Asheesh Shanker *Editor*

Bioinformatics: Sequences, Structures, Phylogeny

Springer

Bioinformatics: Sequences, Structures, Phylogeny

Asheesh Shanker
Editor

Bioinformatics: Sequences, Structures, Phylogeny

Springer

Editor
Asheesh Shanker
Department of Bioinformatics
Central University of South Bihar
Gaya, Bihar, India

ISBN 978-981-13-1561-9 ISBN 978-981-13-1562-6 (eBook)
https://doi.org/10.1007/978-981-13-1562-6

Library of Congress Control Number: 2018956286

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

Preface

Bioinformatics has established itself as an independent discipline and primarily deals with sequence and structure data. This interdisciplinary subject has revolutionized the biological research with potential application in almost all areas including agriculture and medicine. Since my joining as an Assistant Professor (Bioinformatics) in 2004 at Banasthali Vidyapith, Rajasthan, India, where I taught and guided several students, the field of Bioinformatics has changed a lot. The students to whom once I taught now became independent researchers. However, they still remember that which I usually tried to convey last semester students "Always try to serve your roots". With a thought to show the contribution of my decade-old network in the field of Bioinformatics, I planned to edit this book. Starting with the introduction to the field, this book covers sequence, phylogenetic, and structure analysis. I hope it will be beneficial to both a beginner and an experienced researcher.

I am thankful to all the authors for their generosity and willingness to contribute book chapters. I am grateful to the reviewers for their time and efforts in providing suggestions for the improvement of the chapters. I am also thankful to the staff at Springer Nature for their support.

Earlier, I authored/co-authored three books; however, this edited book is the most difficult project. I acknowledge Shefali, Akshita, and Parv for their constant love, support, and cheerful moments.

Once again, I thank one and all who make this book a successful endeavour.

Thank you very much.

Gaya, Bihar, India
Asheesh Shanker

Contents

About the Editor

Asheesh Shanker is Associate Professor and Coordinator of the Bioinformatics Programme at the Centre for Biological Sciences, Central University of South Bihar, India. Prior to that, he served as Assistant/Associate Professor (2004–2015) at the Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan, India, and he has a decade of teaching and research experience. His research focuses on data mining, database and server development, evolutionary analysis, computational genomics, and structure analysis. He is a recipient of the prestigious Indo-Swiss Joint Research Programme Fellowship from the University of Lausanne, Switzerland. He has been a convener/organizing secretary for several national bioinformatics workshops, served as a member of organizing committees of international conferences, and delivered invited talks. He is also a reviewer for a number of peer-reviewed journals. He has published more than 50 research articles in leading national and international journals and authored/co-authored 3 books in the field of bioinformatics, which are popular among students.

Intellectual Property Rights and Bioinformatics: An Introduction

1

Shilpa and Umang Gupta

1.1 Introduction

Bioinformatics is a branch of science that works at an intersection of biology, information technology, mathematics, and chemistry. It deals with the analysis of biomolecules like DNA, RNA, and proteins, biological databases, and related software to analyze the data. The increasing investments in the area of bioinformatics have increased the need of adequate intellectual property laws which is one of the key issues in any emerging area. Intellectual property rights (IPRs) are the legal rights given to the inventor or creator to protect his creation for certain period of time. These include patents, copyrights, designs, trademarks, trade secrets, and geographical indications. Intellectual property (IP) protection is one of the very important measures to be seen for economic growth and advancement in any technological field. It drives the innovation and advancement in the competitive society (Rogers 1998). IP also plays an important role in bridging the "valley of death" by providing access to finance and infrastructure.

For intellectual property protection, bioinformatics is limited to patents, copyrights, and trade secrets. This chapter discusses how different bioinformatics components relate to the intellectual property law system specifically in the context of patent and copyright law. Moreover, the requirements and limitations of intellectual property in the field of bioinformatics and the patent trends in bioinformatics are discussed.

Shilpa (✉) · U. Gupta
Academy of Scientific and Innovative Research, CSIR-National Institute of Science Technology and Development Studies, New Delhi, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_1

1.2 Intellectual Property Rights

Man with his skills, acquired through proper training and practice over the time, is continuously involved in the development of either something which addresses the challenges that exist at the time of development of that product or something which emotionally pleases other persons. These creations in the earlier times fell in the public domain as public goods and anybody could use them, or imitate them, without any royalty paid to the original creator. With the passage of time, the importance and value of such creations was realized, and a threat was perceived against their development. Further, commercialization of these creations also started, and a need arose for the development of a mechanism which protects the rights of creators of such intellectual properties. Some of the very basic uses of IPR include the following: intellectual property creates employment by promoting research and development; it facilitates the development of solutions to existing challenges; it drives economic growth and competitiveness among innovators; it encourages incentivization to the innovators and entrepreneurs; and when properly enforced, it also protects the rights of consumers.

These rights are (or need to be) embraced by all sectors of technology and industry. A proper policy framework is needed for countries consisting of all types of small, medium, and large industries. Obtaining intellectual property protection for bioinformatics and related technologies is a critically important process.

1.2.1 Types of Intellectual Property

1.2.1.1 Patents

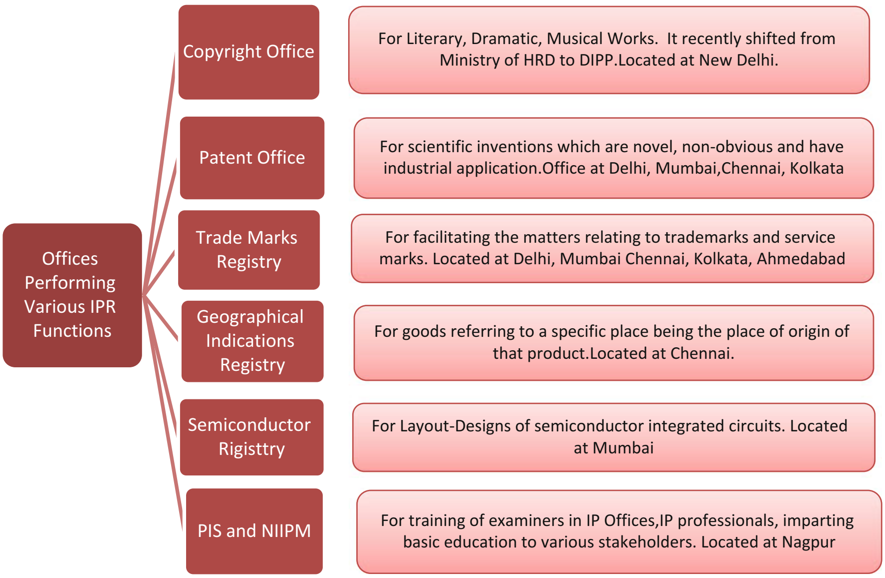
Patent is a document that discloses information to the public. In exchange of disclosing the invention to the public, the government grants the rights to the inventor to exclude others from making, using, or selling the invention claimed in the patent generally for a period of 20 years (this varies from geographic location). Patents are territorial rights and are confined to those regions only in which they have been granted. To obtain rights in a country, it is necessary to apply for patents and follow the procedures as required by the patent regime in that country. However, with the development of more sophisticated systems like Patent Cooperation Treaty (PCT), the applicants can file a single application for many countries. Further, after obtaining patent right on his invention, the patent owner may issue license to another person who provides the right to use his patented invention or sell the invention on mutually agreed terms.

1.2.1.2 Copyrights

Copyright is a form of intellectual property protection given to content creators through the assignment of specific rights to works such as musical compositions, films, software programs, paintings, expression of creative ideas, and other literary work. Although copyright is available on a work by virtue of its creation, still it should be applied for protection by following proper laws. The main purposes of copyright are (1) to promote the progress of science and culture, (2) to provide monetary incentives to copyright holders for their works, and (3) to encourage moral principles of respect and trust. It protects the unauthorized reproduction or exploitation of protected materials. Creators often sell the copyrights to individuals or companies which are best able to market their work in return for royalty. Copyrights in India are valid for life span of the creator and for an additional 60 years.

1.2.1.3 Trade Secrets

A trade secret could be any sort of data or information relating to the business and is not known to the public because owner wants to keep it confidential. It confers some sort of economic benefit on its holder over the competitors. The subject matter of trade secrets may include sales methods, distribution methods, advertising strategies, and manufacturing processes.[1] This type of protection is generally valuable for business corporations to enable them to recoup their investments. It can also be used as an alternative of patent by start-ups which do not possess enough economic resources for patenting. Trade secrets can be protected for an unlimited period of time but a considerable amount of secrecy must exist. Figure 1.1 shows the information of various Indian offices involved in IPR-related operations.



For Literary, Dramatic, Musical Works. It recently shifted from Ministry of HRD to DIPP.Located at New Delhi.

For scientific inventions which are novel, non-obvious and have industrial application.Office at Delhi, Mumbai,Chennai, Kolkata

For facilitating the matters relating to trademarks and service marks. Located at Delhi, Mumbai Chennai, Kolkata, Ahmedabad

For goods referring to a specific place being the place of origin of that product.Located at Chennai.

For Layout-Designs of semiconductor integrated circuits. Located at Mumbai

For training of examiners in IP Offices,IP professionals, imparting basic education to various stakeholders. Located at Nagpur

Fig. 1.1 Indian offices for performing operations of various forms of IPR

[1] http://www.wipo.int/sme/en/ip_business/trade_secrets/trade_secrets.htm

1.3 The Importance of Protecting Intellectual Property

IPR is a legal right given to the inventor or creator of original creation developed through his intellect such as artistic, literary, technical, or scientific creation to protect his creation for a certain period of time (WIPO[2]). These rights enable the original creator to prevent competitors and others from using his creation for their own profit without his consent. In lieu of the rights granted over the creation, the creator is under the obligation to disclose his invention to the government in public interest. It is well established that intellectual property plays an important role in the present knowledge-based economy (Alikhan 2000). With the invigorating investments in research and development, the stakes of innovators in knowledge creation have become very high. There are two fundamental principles for the development of IPR: (1) to provide proper incentives to the inventors and (2) economic and social welfare of people of the state. Therefore, it is important to protect these innovations to recoup costs associated with it (Elliot 2005; Commission on IPR 2002). Moreover, IPR plays an important role not only in the economic development of the nations but also in many other aspects of development. The contribution of IPR framework can be observed in different contexts including networking of the organizations, trust building, building of scientific infrastructure, and development of market rapport, standards development, and sustainable development of technologies. Thus, IPRs promote healthy competition in the markets and thereby support technological, social, and economic development of a country. A number of instances regarding patent disputes enable us to understand the value and importance of intellectual property protection. For example, the verdict in the case between Monsanto and DuPont awarded $1 billion[3] to the patent owner Monsanto, and ultimately technology deal was agreed to between the companies wherein the DuPont had to pay more than $1.75 bn to Monsanto to make use of its patented technology.[4] Similarly, Biogen agreed to pay Forward Pharma $1.25 billion to secure an irrevocable license to all the intellectual property owned by Forward Pharma. Companies use IPR not only for economic compensations and protection of their technologies, but also for their portfolio development, and as a strategy for increasing their brand value. In the India-US Basmati Rice Dispute, RiceTec was forced to give up the title of its patent which had implications related to biopiracy (Jayaraman 1998). In this case, the importance of two major forms of IPR were discussed, i.e., patents and geographical indications. Dr. Vandana Shiva, director of a Delhi-based research foundation which monitors issues involving patents and biopiracy, said that, “the main aim for obtaining the patent by RiceTec Inc. was to trick the consumers in believing there is no difference between spurious Basmati and

[2] http://www.wipo.int/edocs/pubdocs/en/intproperty/450/wipo_pub_450.pdf

[3] Monsanto Company and Monsanto Technology LLC v. E.I. Du Pont de Nemours and Company and Pioneer Hi-Bred International, Inc., Docket No. 2013–1349 (May 9, 2014).

[4] http://www.reuters.com/article/us-monsanto-dupont-gmo-idUSBRE92P0IK20130326

real Basmati" (Mukherjee 2008). A large number of farmers cultivating Indian Basmati Rice would have been affected by this judgment because "Indian farmers export $250 million in Basmati every year and USA is a target market."[5] India would have lost US, European Union, and Middle East import market leading to huge economic loss (Mukherjee 2008). These IPR disputes require a huge drain of resources, and their results impact a considerable size of population; therefore, this position of intellectual property protection can't be overlooked.

1.4 Bioinformatics

Biological data is growing at an exponential rate. For example, number of sequences in GenBank increased from 606 (release 3; Dec. 1982) to 201,663,568 (release 220; June 2017).[6] GenBank has been doubling in size about every 15 months. Similarly, entries in UniProtKB/TrEMBL reached more than 90,050,000 in 2017.[7] In addition to these sequences, data also consists of gene expression records, protein structures, and details on how these structures interact with one another. This exponential growth in the amount of different types of biological data has necessitated the use of information technologies for cataloging and retrieval of this data. Sophisticated technologies (software and hardware) play a critical role in the analysis of this complex data, and the advancement in these technologies has further established this field (Shanker 2012; Gaudet et al. 2011).

Bioinformatics works at an intersection of various branches of science like genomics, biotechnology, information technology, mathematics, and chemistry with applications in various different areas like drug discovery (Blundell et al. 2006), biomarker development (Crichton et al. 2010), forensics (Bianchi and Lio 2007), plant sciences (Rhee et al. 2006), and molecular medicine (Maojo and Kulikowski 2003). Bioinformatics is defined in many different ways as its scope varies with different application areas (Box 1.1).

Box 1.1: Definitions of Bioinformatics

Oxford English Dictionary

"The science of collecting and analyzing complex biological data such as genetic codes."

National Institute of Health

(continued)

[5] Ibid.

[6] https://www.ncbi.nlm.nih.gov/genbank/statistics/

[7] http://www.uniprot.org/statistics/TrEMBL

Box 1.1 (continued)

Bioinformatics is “research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.”

National Center for Biotechnology Information

Bioinformatics is the “field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.”

International Patent Classification (IPC) for Bioinformatics (G06F 19/10)

Bioinformatics, “i.e., methods or systems for genetic or protein-related data processing in computational molecular biology (in silico methods of screening virtual chemical libraries; in silico or mathematical methods of creating virtual chemical libraries).”

National Agriculture Library, US Department of Agriculture

A “field of biology concerned with the development of techniques for the collection and manipulation of biological data, and the use of such data to make biological discoveries or predictions. This field encompasses all computational methods and theories applicable to molecular biology and areas of computer-based techniques for solving biological problems including manipulation of models and datasets.”

Immense opportunities in this area have driven various government initiatives and funding from private sources which is continuously increasing the market size. Bioinformatics is generating revenue of more than a billion dollars every year globally.[8] The “global bioinformatics market which accounted for $4.2 billion in 2014 and is poised to reach $13.3 billion by 2020”[9] is among the fastest-growing areas. The application of bioinformatics in drug development is expected to reduce

[8] Bioinformatics Market – Global Industry Analysis, Size, Growth Trends, Share, Opportunities, and Forecast

http://www.prnewswire.co.uk/news-releases/bioinformatics-market-is-expected-to-reach-revenue-of-128-billion-globally-by-2020%2D%2D-allied-market-research-260945081.html

[9] Bioinformatics Market by Sector (Molecular Medicine, Agriculture, Forensic, Animal, Research & Gene Therapy), Product (Sequencing Platforms, Knowledge Management & Data Analysis) &

the annual cost of developing new drugs by 33% and time for drug discovery by 30% (Jagadish 2013). Various governments and private sectors are investing heavily in this area for the exploitation of biological data and developments in the field of medicines (Shook 2002). It is now an important commodity and will benefit both public and private sectors.

