific, Singapore, pp 513–543

Mathe C, Sagot MF, Schlex T et al (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30(19):4103–4117

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133

Mingoti SA, Lima JO (2006) Comparing SOM neural network with fuzzy c-means, K-means and traditional hierarchical clustering algorithms. Eur J Oper Res 174:1742–1759

Noordewier MO, Towell GG, Shavlik JW (1991) Training knowledge-based neural networks to recognize genes in DNA sequences advances in neural information processing systems, vol 3. Morgan Kaufmann, Burlington, MA

Ozbey Y, Ceylan R, Karlik B (2006) A fuzzy clustering neural network architecture for classification of ECG arrhythmias. Comput Biol Med 36:376–388

Saha I, Mukhopadhyay A, Maulik U (2009) Combining fuzzy clustering with ANN classifier for categorical data. In: IEEE International Advance Computing Conference, (IACC 2009). IEEE, New York, pp 44–49

Shmygelska A, Hoos HH (2005) An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. BMC Bioinform 6:30

Sio-Iong AO (2009) Hybrid intelligent regressions with neural network and fuzzy clustering advances in computational algorithms and data analysis. In: Ao SI, Rieger B, Chen SS (eds) Advances in computational algorithms and data analysis. Lecture notes in electrical engineering, vol 14. Springer, Dordrecht

Su ZX (2011) A hybrid fuzzy approach to fuzzy multi attribute group decision making. Int J Inf Technol Decis Making 10(4):695–711

Tampuu A, Bzhalava Z, Dillner J, Vicente R (2019) ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. PLoS One 14(9):e0222271

Tang B, Pan Z, Yin K, Khateeb A (2019) Recent advances of deep learning in bioinformatics and computational biology. Front Genet 10:1–10

Towell G, Shavlik J, Noordewier M (1990) Refinement of approximate domain theories by knowledge-based artificial neural networks. In: Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90). AAAI, Menlo Park, CA

Tripathi BK, Kalra PK (2010a) The novel aggregation function based neuron models in complex domain. Soft Comput 14(10):1069–1081

Tripathi BK, Kalra PK (2010b) High dimensional neural networks and applications. In: Intelligent autonomous systems, SCI, vol 275. Springer, Berlin, pp 215–233

Tripathi BK, Kalra PK (2011a) On the learning machine for three dimensional mapping. Neural Comput Appl 20:105–111

Tripathi BK, Kalra PK (2011b) On efficient learning machine with root-power mean neuron in complex-domain. IEEE Trans Neural Network 22(5):727–738

Tripathi BK, Kalra PK (2011c) Complex generalized-mean neuron model and its applications. Appl Soft Comput 11(01):768–777

Wang Z, Jiang M, Hu Y, Li H (2012) An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors. IEEE Trans Inf Technol Biomed 16(4):691–699

Wu T (2020) DNA sequence alignment based on ants' colony algorithm. AIP Conf Proc 2208:020030

Yang X, Zhang G, Lu J, Ma J (2011) A kernel fuzzy c-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises. IEEE Trans Neural Network 19(1):05–115

Yuan B, Klir GJ, Stone JF (1995) Evolutionary fuzzy c-means clustering. In: Proceedings of International Conference on Fuzzy System. ACM, New York, pp 2221–2226

Zadeh LA (1994) Fuzzy logic, neural networks and soft computing. Commun ACM 37(3):77–84

Zamani M, Kremer SC (2013) Neural networks in bioinformatics. In: Bianchini M, Maggini M, Jain L (eds) Handbook on neural information processing. Intelligent systems reference library, vol 49. Springer, Berlin

Zhang D, Peng H, Zhou J, Pal SK (2002) A novel face recognition system using hybrid neural and dual Eigenspaces methods. IEEE Trans Syst Man Cybern A 32(6):787–793

Zio E, Baraldi P (2005) Evolutionary fuzzy clustering for the classification of transient in nuclear components. Prog Nucl Energy 46(3–4):282–296