Various components of bioinformatics demand different forms of IPR to get protected. Bioinformatics comprises of (1) biological sequences such as DNA, RNA, and protein, (2) biological databases in which this information is organized, and (3) software designed to analyze data.

1.4.1 Biological Sequences

Bioinformatics is focused to study the DNA, RNA, and proteins which altogether reveal the secret of biology. Each of them are described using simple codes, i.e., A, C,G,T for DNA; A,C,G,U for RNA; and 20 different amino acids for proteins. DNA is the hereditary material which passes information to generations. Coding regions of DNA, genes, transcribe to RNA which further translates to proteins. RNA acts as an intermediator to translate a DNA sequence into an amino acid sequence. It is the amino acid sequence that determines the trait. This is called central dogma of life; DNA − > RNA − > Protein (Crick 1970). Sequencing of these molecules helps to find out the arrangement of constituting units either in case of DNA, RNA, or protein. Sequencing enables modeling of protein structure and function prediction. It also allows extracting knowledgeable information from the data accumulated in molecular biology.[10]

In the USA, with the settlement of patent issue of genetically modified bacteria in the case of Diamond v. Chakrabarty (447 US 30, 1980), the granting of patents to genes was considered by US court. Patents on biological materials became accepted and the Supreme Court justice permitted the patenting of genetically modified bacteria, and since then the US Patent and Trademark Office (USPTO) has permitted the patenting of biological molecules which are isolated and purified from their natural environment (Hultquist et al. 2002). However in 2013, the US Supreme Court overruled that decision and now only cDNA is patentable as it is artificial material. The USPTO now only includes DNA, RNA, and proteins as patentable compositions as DNA and RNA are composed of nucleotides and proteins are composed of amino acids.[11] Contrasting to this, in Europe, the patentability of

Application (Genomics, Proteomics & Metabolomics) – Global Forecast to 2020.Report code – BT 3321.

http://www.marketsandmarkets.com/Market-Reports/bioinformatics-39.html

[10] http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp2_1.shtml

[11] http://www.squirepattonboggs.com/~/media/files/insights/publications/2013/08/us-supreme-court-holds-that-isolated-human-dna-i__/files/ussupremecourtholdsthatisolatedhumandnaisnotpate__/fileattachment/ussupremecourtholdsthatisolatedhumandnaisnotpate__.pdf

isolated genes is expressly accepted in European law after the implementation of the order on the Legal Protection of Biotechnological Inventions (EU Biotechnology Directive[12]) in 1998.

1.4.2 Biological Databases

A database is a storage or collection of information that is organized so that it can be easily accessed and managed. In scientific perspective, these are collection of biological information from scientific experiments, high-throughput screening, and computational analysis (Altman 2004). It includes experimental study of genomics, proteomics, metabolomics, microarray gene expression, etc.

An important feature of biological databases is the requirement of knowledge discovery like identification of links between different chunks of information. This became possible by increasing sophistication of database management systems and with increasing capability to manage large quantities of data and handling complex operations in an efficient manner which is another important reason for the growth of bioinformatics (Eriksson[13]). The trend of biological databases started after the first insulin protein was sequenced in 1956. The first database, the Protein Data Bank (PDB), a database of protein structures compiled in 1971, archived only 7 structures in the beginning and has now grown to more than 134,000 entries. In 1986 another consolidated database, SWISS-PROT,[14] was developed. Currently, several biological databases have been developed, and some of the widely used include DDBJ, EMBL, GenBank, and PIR. These databases comprise not only biological data but also sophisticated query services for data analysis.

In Europe two main forms of protection are available for databases – (1) copyright protection for the structure and form of databases and (2) sui generis right for the contents of database (Chang and Zhu 2010). The main drawback with copyright protection for databases was that the contents of the databases can be copied or reorganized without consent of the inventor. The sui generis right provided by Europe deals with this deficiency in the copyright law. There is no sui generis law on database protection in the USA.

The guidelines (1996) of USPTO state that if the database is merely a "data structure" or "nonfunctional descriptive material," it is not patentable, while the approach of creating the database may constitute a patentable process as it is application of algorithm which provides some concrete results. Therefore, in the USA the process of creating database as database itself consists of numerous applications that are patentable. Similar interpretation is also possible in Indian patent law system.

[12] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31998L0044:EN:HTML

[13] http://www.deroneriksson.com/tutorials/bioinformatics/biological-database-integration-computer-technology

[14] http://vle.du.ac.in/mod/book/view.php?id=8928&chapterid=12678

1.4.3 Biological Software

Another sub-discipline of bioinformatics is to develop systems and programs to utilize the information in these databases and extract the knowledge. Software engineers have developed different programs to analyze this information. For example, BLAST[15] (Basic Local Alignment Search Tool) compares sequences for similarity by aligning the two sequences and then calculating a similarity score. This is useful for predicting the function of a gene or protein which is previously unknown or to draw evolutionary relationships (Feng et al. 2000). Some other widely used softwares include BioPerl, EMBOSS, and BioJava.

Copyright protection is a way of protecting software in most of the countries. In the USA, software is protected as literary work under the copyright act. In Europe computer program is protected as literary work under the European Union Computer Programs Directive 2009/24/EC. In India also, bioinformatics software can be protected as computer program under Section 2 of the Copyright Act, 1957. Copyright protection has some limitations as the protection is available for the original expression of ideas while a change in the aspects of source code and object code provides a scope for an independent copyright.

Another type of available protection is through patents. In the USA, computer program is patentable if it meets the legal requirements for all patentable inventions (USPTO 1996). European approach toward patenting a software invention is different than the USA. To be eligible for patent protection in Europe, an algorithm should solve a technical problem in a novel and nonobvious manner. According to Indian Patent Office (IPO), computer program is not patentable per se, but in combination with new hardware components a computer program can be considered patentable.[16]

Bioinformatics is experiencing an emerging phase where technology is advancing while returns on investments are low. Therefore, it becomes important for public research organizations and enterprises to protect the inventions and innovations with intellectual property rights (Table 1.1). With further advancements in research and development in this field, more sophisticated tools are being developed. In such a scenario, there is no indifference in expecting a legal framework to encourage such developments. This complements the need to discuss IPR in the field of bioinformatics leading to the maximum return on investments in the given setup.

[15] NCBI, BLAST: Basic Overview.
http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul_1.html

[16] Business Standard. 22 February 2016.
http://www.business-standard.com/article/economy-policy/no-patent-if-invention-lies-only-in-computer-programme-says-indian-patent-office-116022200875_1.html

Table 1.1 IPR protection in bioinformatics in the USA, Europe, and India

	USA		Europe		India	
	Copyright	Patent	Copyright	Patent	Copyright	Patent
Biological sequences	√	√Restrictive	√	√Restrictive	√	√
Databases	√	√	√Additionally sui generis right for contents	–	√	√
Software	√	√	√	√If it makes technical contribution	√	√Restrictive

1.5 Patent Trends in Bioinformatics

In 2011, the World Intellectual Property Organization (WIPO) has introduced new International Patent Classification (IPC) search codes to cover patents in bioinformatics. The search methodology for extracting patents granted in bioinformatics is based on this IPC class search. USPTO and European Patent Office (EPO) were used to analyze the patent trends. There is an increasing gap in terms of number of patents granted in USPTO and EPO with 1406 patents granted in USPTO while 261 patents in EPO from year 1991 to 2015 (Fig. 1.2).

The USA is far ahead in innovations as compared to Europe in bioinformatics as reflected from patent trends. There could be many reasons behind this like established market in the USA, availability of infrastructure and resources, mobilization of human resource, and liberal patent laws at the policy level.

Further, data was extracted in different IPC classes to understand the distribution of patents in sub-disciplines of bioinformatics. The patent distribution in different IPC classes shows that patents are filed in wide range of subject areas in bioinformatics including structural biology, gene expression, molecular simulations, machine learning, data visualization, and database development. The difference in preference of patents granted in different sub-disciplines of bioinformatics is also visible in USPTO and EPO (Table 1.2).

The USA and Europe have typically distinct approach for protecting software programs which is visible in their patent trends. The USA has maximum patents granted in machine learning, data mining, and biostatistics (G06F 19/24) with

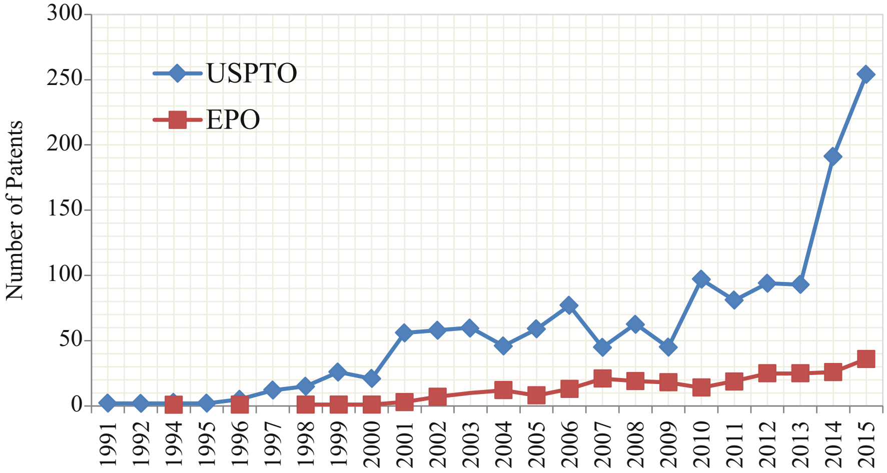


Fig. 1.2 Patent trends in bioinformatics (patents granted). (Source: Data Extracted from Thomson Innovation Patent Database)

Table 1.2 Patent distribution in different IPC classes (1991–2015)

IPC class	Description	Number of patents	
		USPTO	EPO
G06F 19/10	"Bioinformatics, i.e. methods or systems for genetic or protein-related data processing in computational molecular biology (in silico methods of screening virtual chemical libraries; in silico or mathematical methods of creating virtual chemical libraries)"	103	8
G06F 19/12	"For modelling or simulation in systems biology, e.g. probabilistic or dynamic models, gene-regulatory networks, protein interaction networks or metabolic networks"	157	26
G06F 19/14	"For phylogeny or evolution, e.g. evolutionarily conserved regions determination or phylogenetic tree construction"	25	1
G06F 19/16	"For molecular structure, e.g. structure alignment, structural or functional relations, protein folding, domain topologies, drug targeting using structure data, involving two-dimensional or three-dimensional structures"	271	72
G06F 19/18	"For functional genomics or proteomics, e.g. genotype-phenotype associations, linkage disequilibrium, population genetics, binding site identification, mutagenesis, genotyping or genome annotation, protein-protein interactions or protein-nucleic acid interactions"	325	69
G06F 19/20	"For hybridisation or gene expression, e.g. microarrays, sequencing by hybridisation, normalisation, profiling, noise correction models, expression ratio estimation, probe design or probe optimisation"	296	53
G06F 19/22	"For sequence comparison involving nucleotides or amino acids, e.g. homology search, motif or SNP [Single-Nucleotide Polymorphism] discovery or sequence alignment"	342	68
G06F 19/24	"For machine learning, data mining or biostatistics, e.g. pattern finding, knowledge discovery, rule extraction, correlation, clustering or classification"	348	50
G06F 19/26	"For data visualisation, e.g. graphics generation, display of maps or networks or other visual representations"	130	12
G06F 19/28	"For programming tools or database systems, e.g. ontologies, heterogeneous data integration, data warehousing or computing architectures"	222	32

Source: Data extracted from Thomson Innovation Patent Database

348 patents, while only 50 patents are granted in EPO. Another critical difference visible is in molecular structure (G06F 19/16).

The combination of biology, mathematics, and information technology makes bioinformatics a unique and knowledge-based area. It is one of the fastest-growing markets, and intellectual property protection could be important for ensuring the continuation of this growth. A key feature of this area is to analyze the increasing amount of biological information from the molecular biology laboratories. It is important to identify patterns from this data and convert the information into knowledge. This will further help to solve numerous life science issues like improvement in health and treatment of various diseases, increased agriculture production for food security, etc. For such essential objectives, a balanced IP protection system is extremely important which will further lead to the development of this domain. The IP protection system is still emerging in the area of bioinformatics by taking the understanding from the existing laws for biology, computers, and electronics.

Some of the bioinformatics components constitute patentable subject matter, while some are debatable with regard to patentability. The scope of protection varies among countries. This is also visible from patenting trends. Apart from this, copyright protection is also available but only for some elements of the software codes. It needs to be more liberal for open sharing of information. Copyright protection is significant for databases. Some other forms of protection are available like sui generis law in Europe. Intellectual property rights ensure the protection of ideas and have revolutionized science and technology which further motivate the system. Indian patent law does not provide any specific protection for bioinformatics components. The intellectual property protection system needs continuous revision as bioinformatics is an emerging area. It is also important to strike a balance between global and local intellectual property regulations by considering interests of creators, entrepreneurs, society, and other players in the field.

References

Alikhan S (2000) Socio-economic benefits of intellectual property protection in developing countries. World Intellectual Property Organization, Geneva

Altman RB (2004) Editorial: Building successful databases. Brief Bioinform 5(1):4–5

Bianchi L, Lio P (2007) Forensic DNA and bioinformatics. Brief Bioinform 8(2):117–128

Blundell TL, Sibanda BL et al (2006) Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philos Trans R Soc Lond B Biol Sci 361(1467):413–423

Chang J, Zhu X (2010) Bioinformatics databases: intellectual property protection strategy. J Intellect Prop Rights 15:447–454

Crichton DJ, Mattmann CA, Thornquist M, Anton K, Hughes JS (2010) Bioinformatics: biomarkers of early detection. Cancer Biomark 9(1–6):511–530

Crick F (1970) Central dogma of molecular biology. Nature 227:561–563

Elliot SR (2005) Who owns scientific data? The impact of intellectual property rights on the scientific publication chain. Learn Publ 18:91–94

Feng L et al (2000) Amino transferase activity and bioinformatic analysis of 1-aminocyclopropane-1-carboxylate synthase. Biochemistry 39(49):15242–15249

Gaudet P et al (2011) Towards BioDBcore: a community-defined information specification for biological databases. Nucleic Acid Res 39:D7–D10

Hultquist SJ, Harrison R, Yang Y (2002) Patenting bioinformatic inventions: emerging trends in the United States. Nat Biotechnol 20:743–744

Integrating Intellectual Property Rights and Development Policy (2002) Report of the Commission on Intellectual Property Rights

Jagadish AT (2013) Law of intellectual property and bioinformatics. Int J Sci Res Publ 3(1)

Jayaraman KS (1998) India to challenge basmati rice ‘Invention’. Nature 391:728

Maojo V, Kulikowski CA (2003) Bioinformatics and medical informatics: collaborations on the road to genomic medicine? J Am Med Inform Assoc 10:515–522

Mukherjee U (2008) A study of the basmati case (India-US basmati rice dispute): the geographiscal indication perspective. Soc Sci Res Netw. https://doi.org/10.2139/ssrn.1143209

Rhee SY, Dickerson J, Xu D (2006) Bioinformatics and its applications in plant biology. Annu Rev Plant Biol 57:335–360

Rogers M (1998) The definition and measurement of innovation. Melbourne institute working paper, 10/98

Shanker A (2012) Genome research in the cloud. OMICS J Integr Biol 16:422–428

Shook D (2002) Celera: a biotech that needs a boost, Bus. WK. http://www.businessweek.com/bwdaily/dnflash/mar2002/nf2002031_8351.html

USPTO Examination guidelines for computer-related inventions, Effective Date: 29 Mar (1996)

Next-Generation Sequencing: Technology, Advancements, and Applications

2

Gourja Bansal, Kiran Narta, and Manoj Ramesh Teltumbade

2.1 Introduction

The most comprehensive way of obtaining information about the genome of any living organism is to determine the precise order of nucleotides, known as sequencing, in its complete DNA sequence. Earlier, traditional methods which include Sanger's chain termination method and Maxam-Gilbert's chemical degradation method have been used for DNA sequencing. However, it is quite expensive and time consuming to sequence whole genome of an organism using the traditional method. Demand for low-cost and highly efficient sequencing gave rise to massively parallel sequencing technology which is known as "Next-Generation Sequencing (NGS)" (Koboldt et al. 2013). NGS refers to the high-throughput sequencing technologies which can simultaneously sequence millions or billions of DNA molecules.

Completion of Human Genome Project (HGP) in 2003 has marked a history in the field of genetics. Later the significant advancements have been made in sequencing technologies which decreased the cost of sequencing per base and increased the number of bases sequenced effectively per unit time (Metzker 2010).

Using the traditional sequencing method, it took nearly 15 years to sequence J.C. Venter genome as part of Human Genome Project, whereas, with the advent of NGS techniques, same can be completed in a couple of hours (Venter et al. 2015). In the current era, sequencing technologies have advanced to such a level that one can study the genome at a cellular level too. Single-cell sequencing techniques are now available which enable researchers to study cells individually rather than relying on an average signal from aggregate of cells (Wang and Navin 2015). In today's time, NGS methods have been applied to a variety of genomes ranging from singular to multicellular organisms. With the advent of NGS techniques, a number of

G. Bansal (✉) · K. Narta · M. R. Teltumbade
Institute of Integrative and Genomic Biology, New Delhi, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_2

applications and methods that leverage the power of genome-wide sequencing have increased at an exponential pace.

Rapid progress in sequencing techniques, as well as synchronized development in bioinformatics tools, has provided the solution to many different problems in the field of genetics and biology (Lelieveld et al. 2016). It allowed and helped researchers from diverse groups across the globe to generate draft genome sequence of any organism of their interest (Grada and Weinbrecht 2013).

Whole genome sequence of any organism cannot be read by traditional sequencers in a single run. Instead in a typical NGS run, thousands or millions of short overlapping sequences are produced concurrently in a single run. Each of the short sequence is called as a read. Reads are usually short (<= 250 bp) and can contain sequencing errors. To rule out sequencing errors, generally a genomic region is sequenced more than once. The number of reads spanning a genomic region determines the depth at which a genome is sequenced.

Along with whole genome sequencing, NGS techniques can also be applied to transcriptome sequencing (RNA-Seq), whole exome sequencing (WES), candidate gene sequencing (CGS), genotyping by sequencing (GBS), and chromatin immunoprecipitation sequencing also called as Chip-Seq (Furey 2012). Whole exome sequencing captures variations in all coding regions, or exons, of known genes and covers more than 95% of the total exons. As exome represents around 2% of the genome, it is a cost-effective alternative to whole genome sequencing if one is interested in finding variations only in coding part of the genome (Rabbani et al. 2014). RNA sequencing provides transcriptional activity of all coding as well as noncoding segments of the genome. As it can quantify alternative splice isoforms, RNA-Seq provides more accurate and precise measurements of gene expression levels than microarray platforms (Van Verk et al. 2013). Methylome sequencing complements genome sequencing as it determines the active methylation sites and provides a list of epigenetic markers that regulate gene expression, differentiation, and disease state (Schubeler 2015). Along with increasing our understanding toward genome sequence, sequencing methods also provide information about genetic variations, differential gene expression, and different aspects of transcriptional regulation.

There are several companies which make machines on which NGS can be done, such as Illumina (http://www.illumina.com), Roche (http://www.454.com), ABI/Life Technologies (http://www.lifetechnologies.com), Helicos BioSciences (http://www.helicosbio.com), Pacific Biosciences (www.pacificbiosciences.com), and Oxford Nanopore Technologies (http://www.nanoporetech.com). The different platforms vary in their sequencing technologies in terms of the sequencing chemistry, read length, number of reads per run, speed of sequencing, and cost per base pair sequenced (Goodwin et al. 2016). Table 2.1 provides a list of various NGS platforms along with their features.

In this chapter, we will be mainly focusing on the technology, advancements, and the applications of next-generation sequencing in the field of "-omics" development.

Table 2.1 Various NGS platforms and their features

Company	Sequencing principle	Detection	System platform	Read Length	No. of reads	Time/run	Throughput/ run	Accuracy (%)
Illumina	Reversible Terminator sequencing by synthesis	Fluorescence/Optical	HiSeq 2500/ 1500	36/50/ 100	3 billion (SE)	2~11 days	600 GB	>99
			Genome					
			Analyzer llx	35/50/75/ 100	320 billion (SE)	2~14 days	95 GB	>99
				25/36/ 100/25				
			MiSeq	0	17 million (SE)	4~27 h	8.5 GB	>99
Roche	Pyrosequencing	Optical	454 GS FLX+	700	1 million	23 h	0.7 GB	99.99
			455 GS Junior	400	1 million	10 h	0.035 GB	>99
Helicos Biosciences	Single molecule sequencing	Fluorescence/Optical	Heliscope	25~55	600~800 million	8 days	37 GB	99.99
ABI Life Technologies	Ligation	Fluorescence/Optical	5500 SOLiD	75 + 35	1.4 billion	7 days	90 GB	99.99
			5500×1 SOLiD	75 + 35	2.8 billion	7 days	180 GB	99.99
		Change in pH detected by Ion Sensitive Field effect	Ion Personal					
	Proton detection	Transistors	Genome Machine	35/200/ 400	12 million	2 h	2 GB	>99

(continued)

Table 2.1 (continued)

Company	Sequencing principle	Detection	System platform	Read Length	No. of reads	Time/run	Throughput/ run	Accuracy (%)
Pacific Bioscience	Real Time single molecule DNA sequencing	Fluorescence/Optical	PacBio RS	Average: 3000	~50 k	2 h	13 GB	84~85
Oxford Nanopore	Nanopore Exonuclease Sequencing	Electrical Conductivity	gridION	Tens of Kb	4~10 million	According to experiment	Tens of GB	96

2.2 History of NGS

The seeds of the genomics era were sown with the identification of DNA as the genetic material in 1952 by Alfred Hershey and Martha Chase (Hershey and Chase 1952). One year later, another breakthrough came with the discovery of the DNA double-helical structure by James Watson, Francis Crick, and Rosalind Franklin (Watson and Crick 1953). The basic knowledge of the genetic material led to the curiosity to know how a four-letter code (nucleotides) could govern all biological processes. The order of occurrence of the nucleotides (sequence) in the DNA seemed to play a major role in encoding the information in living organisms. This sequence in the DNA is conserved across generations of a species and sometimes even across different species. To unravel this information, many scientists focused on developing methods to decode the sequence of the genes and genomes that constitute organisms.

In 1977 Frederick Sanger and Alan Coulson sequenced the first genome, Phage Phi X-174 (PhiX), using a "plus and minus" method of sequencing (3). This technique used polyacrylamide gel for identification of the varied lengths of the amplified products. Since then the sequencing technologies have come a long way and have revolutionized the field of genomics. Aggressive research in the area of NGS has led to the development of novel chemistries and technologies, thereby increasing the speed and reducing the cost and time of sequencing (Grada and Weinbrecht 2013). This has led to an affordable sequencing of the human genome. Veritas Genetics has sequenced the human genome at a price of 1000 USD. This is a huge step in predictive and personalized medicine at an affordable price (Mardis 2006; Service 2006).

The efficiency of a sequencing technique is measured by its accuracy, speed, cost, and automation. Although there is no official classification, depending on the above parameters, the sequencing techniques can be broadly classified into three generations. The first generation includes the earliest sequencers (developed by Maxam-Gilbert and Sanger) where only small stretches of amplified DNA regions were sequenced. After that came the high-throughput sequencing technologies, which could sequence multiple DNA regions from multiple samples in one go and generate huge data output. These include the second-generation and the third-generation sequencers. Examples of second-generation technologies include Roche's 454, Illumina's Hiseq, and Life Technologies' SOLiD. The first- and second-generation sequencing techniques are dependent on amplification steps. This is mainly to ensure sufficient signal detection by the sequencer. The amplification comes with inherent biases and errors, which get incorporated into the resulting sequence. The third-generation sequencers do not require the amplification process. These sequencers can detect signals generated from a single molecule of DNA. They have longer reads, faster turnaround time, and higher output. The examples of third-generation sequencers include Helicos BioSciences' tSMS, Pacific Biosciences' SMRT sequencing, Illumina's Tru-seq Synthetic Long-Read technology, and Oxford Nanopore Technologies' sequencing platform. On the basis of advancements in sequencing technologies, sequencing era has been divided into multiple generations (Fig. 2.1).

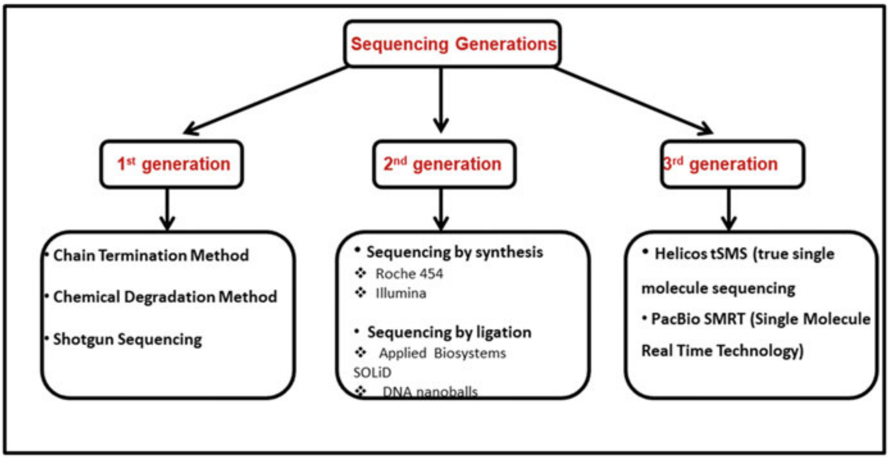

Sequencing Generations
1st generation
2nd generation
3rd generation
• Chain Termination Method
• Chemical Degradation Method
• Shotgun Sequencing
• Sequencing by synthesis
❖ Roche 454
❖ Illumina
• Sequencing by ligation
❖ Applied Biosystems SOLiD
❖ DNA nanoballs
• Helicos tSMS (true single molecule sequencing
• PacBio SMRT (Single Molecule Real Time Technology)

Fig. 2.1 Different sequencing generations

2.3 First-Generation Sequencers

First-generation sequencing techniques mainly used methods that could generate DNA fragments of different sizes. These methods used chemicals/enzymes to cleave DNA at specific sites or modified bases that could terminate the DNA amplification. The fragments hence generated were separated by electrophoresis using polyacrylamide (PAA) gel slabs. The two methods primarily used in the first-generation sequencers were chain termination method developed by Frederick Sanger and the chemical degradation method developed by Maxam-Gilbert (Morey et al. 2013).

2.3.1 Chemical Degradation Method

This approach was developed by Allan Maxam and Walter Gilbert (1977). In this method, each strand is chemically modified randomly, such that the backbone is exposed for degradation by alkali treatment at specific points. This process generates fragments of variable size. The fragmented DNA is terminally (both 3′ and 5′) radiolabeled with 32P and denatured. This is carried out as four reactions depending on the chemical treatment (e.g., G, A + G, C, C + T). The cleaved 32P-ssDNA is run on PAA gel. Then the autoradiograph is obtained from which the sequence of the fragment can be identified. This method involves the use of hazardous chemicals like dimethyl sulfate (for G and with acidic conditions release A), hydrazine, piperidine (for T and C), and NaCl (only C) (Maxam and Gilbert 1977). Also a huge amount of DNA is required. This technique is now obsolete.

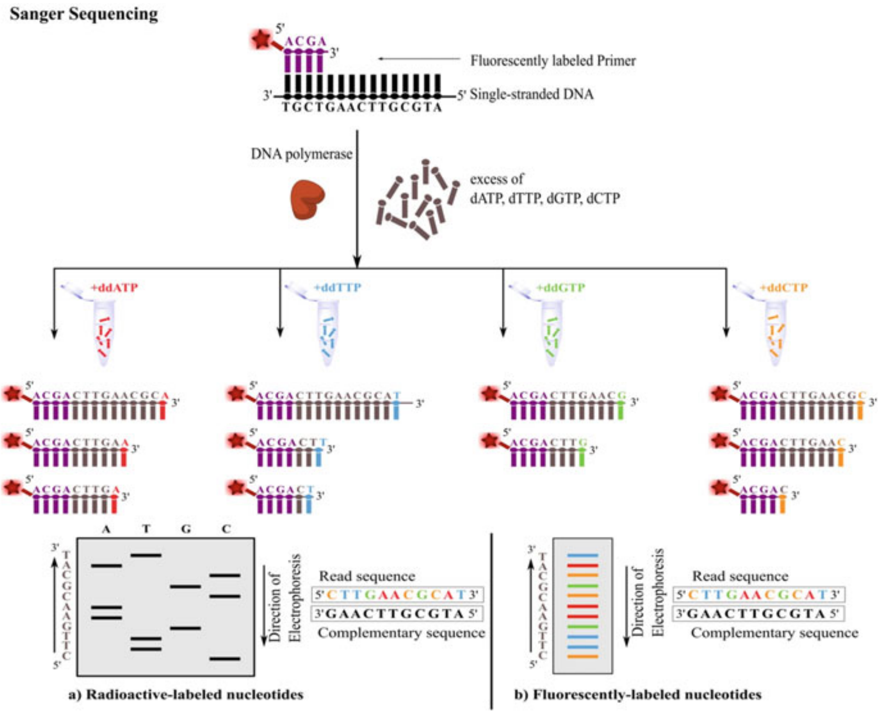


Fig. 2.2 Sanger di-deoxy chain termination method for sequencing: in Sanger sequencing, the template DNA is first primed with a fluorescently labeled (optional) primer. Then a sequencing reaction is carried out. The essential ingredients include a DNA template, primer, DNA polymerase, excess of dNTPs, and chain-terminating ddNTPs. The sequencing reaction can be performed either in four separate tubes, one for each ddNTP (**a**), or if fluorescently labeled ddNTPs are used, then all reactions can be performed in a single tube (**b**). After multiple rounds of template extension, the DNA fragments are denatured. The denatured fragments are run in gel slabs (now capillaries containing polymer) that separate the amplified products depending on their size. The sequence is then deciphered by the relative position of the bands from bottom to top

2.3.2 Chain Termination Method

In 1977, Frederick Sanger and group published the chain termination method of sequencing, also called as sequencing by chain termination (Sanger et al. 1977). This is now known as the Sanger method. Even today variations of this method are widely used by different sequencing techniques.

In the Sanger method (Fig. 2.2), the sample is first denatured by heat, and then four reactions are performed with ssDNA. Each tube contains a primer, DNA polymerase 1(Klenow enzyme), and all four dNTPs and one of four ddNTPs. The ddNTPs have hydrogen (instead of hydroxy-) group at the 3′ terminal. The amplification is carried out by extension of the primer, using single-stranded DNA (ssDNA) as a template. The presence of dNTPs and specific ddNTPs can randomly terminate the extending DNA chain. The DNA is then denatured and run on PAA gel. The bands hence obtained are combined to form a single sequence (Sanger et al. 1977). Initially, radioisotopes were used to label dNTPs (Fig. 2.2a); later it was completely replaced by fluorescent dyes (Fig. 2.2b) (Smith et al. 1986).

Sanger sequencer ABI 370 was the first automated sequencer and it was launched in 1986 by Applied Biosystems (now Life Technologies). This included some significant improvements in the method like running DNA in capillaries instead of gel slabs, introduction of dye-labeled ddNTPs making way for one tube reactions, multicapillary electrophoresis, and automatic loading of samples (Ansorge et al. 1986; 1987).

Sanger sequencing was widely used in the 1990s to sequence genes and genomes. Even today it plays a very important part in screening genes for disease mutations and validation of data from the next-generation sequencers. The average read length of sequence from Sanger data is still more than most of the second-generation sequencers (Treangen and Salzberg 2011). Sanger sequencing formed the basis of the first draft of the human genome. In 2001 two landmark papers were published, which reported the sequencing of the human genome (Lander et al. 2001; Venter et al. 2001). Celera Genomics used shotgun sequencing in which a large piece of DNA is fragmented mechanically. Each fragment is sequenced independently, using the Sanger method. The sequences obtained are then assembled using the overlapping regions to get a complete sequence (Anderson 1981). Shotgun sequencing can be considered as the bridge between the first-generation and the second-generation sequencers.

After the completion of the Human Genome Project, scientists everywhere realized the enormous potential in identifying the DNA/RNA sequence information of an organism. The primary limitation of the first-generation sequencers was their low output and inability to scale up. The cost per base sequenced is also very high as compared to the high-throughput methods (Mardis 2011). To overcome these issues, automated, faster, and cheaper sequencers were developed. They were primed to sequence longer and large number of DNA molecules parallelly.

2.4 Second-Generation Sequencers

The second-generation sequencers can generate a huge amount of data in one run at a much lower cost and higher speed as compared to the first-generation sequencers. These sequencers use amplified DNA fragments and the sequencing is performed in parallel for millions of DNA fragments which is why it is also called as the massively parallel sequencing. The second-generation sequencers include three major processes: library preparation, amplification, and imaging/sequencing (Mardis 2008). These steps may vary in different sequencers of this generation.

The basic steps in next-generation sequencing are represented in Fig. 2.3.

Library Preparation It primarily involves fragmentation of the DNA and adapter ligation.

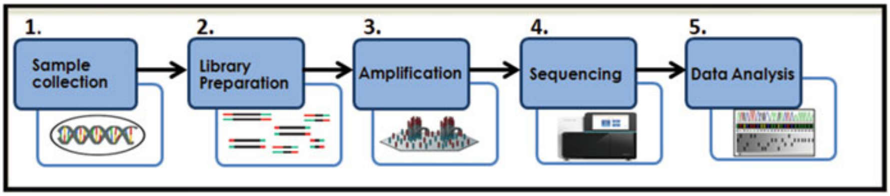


Fig. 2.3 Basic steps in next-generation sequencing

The second-generation sequencing techniques can only sequence small stretches of DNA molecules. Therefore, it is important to fragment the DNA molecules such that they can be sequenced and after that reassembled. Fragmentation can be either mechanical shearing or enzymatic cleavage (Morey et al. 2013).

The fragmented DNA is attached to universal adapters to facilitate amplification and attachment to a surface for sequencing. Adapters are double stranded and have sites for primer binding. They also have index sequences which are used to differentiate the reads coming from various samples, if multiple samples are pooled together during sequencing. Primer binding sites are used to prime the sequencing reaction (Morey et al. 2013). The steps involved in library preparation are shown in Fig. 2.4.

Amplification of Template Most of the second-generation sequencers are not able to detect fluorescence from a single DNA molecule. To overcome this, the DNA molecules are attached/immobilized on a surface, which are then amplified (Morey et al. 2013). This enables the sequencers to capture a clear signal while imaging. Two major amplification techniques are:

(i) Emulsion PCR: Used by 454 (Roche), SOLiD, and Ion Torrent (Thermo Fisher)
(ii) Solid-phase amplification: Illumina

This step produces clonal templates for sequencing (Goodwin et al. 2016).

Sequencing The amplified products from the previous step are sequenced in this step. A sequencing primer is added to the templates to start the addition of bases and their simultaneous imaging. These steps are carried out in a cyclic fashion. Sequencing can be on the basis of one of the following two principles (Goodwin et al. 2016):

(a) *Sequencing by synthesis*: This technique makes use of the DNA polymerase to add bases sequentially. The major platforms which use this approach are 454, Illumina, Qiagen, and Ion Torrent.
(b) *Sequencing by ligation*: In this technique, a fluorophore-bound probe is hybridized to the template and ligated to the former previous base. Once ligation is complete, the probes are imaged to identify the bases. SOLiD and Complete Genomics use this method of sequencing.

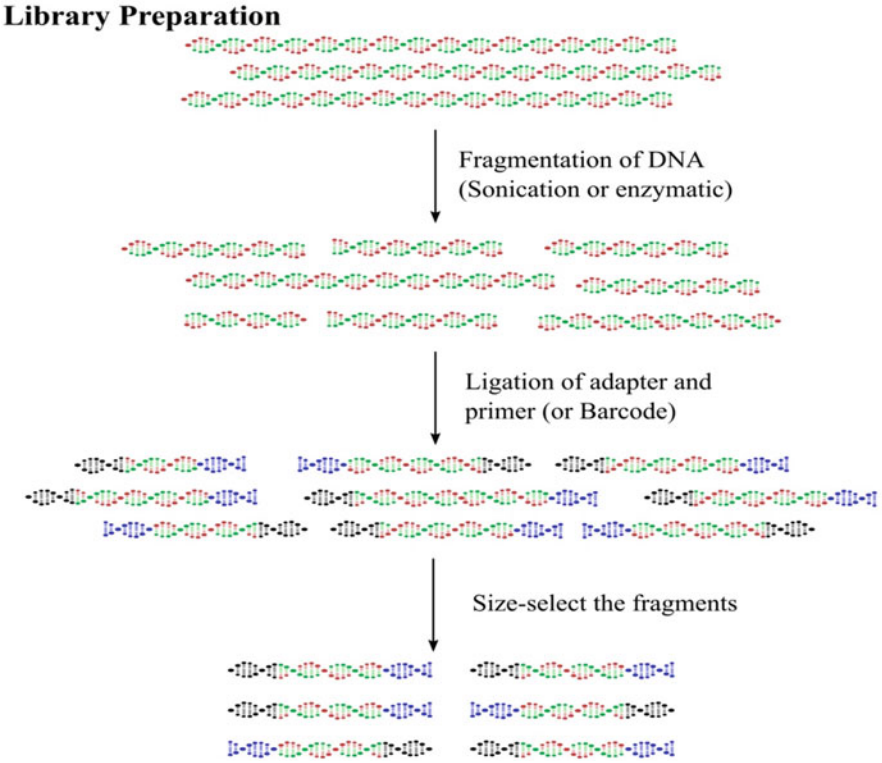

Library Preparation
Fragmentation of DNA (Sonication or enzymatic)
Ligation of adapter and primer (or Barcode)
Size-select the fragments

Fig. 2.4 Diagrammatic view of library preparation: The first step involves fragmentation of the DNA molecule. It can be carried out by (a) enzymatic methods by nonspecific endonuclease treatments or tagmentation using transposase and (b) physical methods using sonication or acoustic shearing using Covaris. In the case of RNA sequencing, the fragmentation is usually done by heating in the presence of divalent cations and then the fragmented RNAs are converted to cDNA After fragmentation the ends are repaired, i.e., the ends are blunted, the 5′ ends are phosphorylated, and the 3′ ends are adenylated. The adapters are ligated to the fragments (represented here by blue and black). The adapters are barcoded, using index sequences, so that multiple samples can be sequenced together. The part of the template between the adapters (they are of constant length) is called the insert, and it determines the library size. The insert size is in turn determined by the application and the technology used for sequencing
Size selection involves the steps to obtain the library size within a desired range and to remove adapter dimers. This is usually done with the help of magnetic beads (e.g., Agencourt AMPure XP beads) or the library is run on agarose gel. Once the library is prepared, its quality and quantity is checked before moving on to sequencing

The basic principle involves the cyclic process of addition of bases to the primer until the required number of bases is read (imaged) from the template. All the reactions are performed in parallel.

Imaging of Bases and Data Analysis After the sequencing is performed, the information from images is converted to bases. This is generally carried out using platform-specific software which generates raw files of sequence data for further processing and analysis.

This is the step from where data comes out of the experimental lab to a high configuration computer. Sequence analysis is done by a bioinformatician who analyzes the data and draws meaningful insights out of it. Different open-source algorithms are available for each step of analysis workflow.

Some of the widely used second-generation sequencers are discussed here:

2.4.1 Roche 454 Genome Sequencer

Genome Sequencer GS20 was the first commercialized second-generation sequencer launched in 2005 (www.454.com). It was developed in 1996 at the Stockholm Royal Institute of Technology *and* has been used to sequence Neanderthal, barley, and *Helicobacter pylori* genomes (Rothberg and Leamon 2008).

Library Preparation

Nebulizer randomly fragments the DNA. These fragments are then flanked by two types of adapters on different sides and denatured. One of the adapters contains the sequencing primer binding site and the other adapter has a biotin label. Only the fragments containing different adapters are selected and mixed with capture beads. The capture beads have probes complementary to adapter containing the biotin label so DNA fragments can bind to them. Excess of capture beads are added to bind only one DNA molecule to each bead (Fig. 2.5b).

Amplification

This technique uses emulsion PCR to amplify the DNA fragments bound to the beads clonally. By the end of this process, there are millions of clonal molecules on the beads, which are denatured to obtain only ssDNA molecule. Sequencing primer is attached to the ssDNA adapters.

Sequencing

The beads are then loaded on a picotiter plate containing wells. The dimensions of wells are such that only one bead enters the individual well. Later smaller packing beads containing immobilized sulfurylase and luciferase are added to the wells. 454 sequencing is based on pyrosequencing (Fig. 2.5b). In this technique, as the nucleotides are incorporated, pyrophosphate (PPi) is released. PPi is then converted to ATP using ATP sulfurylase and adenosine phosphosulfate. This ATP combines with luciferase to convert luciferin to oxyluciferin. This generates a light signal, which is detected by a charge-coupled device (CCD) at the bottom of the plate. Each nucleotide is given one at a time in a predesigned order. Incorporation of more than one nucleotide on the same molecule is read as stronger intensity. Therefore, the

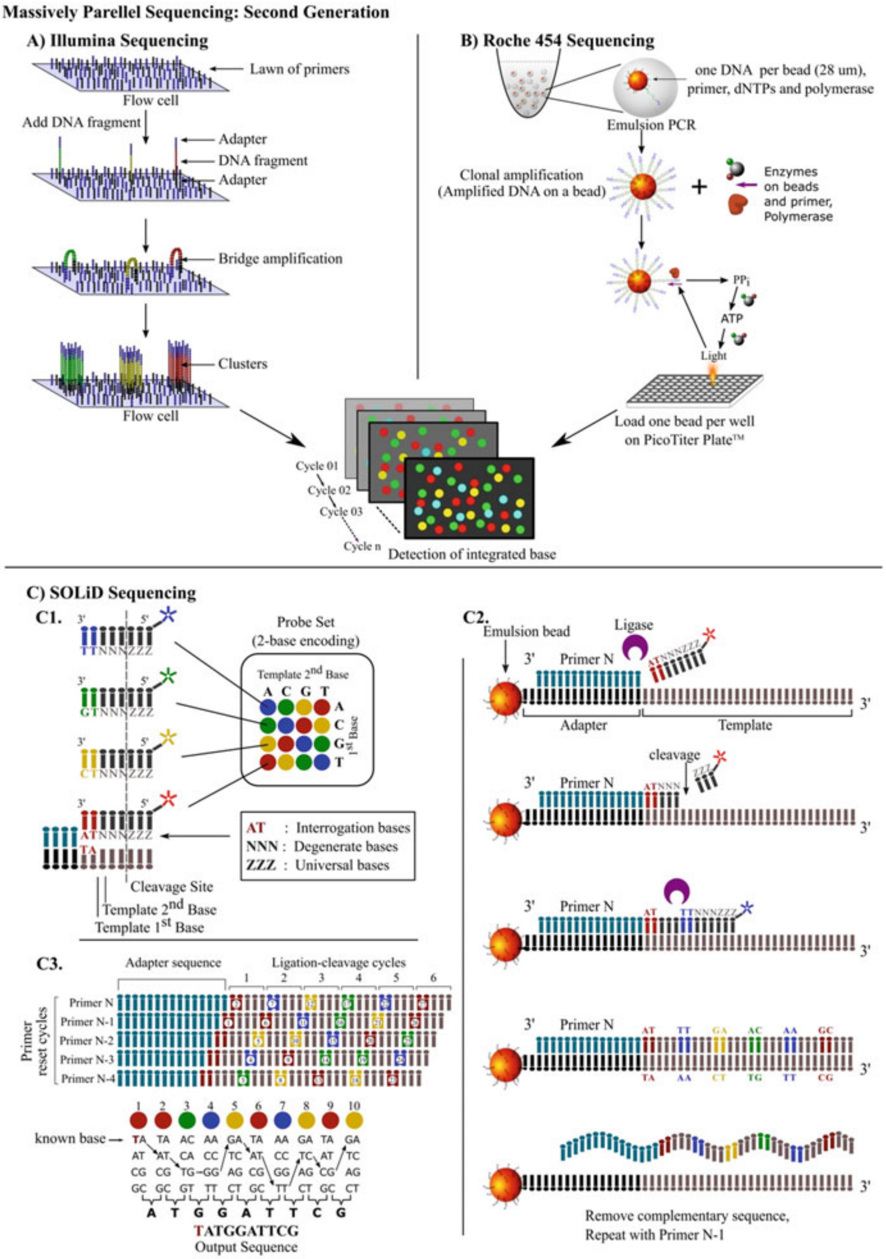


Fig. 2.5 Different platforms for second-generation sequencing

(**a**) Illumina sequencing: The prepared library is distributed over a flow cell. The flow cell contains a lawn of primers that are complementary to the ends of the adapters, as a result of which the DNA fragments bind to them. Solid-phase bridge amplification involves amplification of the attached templates in the presence of unlabeled nucleotides, polymerase, and buffer. The dsDNA is then amount of light can be regressed to calculate the number of similar nucleotides incorporated. After the completion of required number of cycles, the sequence is acquired by combining the information generated (Rothberg and Leamon 2008).

This technique can generate reads up to 1 Kb in length and faces major problems in identifying homopolymer stretches in the genome. Moreover, it has lower output and sequence per base cost is higher than its competitors (Mardis 2008; Margulies et al. 2005).

2.4.2 Illumina

It is one of the most successful sequencing platforms. The technology, developed by Solexa, was first commercialized as Genome Analyzer (GA) in 2006. One year later,

Fig. 2.5 (continued) denatured and the original template is washed away, leaving behind the primer and the elongated strand. This process is repeated for a number of cycles to generate clusters from each attached template. After cluster generation, the strands are sequenced. The sequencing is initiated once the sequencing primers are attached and the reagents containing labeled nucleotides and polymerase are added. As each base is added, it is imaged after removing the unincorporated bases. The fluorescent signal is then removed from the incorporated nucleotides. Multiple such cycles are carried out to obtain sequential images of bases of each cluster
(**b**) Roche 454 sequencing: In this, the adapter-ligated fragments are attached to the beads. Then amplification process takes place by emulsion PCR. One bead per well is distributed onto a 454 picotiter plate. The amplification is then carried out by pyrosequencing in which, unlike Illumina, only one type of deoxynucleotide triphosphate base is provided to be incorporated by the polymerase into a cycle. This addition is accompanied by the release of pyrophosphate (PPi). With the help of enzymes (ATP sulfurylase, luciferase, and luciferin attached to the bead), the PPi is converted to light, which is detected. The base addition in each well is recorded for the desired number of cycles, and simultaneously the sequences on the template are deciphered
(**c**) SOLiD sequencing: The emulsion PCR (emPCR) amplification process is similar to Roche. Then the beads with amplified template are deposited on a slide. SOLiD uses the sequencing by ligation method. (c1) SOLiD uses two-base encoding and uses four different colored fluorescence probes. In the probe color matrix, each color represents 4 out of 16 possible combinations. The probes are made up of eight bases. The first two are the template bases, which match read positions on the sequence to be read. The next three are degenerate bases that match the three unread bases upstream to the template bases. The identity of these three bases is not needed for sequencing. Finally, the three universal bases can bind to any of the nucleotides. They have a fluorescent dye attached at 5′ end and has cleavage site at the 3′ end. (c2) The basis of SOLiD is that the labeled probes get ligated to the primer only if they are perfectly matched. Once ligated, the three universal bases at the 5′ end are removed. The remaining ligated probe acts as a primer for the next probe. This process is carried out for the desired number of ligation cleave cycles. Then the extension product is removed and the template is reset with shorter $n-1$ primer revealing a thymidine (T) at the adapter for next round. (c3) The numbers enclosed in white circles represent the sequence of base position in the template. The cycle numbers are denoted on the top. It requires five iterations (n, $n-1$, $n-2$, n-3, $n-4$) to decipher the complete sequence and to fill the gaps (of the two template bases and three degenerate bases) formed during the first iteration. Also represented here is the walk-through to obtain sequence information from multiple iterations

Solexa was purchased by Illumina (Treangen and Salzberg 2011). It uses reversible terminator reactions to carry out sequencing by synthesis.

Library Preparation DNA is first randomly broken into fragments, by mechanical shearing (using Covaris) or by enzymes/transposomes. Generally 200–300 bp fragments are selected using gel or SPRI/Ampure beads and adapters are ligated to them. These are then denatured and injected onto a solid surface called flow cell. The adapters are ligated to the end of the fragments such that one end serves to attach to the flow cell and other is used for sequencing (Fig. 2.5a).

Amplification The surface of a flow cell has a lawn of probes, which has sequences complementary to one end of the adapters bound to the DNA fragments. When the adapter-ligated fragments are distributed over a flow cell, they bind to complementary probes. To obtain clonal copies of the DNA, a process called bridge amplification is used. By the end of this process, thousands of copies are generated for each attached fragment, which corresponds to a cluster. Reverse strands are cleaved and the single strands are primed for sequencing (Adessi et al. 2000).

Sequencing by Synthesis The amplified products are sequenced by a process similar to the Sanger's chain termination reaction. Here the nucleotides are labeled with four different fluorescent dyes and are capable of reversible termination. During sequencing, the flow cell is flushed with all four nucleotides. A complementary fluorescently labeled base is added to the primed template by the polymerase bound to the template. After addition of a single base, the reaction is terminated. To obtain signal from the bound nucleotide, the remaining unbound nucleotides are washed away. The bound-labeled nucleotides are then illuminated with the help of lasers. Imaging is performed to identify location and type of base incorporated in each cluster. Then the fluorescent label is cleaved thereby exposing the 3′-OH group, to which a base can be integrated. This process is carried out with the help of Tris (2-carboxyethyl) phosphine (TCEP, reducing agent).

This cycle of addition of dNTPs, imaging, and cleavage is carried out until all the bases are read. Then all the images can be simultaneously combined to create the sequence for each cluster.

Illumina provides with the ability to sequence DNA from one end (single-end sequencing) or both ends of the fragment (paired-end sequencing) and mate pair sequencing (used to sequence the ends of long fragments, ignoring the bases in between) (Fuller et al. 2009; Pettersson et al. 2009; Rothberg and Leamon 2008).

Over the past decade, this technology is continuously improving its sequencers in terms of efficiency and accuracy. It generates one of the highest outputs with lowest reagent cost among all the sequencers present to date (Liu et al. 2012).

Applied Biosystems' SOLiD

George Church in 2005 developed small oligonucleotide ligation and detection (SOLiD) system for high-throughput DNA sequencing (Shendure et al. 2005). It was commercialized by Applied Biosystems (now Life Technologies) in 2007. Its principle involves sequencing by ligation (SBL).

Sample Preparation and Amplification

SOLiD uses emulsion PCR, similar to Roche's 454. Fragmentation is achieved by nebulization/sonication or digestion. Universal adapters are attached to the ends of the fragmented template which are then deposited onto microbeads. The templates undergo clonal amplification reaction in water/oil emulsion microdroplets. The microbeads are then distributed on a glass slide to which they bind covalently. Based on application, slides may have one, four, or eight compartments**.**

Sequencing and Imaging:

The sequencing reaction starts by annealing of primer to the amplified template. In SOLiD sequencing, each cycle constitutes the following four steps:

1. The chemical reaction involves the binding of eight-nucleotide-long probes. Only the first two bases (di-base) at 3′end have a known sequence. The rest of the probe is degenerate. The probe is fluorescently labeled at the last base on 5′ end such that it corresponds to a specific di-base. The probe binds to complementary sequence next to the primer. Due to restriction in available fluorescent dyes, the complementary and reverse di-bases are encoded by the same color (FAM for AA, CC, GG, TT; Cy3 for AC, CA, TG, GT; TXR for AG, GA, TC, CT; Cy5 for AT, TA, CG, GC) (Fig. 2.5c1). After the primer is ligated to the adapter, octamer probes with same fluorescent label are added (Fig. 2.5c2).
2. When a complementary probe binds to the template, DNA ligase hybridizes the probe to the primer. During this process, a fluorescent signal is emitted, which is captured by the detector.
3. After ligation, the three bases (including the dye) at the 5′end of the probe are cleaved.
4. These steps are repeated with the three remaining fluorescent dye pool of probes. After each successful annealing, each probe is ligated to the previous probe in the second step. So at the end of one cycle, two bases are read per three skipped bases per probe. This process can be carried out for the desired number of times (Fig. 2.5c2).

After this, the primer along with all the probes is removed. Then a new primer is added such that it anneals to the penultimate base from the adapter-template junction ($n-1$). The abovementioned steps are repeated. This cycle is carried out four times (for $n-2$, $n-3$, and $n-4$ also) and every time primer shifted one base toward 5′ end (Fig. 2.5c3) (Mardis 2008; Valouev et al. 2008).

Data Analysis: Exact Call Chemistry It uses eight-base interrogation system, with four different colored primers to map possible combinations in sequences.

SOLiD sequencers are known to have problems while handling palindromic sequences. However, they are less error prone, as each base is read twice as compared to other second-generation sequencers. They are flexible, allowing for sequencing in different applications. It detects single nucleotide variants (SNVs) and insertion/deletion (indels) with ease.

Despite a high output and multi-sample processing by the second-generation sequencers, there is still scope for improvement. Second-generation sequencers face issues related to errors due to amplification and need for repeated "wash and scan" cycles (Metzker 2010) which are not only time consuming but also lead to asynchronous (dephasing) sequencing (Whiteford et al. 2009). These issues result in erroneous base calls and also limited read lengths (Metzker 2010).

2.5 Third-Generation Sequencing

Given a very rapid evolution of the successive sequencing technologies over the past few years, and therefore small time lapse, there has been a continuum of improvements among the successive next-generation sequencers. The next-generation sequencers are considered as the third generation primarily on the basis of the following features: First, they do not require amplification of template DNA. Second, the sequencing is performed in real time. They also do not require repeated "wash and scan" cycles. These sequencers generate longer reads and have higher speed and accuracy with lower cost and effort (Gut 2013; Heather and Chain 2016; Morey et al. 2013; Niedringhaus et al. 2011; Pareek et al. 2011; Schadt et al. 2010).

Three of the third-generation techniques are discussed below.

2.5.1 Helicos: tSMS (True Single-Molecule Sequencing)

Helicos BioSciences' tSMS was the first commercially available third-generation single-molecule sequencer (Heather and Chain 2016).

Library Preparation The DNA is broken down into 100–200 bp fragments, and a poly A sequence of approximately 50 bp is attached to 3′ end of each fragment. The fragments are labeled with fluorescent adenosine. These labeled fragments serve as templates for sequencing and are hybridized on the surface of a poly-T-containing flow cell. The flow cell containing 25 channels has oligo-dT (50 bases) primer attached to the surface. The 3′OH of the tailed molecules are blocked by terminal transferase and dideoxynucleotides to prevent extension.

Sequencing Before sequencing begins, the location of each fluorescently labeled template is captured, by illuminating with a laser. After imaging the fluorescent label is washed away. To start sequencing reaction reversible terminator fluorescently labeled nucleotides (Bowers et al. 2009) and DNA polymerase is added to the flow cell. The sequencing chemistry is similar to Illumina, where signals are captured after a laser illuminates the flow cell. In the case of Helicos, a single type of nucleotide is added at a time (e.g., A). The camera records the addition of each nucleotide on a single DNA fragment. After imaging, the labels are cleaved and washed. This process takes place for the remaining three bases also (C, G, T). This cycle of sequencing is repeated until the required read length is achieved (Fig. 2.6a).

This method requires shorter sample preparation time and can be used to sequence degraded molecules also. A higher accuracy is achieved, as there is no PCR amplification step, but the sequencing time is long due to repeated cleaving and washing steps and also per base cost is high.

2.5.2 Single-Molecule Real-Time Technology (SMRT)

Single-molecule real-time technology was developed by Pacific Biosciences**.**

Library Preparation The DNA fragmentation is performed depending on required insert size, with a range from 500 bp to 10 kb. End repair is carried out to create blunt ends and addition of dA tail. Then SMRTbell hairpin loop adapters are ligated to both ends of the double-stranded fragments. A SMRT library is prepared after purification steps which ensure that only the fragments having adapter ligated to both ends are selected. A Φ29 DNA polymerase is attached to the DNA molecules of the library. This enzyme also has a strand displacement property, so the double-stranded DNA can be opened up into circular template (Eid et al. 2009).

Sequencing The sequencing reactions are carried out in a chip containing small wells (10^–21 L). Each of these reaction cells, also called zero mode waveguide (ZMW), has a molecule of Φ29 DNA polymerase attached to the bottom. ZMW are small pores surrounded by metal film and silicon dioxide (Foquet et al. 2008)**.** Once the template is added, it binds to the DNA polymerase. Then the fluorescently labeled dNTPs are added to the wells. All four nucleotides are phospho-linked and have different colored fluorophores (Korlach et al. 2008a). In this method, the fluorescence is attached to the terminal phosphate of the nucleotide (instead of the base as in previous cases).

During sequencing, the complementary dNTP enters the polymerase and emits a fluorescence signal in the ZMW. This signal is detected as a light pulse in the detection volume of 20 zeptoliters (Korlach et al. 2008b). The fluorescence label is released after cleaving the phosphate chain. Then a new base is incorporated (Fig. 2.6b).

This is a high-speed process as ~10 nucleotides can be added in a second.

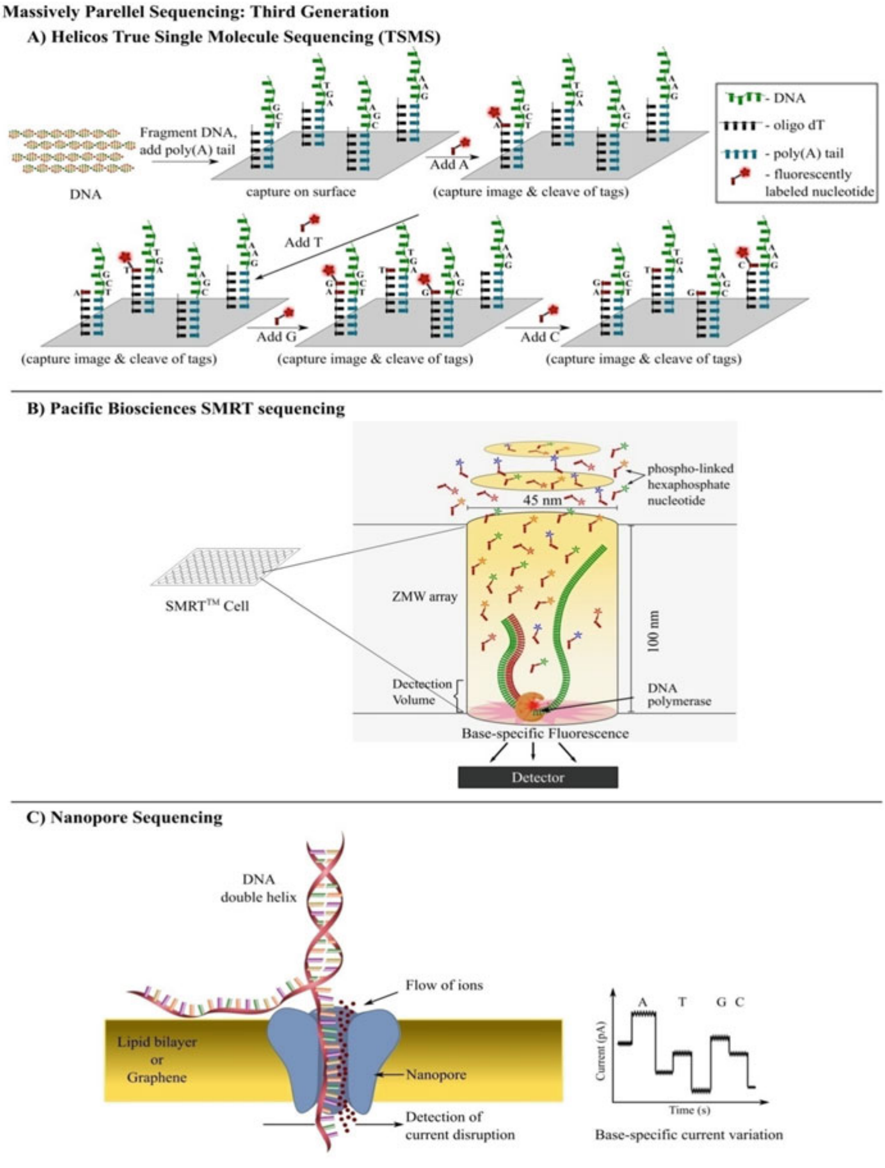

Massively Parellel Sequencing: Third Generation
A) Helicos True Single Molecule Sequencing (TSMS)
Fragment DNA, add poly(A) tail
DNA
capture on surface
Add A
(capture image & cleave of tags)
- DNA
- oligo dT
- poly(A) tail
- fluorescently labeled nucleotide
Add T
Add G
Add C
(capture image & cleave of tags)
(capture image & cleave of tags)
(capture image & cleave of tags)
B) Pacific Biosciences SMRT sequencing
phospho-linked hexaphosphate nucleotide
45 nm
SMRT™ Cell
ZMW array
100 nm
Dectection Volume
DNA polymerase
Base-specific Fluorescence
Detector
C) Nanopore Sequencing
DNA double helix
Flow of ions
Lipid bilayer or Graphene
Nanopore
Detection of current disruption
A T G C
Current (pA)
Time (s)
Base-specific current variation

Fig. 2.6 Detailed view of third-generation sequencing platforms

(**a**) TSMS: The poly A tailed fragmented strands are hybridized to the poly-T-bound Helicos flow cell plate. Then the fluorescent labeled nucleotides are added one at a time. The addition is done in a cycle of "quads," where a quad consists of adding each base (A, T, G, and C) once. The labeled bases are then illuminated by a laser, and the images are taken, which helps in detecting the strands that have bound nucleotides. Before adding the new labeled bases, the labels from the hybridized *bases are cleaved*

(**b**) SMRT: A SMRT cell with ZMW nanostructures has a DNA polymerase immobilized to the bottom of the well. The fluorescently labeled phospho-linked nucleotides are added to the primed

2.5.3 Nanopore Sequencing

Nanopore-based chemical and biological molecule detection is among the most advanced sequencing technologies available today (Morey et al. 2013). The nanopores can be synthetic solid-state or biological in nature. The biological nanopores are modeled on the transporters and ion channels present inside the living cell (Haque et al. 2013). The nanopores decode a sequence, as the string of DNA is transported through it. They capture the modulation in the ion flow through these channels or optical signals in real time (Astier et al. 2006).

The α-hemolysin ion channels were the first ones to be used for this purpose (Kasianowicz et al. 1996) and were commercialized by Oxford Nanopore Technologies (Schadt et al. 2010). The modified α-hemolysin protein is embedded in a lipid bilayer and has an exonuclease on the outer surface, and a cyclodextrin sensor is attached on the inside. The exonuclease cleaves each base as it enters the pore. As the base crosses the channel, the variation in current is detected, which correlates with the specific parameters of the nucleotide (Fig. 2.6c).

Improvements in this technique and various other approaches have led to more accurate and a variety of nanopore sequencing platforms. These include:

(a) Using *Mycobacterium smegmatis* porin A (MspA) protein to sequence an intact ssDNA (Derrington et al. 2010).
(b) Optical detection in nanopore sequencing via multi-colored readouts using synthetic DNA (McNally et al. 2010).
(c) Synthetic material has also been incorporated for improvements in this technology; among them solid-state graphene nanopores and carbon nanotubes are of particular interest (Bayley 2010; Liu et al. 2010; Schneider et al. 2010; Zhao et al. 2012).

The nanopore sequencing is inexpensive as there is no addition of modified/fluorescent bases. Nanopore sequencing is marketed by Oxford Nanopore Technologies through their sequencing platform GridION along with a portable device MinION and the scalable PromethION (https://nanoporetech.com).

Fig. 2.6 (continued) DNA template (green). A signal is recorded when a base is bound to the template in the active site. The fluorophores are activated by the lasers only when they are in the detection volume. As the detection volume is minimal, bottom 20–30 nm, therefore only the correctly bound nucleotide is detected. After the phosphodiester-bond formation, the template is translocated so that next base can be attached. The location of the detector for the optical image is under the nanostructure. (**c**) A voltage-biased membrane (lipid bilayer/graphene) separates two aqueous electrolytes containing chambers. A flow of ionic current occurs through pores (blue) present across the membrane. The passage of the DNA is controlled by the enzymes present in the nanopore, as a result of which there is a disruption in the passage of the ions, which is measured by the very sensitive ammeter. A record showing the measurement of the passage of ions corresponding to the type of nucleotide crossing the pore is also represented alongside

The third-generation sequencing techniques promise to deliver longer reads than any of the previous technology. These long reads (5–15 kb) are proving to be important in many areas (Lee et al. 2016)**.** They are instrumental to fill the gaps in the human genome (Chaisson et al. 2015; Pendleton et al. 2015) *and also* used to get highly accurate reassembly and reconstructs of many bacterial, plant, and animal genomes (Berlin et al. 2015; Chen et al. 2014; Gordon et al. 2016; Koren et al. 2013; Loman et al. 2015). The third-generation sequencers are of particular importance in deciphering the diversity of the metagenome and identifying novel transcript isoforms and gene fusion events (Oulas et al. 2015; Sharon et al. 2013).

2.6 Bioinformatics Analysis Pipeline

Through next-generation sequencing technologies, it is now possible to describe methylated regions in the genome sequence, sequencing whole genomes, transcriptome, catalog noncoding RNA, and protein-DNA interaction sites. Each of these applications generates gigabases of sequence information which imposed an increasing demand on statistical methods and bioinformatics tools for analysis and management of enormous data produced by different sequencing platforms (Grada and Weinbrecht 2013).

The first step in sequence data analysis is to produce short nucleotide sequence also referred as reads and their associated quality scores from raw light intensity signals. This is called as base calling and the related software are usually provided by the manufacturer of the sequencing platforms (McGinn and Gut 2013). For example, CASAVA is a base calling software provided by Illumina for converting intensity files to human readable file format, e.g., FASTQ. Short reads generated are stored in the short read archive (SRA) in FASTQ format.

SRA compact design allows storage and retrieval of sequence data including metadata from experiments and reads with associated quality scores in a very effective manner. We can convert SRA to FASTQ file using SRA Toolkit.

FASTQ is a text-based format for storing biological sequences. It is basically a FASTA file associated with the quality score for each base (Mills 2014). A FASTQ file normally uses four lines per sequence.

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

Line 2 is the raw sequence letters.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the sequence in Line 2 and must contain the same number of symbols as letters in the sequence.

A FASTQ file containing a single sequence might look like this:

```
@SEQ_ID
TTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACT
   CACAGTTT
+
*((((***+))%%%++)(%%%%).1***-+*”))**55CCF>>>>>>CCCCCCC65
```

Phred Score:

A Phred score is a measure of quality of nucleotide bases identified using DNA sequencing (Table 2.2). It is commonly defined as

$$Q = -10\log10P$$

where P is the probability that base is incorrectly called.

Quality Filtering

Before doing any downstream analyses, it is always advisable to check for read quality. Sequencing artifacts like base calling errors, poor quality reads, adapter contamination, PCR duplication, and GC biasedness are common factors that need to be checked. Filtering is the most crucial step to remove low-quality reads as during further analysis one cannot have control on quality of reads (Watson 2014). Alignment or mapping is the next step in NGS analysis. Two different ways are possible for mapping millions of reads. One is a comparative mapping of reads with the reference genome (DNA sequence of species under consideration), and another is de novo assembly. Reference genome is the DNA sequence database of an organism representing species set of genes. Mapping of reads onto reference genome provides a tentative map indicating regions from where the reads belong. A reference genome for individual organisms can be accessed from various web resources like NCBI, Ensembl, or UCSC genome browser. In the absence of reference genome, de novo genome assembly is done with the help of overlapping reads to stitch consecutive regions in the genome (Fonseca et al. 2012). A variety of tools are available for both comparative mapping and for de novo assembly. Alignment results are stored in BAM (Binary Alignment Map)/SAM (Sequence Alignment Map) file (Li et al. 2009). Best mapping hits can be filtered out using multiple parameters like mapping quality score. For every study, filtering and alignment are common steps in sequence

Table 2.2 Phred score and corresponding incorrect base call probability

Phred Quality Score	Probability of incorrect base call	Base call accuracy (in %)
10	1 in 10	90
20	1 in 100	99
30	1 in 1000	99.90
40	1 in 10,000	99.99
50	1 in 100,000	100.00

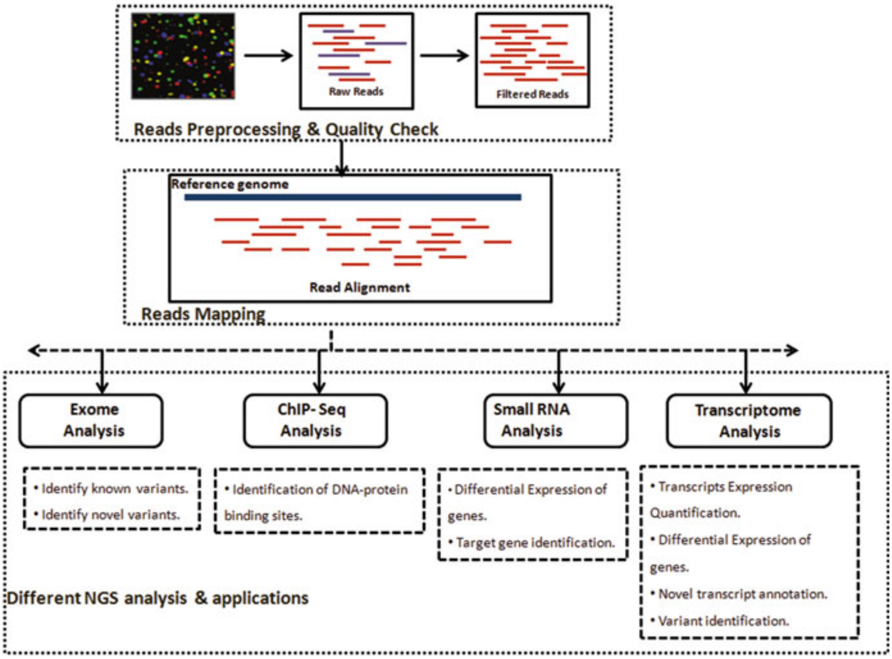

Raw Reads
Filtered Reads
Reads Preprocessing & Quality Check
Reference genome
Read Alignment
Reads Mapping
Exome Analysis
ChIP- Seq Analysis
Small RNA Analysis
Transcriptome Analysis
• Identify known variants.
• Identify novel variants.
• Identification of DNA-protein binding sites.
• Differential Expression of genes.
• Target gene identification.
• Transcripts Expression Quantification.
• Differential Expression of genes.
• Novel transcript annotation.
• Variant identification.
Different NGS analysis & applications

Fig. 2.7 Basic workflow in NGS data analysis and their applications

data analysis. Real fate of sequence analysis is decided only after alignment. Using different kinds of high-throughput data, various analyses can be done (Fig. 2.7).

Exome or whole genome sequencing can be used to detect structural variations that can give rise to different phenotypes in different individuals. Somatic as well as germline mutations can be identified with high precision using exome sequencing. Genes that are expressed differentially in various conditions/groups can be identified through transcriptome analysis. Different software are available for each step of data analysis. Table 2.3 describes freely available commonly used software for various purposes.

For data visualization, a graphical interface usually called as genome browser is required to display analysis results. Comparative analysis with other genomic resources (dbSNP, 1000 genome), expression changes, and peak folding is also possible on these browsers. The real strength of analysis is reflected from the power to display the results in an easy-to-interpret manner. Common IGV, UCSC, Tablet, and MapView are examples of genomic browsers.

Table 2.3 List of software commonly used in NGS data analysis

Commonly used softwares in next generation data analysis	
Reads quality control softwares	*Web URL*
fastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
filteR	http://scbb.ihbt.res.in/SCBB_dept/filter.php
Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic
FastX toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
DNA/RNA alignment	
BWA	http://bio-bwa.sourceforge.net/
Bowtie	http://bowtie-bio.sourceforge.net/index.shtml
Stampy	http://www.well.ox.ac.uk/stampy
TopHat	https://ccb.jhu.edu/software/tophat/index.shtml
STAR	https://github.com/alexdobin/STAR
HiSAT2	http://ccb.jhu.edu/software/hisat2/index.shtml
Denovo assembly DNA/RNA	
Vcake	https://sourceforge.net/projects/vcake/
Velvet	https://www.ebi.ac.uk/~zerbino/velvet/
Trinity	https://sourceforge.net/projects/trinityrnaseq/
Trans-Abyss	https://github.com/bcgsc/transabyss
Variant detection and annotation	
GATK	https://software.broadinstitute.org/gatk/
VarScan	http://varscan.soureeforge.net/
SnpEff	http://snpeff.sourceforge.net/
SeattleSeq	http://snp.gs.washington.edu/SeattleSeqAnnotation138/
Differential expression	
Limma	http://bioconductor.org/packages/release/bioc/html/limma.html
Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/
DeSeq2	http://bioconductor.org/packages/release/bioc/html/DESeq2.html
EBSeq	http://bioconductor.org/packages/release/bioc/html/EBSeq.html
Metagenomics	
QIIME	http://qiime.org/
RDP-Pyro	https://rdp.cme.msu.edu/
Visualization	
IGV	http://software.broadinstitute.org/software/igv/
Circos	http://www.circos.ca/
Tablet	https://omictools.com/tablet-tool
Brig	http://brig.sourceforge.net/brig-in-action/
Cytoscape	http://www.cytoscape.org/
Web resources for NGS data	
Ensembl	http://asia.ensembl.org/index.html
ExAC Browser	http://exac.broadinstitute.org/

(continued)

Table 2.3 (continued)

Commonly used softwares in next generation data analysis	
1000 Genome	http://www.internationalgenome.org/
TCGA	http://cancergenome.nih.gov/
Gene Expression Omnibus	https://www.ncbi.nlm.nih.gov/geo/
Genome 10K	https://genome10k.soe.ucsc.edu

2.7 Applications of Next-Generation Sequencing

Earlier, hundreds of publications have been published in which next-generation sequencing is applied for a variety of applications in the field of genomics and transcriptomics. In accelerating biological and biomedical research, NGS technologies have been useful in a number of ways including whole genome sequencing, targeted sequencing, gene expression profiling, novel gene discovery, chromatin immunoprecipitation, etc. (Buermans and den Dunnen 2014).

2.7.1 Genomic Applications

NGS technology enabled a new era of genomic research through massively high-throughput sequencing data which has solved various research problems. It has provided the most comprehensive view of genomic information and associated biological implications. Using comparative genomics, one can obtain a correlation between variations and its associated clinical features. Through international efforts like UK10K and 1000 genome projects, it is now possible to find variations present in normal healthy individuals belonging to different ethnicities. Various applications of NGS in genomics can be described as:

- *Whole genome sequencing*: With the advent of NGS technologies, it is now possible to sequence genome of simple as well as complex organisms at a faster rate with much low cost. Personalized treatment plans can be offered by healthcare providers using WGS. Based on the variations present in genomic sequence with respect to controls or reference genome, one can predict probable predispositions toward disease in future. Healthcare professionals can suggest any lifestyle-related changes to avoid future complications (Shapiro et al. 2013).
- *Detection of rare variations*: Many international efforts are going on with a sole aim to catalog various kinds of variations like SNPs, mutations, indels, and copy number variations present in the genome. One such global effort is 1000 genome consortium which has cataloged more than 79 million variations in around 2500 individuals (Consortium 2015). Through exome sequencing, one can find mutations in genes which are responsible for rare Mendelian disorders such as sickle cell anemia, Miller's syndrome, as well as common diseases like obesity, diabetes, etc. Structural variations including copy number variations and indels

have been successfully identified in diseased vs. non-diseased individuals. Genome-wide variations identified using NGS techniques help in understanding why some people respond to some therapy easily while many cannot (Boycott et al. 2013).

- *Prenatal diagnosis of genetic diseases*: Sequencing technologies have been applied to detect biomarkers for genetic disorders including Down's syndrome (Palomaki et al. 2011), Edwards' syndrome, and many others. It is now possible to test for genetic abnormalities before the birth itself (Cram and Zhou 2016). Maternal cell-free plasma sequencing is done to detect various chromosomal anomalies in the fetus (Canick et al. 2013). This technique successfully detected 22q11.2 deletion syndrome, Down's syndrome, myotonic dystrophy, and various single gene disorders.
- *Transplantation*: The human leukocyte antigen (HLA) system is a gene complex encoding the major histocompatibility complex (MHC) proteins in humans. These proteins are responsible for the regulations of the immune system in humans. Differences in HLAs are the major cause of organ transplant rejection. Mapping the variations is important to identify the possible course of patient body in accepting or rejecting the transplant. Nowadays, doctors go for HLA-typing to find the suitable match for transplantation (Lan and Zhang 2015).
- *Forensics*: Genome sequencing can be used to find the suspected criminal from the proof like blood and hair obtained from the crime site. As every individual has unique DNA sequence, patterns obtained from sample can be used as proof to identify criminal. Similarly, DNA sequencing has been applied to find paternity of the child (Yang et al. 2014).
- *Population adaptation*: People are adapted to diverse environmental conditions. It is possible with NGS to catalog variations which help them to survive under extreme environments (Long et al. 2015). Classic example of one such kind of gene is EGLN1 whose variations are reported in literature which makes a person able to adapt in low oxygen conditions (Aggarwal et al. 2015).
- *Disease gene identification*: Different gene panels for disease like cancer are available which can detect presence of specific tumor in patients and help in planning a proper treatment for the same.

2.7.2 Transcriptomics Applications

All transcripts expressed by the genome in different tissues at different time points can be captured using RNA sequencing. It is now possible to map all transcribed regions in the genome with a great precision. Currently, two important publicly available databases, the Encyclopedia of DNA Elements (ENCODE) and Genotype-Tissue Expression (GTEx; The GTEx Consortium 2013), are used to map functional elements that can regulate gene expression in different human tissues. Various applications of transcriptome data sequencing are:

- *Gene expression quantification*: NGS technologies have been successfully applied to measure gene expression of thousands of genes at any given point of time. Through RNA sequencing, it is possible to measure expression levels of different transcripts as well. It is now possible to find quantitative expression in different biological conditions, in different cells as well as in different tissues. Genes which are expressed differently in diseased conditions as compared to a normal control can be identified using high-throughput expression quantification techniques (Chen et al. 2012).
- *Noncoding RNA quantification*: Noncoding RNAs (ncRNA) which include transfer RNA (tRNA), ribosomal RNA (rRNA), small nucleolar RNA, micro-RNA (miRNA), and small interfering RNA (siRNA) are not translated into proteins. However, ncRNA plays an important role in various posttranscriptional modifications. It is now possible to measure ncRNA with great precision. Long noncoding RNAs can be easily identified and are found to be associated with various neurological diseases like Alzheimer's and different cancer types (Brunner et al. 2012).
- *Transcript annotation*: RNA sequencing is capable of detecting novel transcript isoforms, promoter elements, and untranscribed regions which can be of functional importance in the genome (Trapnell et al. 2010).
- *Variant detection*: Allele-specific expression detection is very useful to find causal variations in various case control studies. It is now possible to detect tissue-specific transcript variants in different samples accurately.
- *Fusion detection*: A fusion transcript is a chimeric RNA containing exons from two or more different genes and has the potential to code for novel proteins. Through different RNA sequencing experiments, fusion transcripts have been found to be associated with different cancer types including breast and prostate cancer (Bao et al. 2014).

2.7.3 Epigenetics

The study of heritable gene regulation that does not involve DNA sequence itself is called epigenetics. Two major kinds of epigenetic modifications are DNA methylation and histone tail modifications. Epigenetic modifications are of prime importance in oncogenesis and development. These changes decide whether the genes will be turned on or off and ensure proper production of proteins in specific cells only (Holliday 2006).

- *DNA methylation*: Methylomics is the study of genome-wide DNA methylation patterns and their effect on gene regulation. In methylation, when methyl groups are added to a particular gene, that gene is turned off, and no protein is produced from it. Bisulfite sequencing is done to determine methylation patterns of DNA. Bisulfite treatment of DNA converts cytosine to uracil but leaves

5-methylcytosine residues unaffected. Therefore, only methylated residues will be retained. The Human Epigenome Project is an initiative to identify, catalog, and interpret genome-wide methylation patterns of all human genes in all major tissues (Eckhardt et al. 2004). Different methylation clusters are found to be present in cancer patients as compared to control (Soto et al. 2016).

- *Histone tail modification*: Histones are the proteins which package and order the DNA into structural units called nucleosomes. Chromatin immunoprecipitation sequencing (Chip-Seq) is used to analyze histone modifications which determine the accessibility of DNA to transcriptional regulators. It is widely used in gene regulatory networks to find transcription factors and any other protein interactions with DNA on a genome-wide scale. Transcription factors controlling the progression of disease in an individual have been identified through Chip-Seq; for example, GABP is a transcription factor and is a promoter for TERT gene and is found to be associated with multiple cancer types (Messier et al. 2016).

2.7.4 Metagenomics

Metagenomics is the branch of genomics which involves genetic analysis of microbial genomes contained within an environmental sample. NGS-based metagenome analysis has revolutionized our understanding of ecology around us. To reveal the importance of microorganisms that surrounds us, various international efforts such as the Human Microbiome Project (HMP) (Turnbaugh et al. 2007) and Human Gut Microbiome Project have been initiated worldwide. The main goal of all these efforts is to find out the association of changes in the human microbiome with human health and diseases. Various studies have shown the applications of metagenomics to microbial ecology and industrial biotechnology.

- *Human health*: Diet and nutrition intake are the most important identifiers of human health and both govern human microbiome too. Gut microbiome plays a major role in metabolic, nutritional, physiological, and immunological processes in the human body. Studies have shown that perturbations in intestinal microbiome have been associated with various diseases including obesity (Flint et al. 2014), inflammatory bowel syndrome (Kostic et al. 2014), and celiac disease (David Al Dulaimi 2015).
- *Bioremediation*: Biosurfactants are low molecular weight surface-active compounds mainly produced by bacteria, yeast, and fungi. They are used in agriculture for plant pathogen elimination and for increasing the bioavailability of nutrients for beneficial plant microbes. As metagenomics is culture-independent technique, it is used these days to find novel compounds associated with natural ecosystems (Edwards and Kjellerup 2013).
- *Ecology*: Microorganisms play an integral part in history and function of life on earth. Studies in metagenomics have provided valuable insights into the

functional ecology of the microbial community. For example, bacterial communities found in defecations of sea lions in Australia suggest that nutrient-rich feces of these lions are an important nutrient source for coastal ecosystems (Lavery et al. 2012).

- *Biofuel*: Due to diminishing fossil fuel reserves and increased CO_2 accumulation in the atmosphere, biofuels have been viewed as an alternative for sustainability and protecting the environment. Biofuels are fuels derived from biomass conversion like in the conversion of cellulose into cellulosic ethanol (Morrison et al. 2009). Lignocellulose represents the largest terrestrial carbon source on earth, but cannot be broken down without a combination of acids, industrial chemicals, and heat. Various fungi and bacteria have been identified that can enzymatically decompose lignocellulose to its monomeric compounds for use as carbon sources. Various metagenomic studies have identified the key genes and enzymes involved in lignocellulose digestion and conversion into biofuels (Chandel and Singh 2011; Hess et al. 2011; Xing et al. 2012).

NGS technologies are now considered a routine part in multi-omics research. Reduction in cost of sequencing per base facilitated sequencing technologies at different genomic centers and private companies. Low-cost and high-throughput methods are providing physicians with the tools to translate genomic knowledge into clinical practice. Due to current NGS technologies, major advances are possible in many areas especially in understanding and diagnosis of complex and rare diseases. As our understanding of genome variability increases, functional annotation of the genome will also rise. However, no advancements will prove fruitful without developing efficient algorithms which can transform sequence reads and data into meaningful information. There is a need for innovative bioinformatics methods for analysis and infrastructure to store available wealth of data. A year-by-year rise in the number of publications related to the field of NGS is a proof of its wide applicability and advancements. In the coming years, novel sequencing solutions are expected from additional sequencing providers.

References

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ et al (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res 28:E87

Aggarwal S, Gheware A, Agrawal A, Ghosh S, Prasher B, Mukerji M (2015) Combined genetic effects of EGLN1 and VWF modulate thrombotic outcome in hypoxia revealed by Ayurgenomics approach. J Transl Med 13:184

Anderson S (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res 9:3015–3027

Ansorge W, Sproat BS, Stegemann J, Schwager C (1986) A non-radioactive automated method for DNA sequence determination. J Biochem Biophys Methods 13:315–323

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M (1987) Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res 15:4593–4602

Astier Y, Braha O, Bayley H (2006) Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5′-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. J Am Chem Soc 128:1705–1710

Bao ZS, Chen HM, Yang MY, Zhang CB, Yu K et al (2014) RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. Genome Res 24:1765–1773

Bayley H (2010) Nanotechnology: holes with an edge. Nature 467:164–165

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol 33:623–630

Bowers J, Mitchell J, Beer E, Buzby PR, Causey M et al (2009) Virtual terminator nucleotides for next-generation DNA sequencing. Nat Methods 6:593–595

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14:681–691

Brunner AL, Beck AH, Edris B, Sweeney RT, Zhu SX et al (2012) Transcriptional profiling of long non-coding RNAs and novel transcribed regions across a diverse panel of archived human cancers. Genome Biol 13:R75

Buermans HP, den Dunnen JT (2014) Next generation sequencing technology: advances and applications. Biochim Biophys Acta 1842:1932–1941

Canick JA, Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE (2013) The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies. Prenat Diagn 33:667–674

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M et al (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611

Chandel AK, Singh OV (2011) Weedy lignocellulosic feedstock and microbial metabolic engineering: advancing the generation of 'biofuel'. Appl Microbiol Biotechnol 89:1289–1303

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY et al (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148:1293–1307

Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM et al (2014) The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell 158:1187–1198

Consortium GP (2015) A global reference for human genetic variation. Nature 526:68–74

Cram DS, Zhou D (2016) Next generation sequencing: coping with rare genetic diseases in China. Intract Rare Dis Res 5:140–144

David Al Dulaimi M (2015) The role of infectious mediators and gut microbiome in the pathogenesis of celiac disease. Arch Iran Med 18:244

Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M et al (2010) Nanopore DNA sequencing with MspA. Proc Natl Acad Sci U S A 107:16060–16065

Eckhardt F, Beck S, Gut IG, Berlin K (2004) Future potential of the human epigenome project. Expert Rev Mol Diagn 4:609–618

Edwards SJ, Kjellerup BV (2013) Applications of biofilms in bioremediation and biotransformation of persistent organic pollutants, pharmaceuticals/personal care products, and heavy metals. Appl Microbiol Biotechnol 97:9909–9921

Eid J, Fehr A, Gray J, Luong K, Lyle J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

Flint HJ, Duncan SH, Louis P (2014) Gut microbiome and obesity. In: Treatment of the obese patient. Springer, New York, pp 73–82

Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28:3169–3177

Foquet M, Samiee KT, Kong X, Chauduri BP, Lundquist PM et al (2008) Improved fabrication of zero-mode waveguides for single-molecule detection. J Appl Phys 103:034301

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T et al (2009) The challenges of sequencing by synthesis. Nat Biotechnol 27:1013–1023

Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet 13:840–852

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351

Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN et al (2016) Long-read sequence assembly of the gorilla genome. Science 352:aae0344

Grada A, Weinbrecht K (2013) Next-generation sequencing: methodology and application. J Invest Dermatol 133:e11

Gut IG (2013) New sequencing technologies. Clini Transl Oncol Off Publ Fed Span Oncol Soc Nat Cancer Inst Mex 15:879–881

Haque F, Li J, Wu HC, Liang XJ, Guo P (2013) Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. Nano Today 8:56–74

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107:1–8

Hershey AD, Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol 36:39–56

Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H et al (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331:463–467

Holliday R (2006) Epigenetics: a historical overview. Epigenetics 1:76–80

Kasianowicz JJ, Brandin E, Branton D, Deamer DW (1996) Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci U S A 93:13770–13773

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. Cell 155:27–38

Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM et al (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14:R101

Korlach J, Bibillo A, Wegener J, Peluso P, Pham TT et al (2008a) Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. Nucleosides Nucleotides Nucleic Acids 27:1072–1083

Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL et al (2008b) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci U S A 105:1176–1181

Kostic AD, Xavier RJ, Gevers D (2014) The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology 146:1489–1499

Lan JH, Zhang Q (2015) Clinical applications of next-generation sequencing in histocompatibility and transplantation. Curr Opin Organ Transplant 20:461–467

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lavery TJ, Roudnew B, Seymour J, Mitchell JG, Jeffries T (2012) High nutrient transport and cycling potential revealed in the microbial metagenome of Australian sea lion (Neophoca cinerea) faeces. PLoS One 7:e36478

Lee H, Gurtowski J, Yoo S, Nattestad M, Marcus S et al (2016) Third-generation sequencing and the future of genomics. bioRxiv:048603

Lelieveld SH, Veltman JA, Gilissen C (2016) Novel bioinformatic developments for exome sequencing. Hum Genet 135:603–614

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Liu H, He J, Tang J, Liu H, Pang P et al (2010) Translocation of single-stranded DNA through single-walled carbon nanotubes. Science 327:64–67

Liu L, Li Y, Li S, Hu N, He Y et al (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364

Loman NJ, Quick J, Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12:733–735

Long A, Liti G, Luptak A, Tenaillon O (2015) Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nat Rev Genet 16:567–582

Mardis ER (2006) Anticipating the 1,000 dollar genome. Genome Biol 7:112

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402

Mardis ER (2011) A decade's perspective on DNA sequencing technology. Nature 470:198–203

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74:560–564

McGinn S, Gut IG (2013) DNA sequencing – spanning the generations. New Biotechnol 30:366–372

McNally B, Singer A, Yu Z, Sun Y, Weng Z, Meller A (2010) Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. Nano Lett 10:2237–2244

Messier TL, Gordon JA, Boyd JR, Tye CE, Browne G et al (2016) Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes. Oncotarget 7:5094–5109

Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11:31–46

Mills L (2014) Common file formats. Current Protocols in Bioinformatics 45:A 1B 1–A 1B18

Morey M, Fernandez-Marmiesse A, Castineiras D, Fraga JM, Couce ML, Cocho JA (2013) A glimpse into past, present, and future DNA sequencing. Mol Genet Metab 110:3–24

Morrison M, Pope PB, Denman SE, McSweeney CS (2009) Plant biomass degradation by gut microbiomes: more of the same or something new? Curr Opin Biotechnol 20:358–363

Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE (2011) Landscape of next-generation sequencing technologies. Anal Chem 83:4327–4341

Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N et al (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinf Biol Insights 9:75–88

Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE, Neveux LM et al (2011) DNA sequencing of maternal plasma to detect down syndrome: an international clinical validation study. Genet Med 13:913–920

Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52:413–435

Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O et al (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 12:780–786

Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. Genomics 93:105–111

Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. J Hum Genet 59:5–15

Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. Nat Biotechnol 26:1117–1124

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

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19:R227–R240

Schneider GF, Kowalczyk SW, Calado VE, Pandraud G, Zandbergen HW et al (2010) DNA translocation through graphene nanopores. Nano Lett 10:3163–3167

Schubeler D (2015) Function and information content of DNA methylation. Nature 517:321–326

Service RF (2006) Gene sequencing. The race for the $1000 genome. Science 311:1544–1546

Shapiro E, Biezuner T, Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14:618–630

Sharon D, Tilgner H, Grubert F, Snyder M (2013) A single-molecule long-read survey of the human transcriptome. Nat Biotechnol 31:1009–1014

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C et al (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679

Soto J, Rodriguez-Antolin C, Vallespin E, de Castro CJ, Ibanez de Caceres I (2016) The impact of next-generation sequencing on the DNA methylation-based translational cancer research. Transl Res J Lab Clin Med 169(1–18):e11

The GTEx Consortium (2013) The genotype-tissue expression (GTEx) project. Nat Genet 45:580–585

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan 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

Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36–46

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI (2007) The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature 449:804

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S et al (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 18:1051–1063

Van Verk MC, Hickman R, Pieterse CM, Van Wees SC (2013) RNA-Seq: revelation of the messengers. Trends Plant Sci 18:175–179

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

Venter JC, Smith HO, Adams MD (2015) The sequence of the human genome. Clin Chem 61:1207–1208

Wang Y, Navin NE (2015) Advances and applications of single-cell sequencing technologies. Mol Cell 58:598–609

Watson M (2014) Quality assessment and control of high-throughput sequencing data. Front Genet 5:235

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737–738

Whiteford N, Skelly T, Curtis C, Ritchie ME, Lohr A et al (2009) Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics 25:2194–2199

Xing M-N, Zhang X-Z, Huang H (2012) Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis. Biotechnol Adv 30:920–929

Yang Y, Xie B, Yan J (2014) Application of next-generation sequencing technology in forensic science. Genomics Proteomics Bioinformatics 12:190–197

Zhao Q, Wang Y, Dong J, Zhao L, Rui X, Yu D (2012) Nanopore-based DNA analysis via graphene electrodes. J Nanomater 2012:4

Sequence Alignment

3

Benu Atri and Olivier Lichtarge

3.1 Introduction

Life originated on Earth about 3.5 billion years ago and evolved from the last universal common ancestor (LUCA) through processes of speciation and extinction (Fig. 3.1). The diversity that we see around us today is a result of years of evolution. A shared ancestry is evident in the morphological, biochemical, and genetic similarities of the organisms as we see today. Based on this observation, Charles Darwin, in 1859, submitted his groundbreaking theory of evolution by natural selection, in his book *On the Origin of Species*. Natural selection (Darwin 1859) (Fig. 3.2), mutation theory (de Vries 1900–1903), and the laws of inheritance (Mendel 1865; rediscovered by Correns 1950), along with the works of many notable geneticists, laid the foundation of what would later become the *modern evolutionary synthesis* or *modern synthesis theory*, which provides a recognized explanation of evolution. In the simplest of terms, evolution (species or macromolecular level) can be defined as the change in heritable traits of living organisms over generations. An updated version of the universal tree of life is presented by Forterre (2015).

Application of Darwin's speciation concept to a molecular level gave rise to the field of molecular evolution, which is the study of variations and evolution in the molecular components of a cell, e.g., DNA, RNA, and proteins (Zuckerland and

B. Atri (✉)
Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX, USA

O. Lichtarge
Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX, USA

Center for Computational and Integrative Biomedical Research (CIBR), Baylor College of Medicine, Houston, TX, USA

Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_3

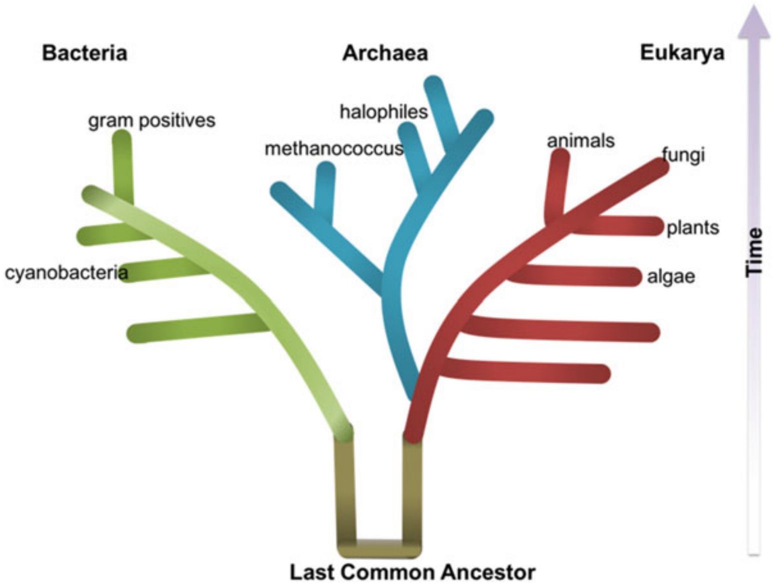


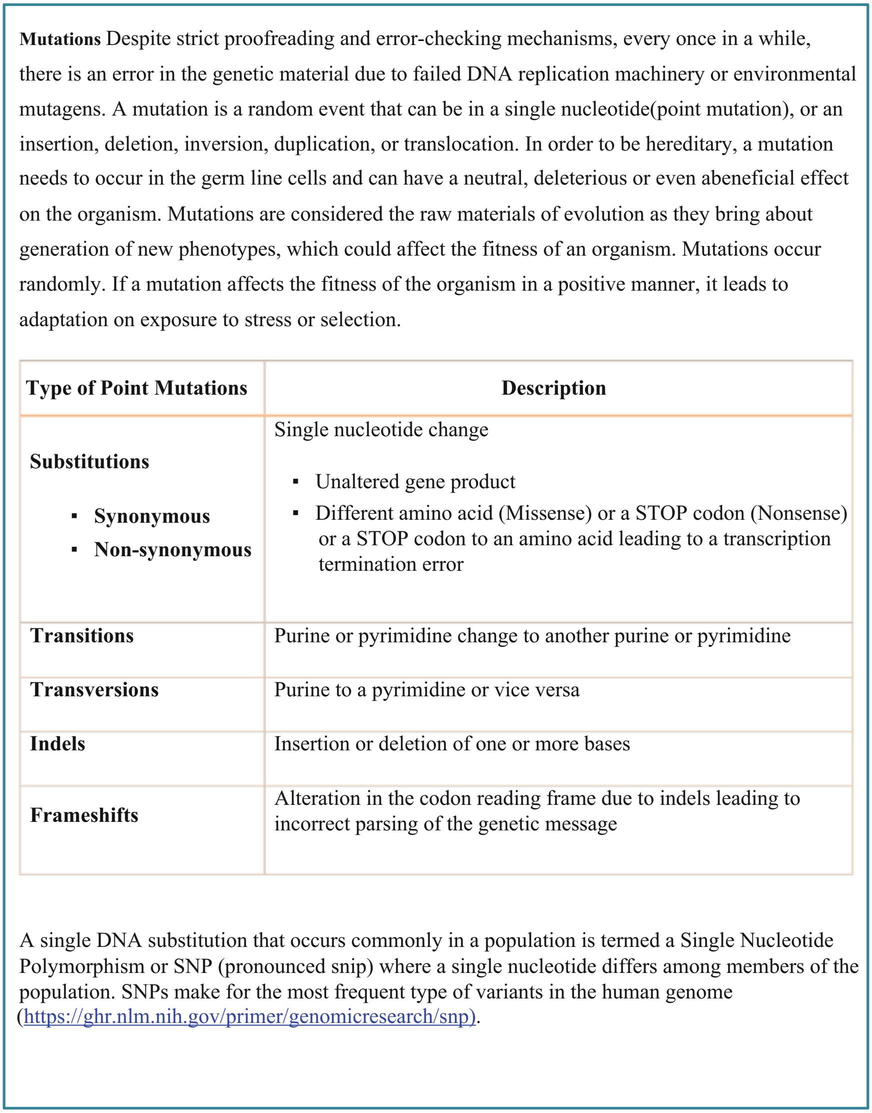
Fig. 3.1 The tree of life. (Adapted from Woese et al. 1990)

Natural Selection Algorithm

If, during the long course of ages and under varying conditions of life, organic beings vary at all in the several parts of their organization, and I think this cannot be disputed; if there be, owing to the high geometric powers of increase of each species, at some age, season or year, a severe struggle for life, and this certainly cannot be disputed; then, considering the infinite complexity of the relations of all organic beings to each other and to their conditions of existence, causing an infinite variety in structure, constitution, and habits, to be advantageous to them, I think it would be a most extraordinary fact if no variation ever had occurred useful to each being's own welfare, in the same way as so many variations have occurred useful to man. But if variations useful to any organic being do occur, assuredly individuals thus characterized will have the best chance of being preserved in the struggle for life; and from the strong principle of inheritance they will tend to produce offspring similarly characterized. This principle of preservation, I have called, for the sake of brevity, ***Natural Selection****.*

Fig. 3.2 Charles Darwin submitted his theory of evolution by natural selection. Darwin and Alfred Russel Wallace (On the Origin of Species, 1859) proposed a common descent and the tree of life, wherein divergence between species could have emerged from a common ancestor

Pauling 1965). These variations are usually the result of an accumulation of mutations, which are rightly called the raw material for evolution (Fig. 3.3). Molecular evolution studies brought to light that biological sequences (DNA, RNA, proteins) contain important evolutionary information, revealed by comparison. Technological breakthroughs such as DNA sequencing (Sanger et al. 1977)



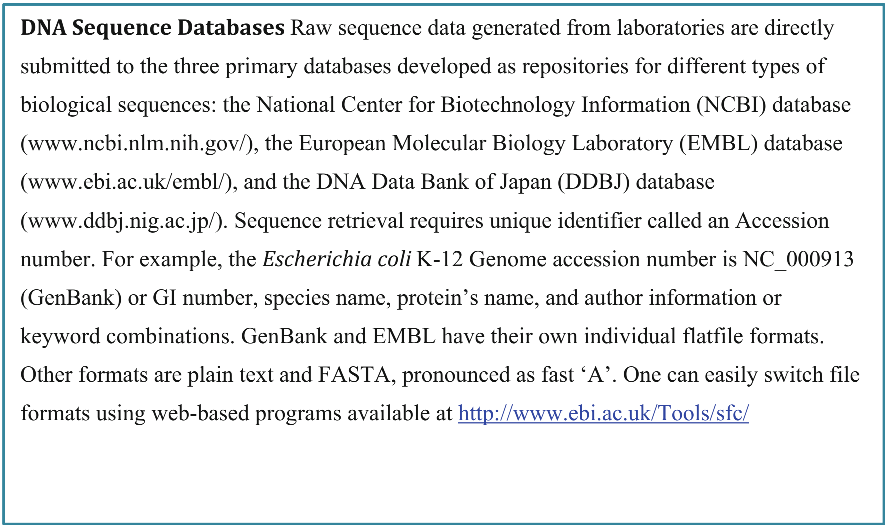
Mutations Despite strict proofreading and error-checking mechanisms, every once in a while, there is an error in the genetic material due to failed DNA replication machinery or environmental mutagens. A mutation is a random event that can be in a single nucleotide(point mutation), or an insertion, deletion, inversion, duplication, or translocation. In order to be hereditary, a mutation needs to occur in the germ line cells and can have a neutral, deleterious or even abeneficial effect on the organism. Mutations are considered the raw materials of evolution as they bring about generation of new phenotypes, which could affect the fitness of an organism. Mutations occur randomly. If a mutation affects the fitness of the organism in a positive manner, it leads to adaptation on exposure to stress or selection.

Type of Point Mutations	Description
Substitutions ▪ **Synonymous** ▪ **Non-synonymous**	Single nucleotide change ▪ Unaltered gene product ▪ Different amino acid (Missense) or a STOP codon (Nonsense) or a STOP codon to an amino acid leading to a transcription termination error
Transitions	Purine or pyrimidine change to another purine or pyrimidine
Transversions	Purine to a pyrimidine or vice versa
Indels	Insertion or deletion of one or more bases
Frameshifts	Alteration in the codon reading frame due to indels leading to incorrect parsing of the genetic message

A single DNA substitution that occurs commonly in a population is termed a Single Nucleotide Polymorphism or SNP (pronounced snip) where a single nucleotide differs among members of the population. SNPs make for the most frequent type of variants in the human genome (https://ghr.nlm.nih.gov/primer/genomicresearch/snp).

Fig. 3.3 Genetic variations are the genotypic differences between individuals of a population or between different populations. In order to be hereditary, these variations must occur in a germ line cell. There are several sources of genetic variations and they can occur at the genomic, chromosomal, or genic level

followed. Eventually, the deluge of data from sequencing and experiments made manual comparison impractical and slow, therefore requiring automation. Remarkable feats in computational biology (Hagen 2000) have made possible not only the storage and retrieval of large sequence data (Fig. 3.4) but also simultaneous



DNA Sequence Databases Raw sequence data generated from laboratories are directly submitted to the three primary databases developed as repositories for different types of biological sequences: the National Center for Biotechnology Information (NCBI) database (www.ncbi.nlm.nih.gov/), the European Molecular Biology Laboratory (EMBL) database (www.ebi.ac.uk/embl/), and the DNA Data Bank of Japan (DDBJ) database (www.ddbj.nig.ac.jp/). Sequence retrieval requires unique identifier called an Accession number. For example, the *Escherichia coli* K-12 Genome accession number is NC_000913 (GenBank) or GI number, species name, protein's name, and author information or keyword combinations. GenBank and EMBL have their own individual flatfile formats. Other formats are plain text and FASTA, pronounced as fast 'A'. One can easily switch file formats using web-based programs available at http://www.ebi.ac.uk/Tools/sfc/

Fig. 3.4 DNA sequence databases are repositories of different types of nucleotide sequences. All published sequence data are submitted to one of these three databases. These three organizations exchange data on a daily basis, so they essentially contain the same data. (International Nucleotide Sequence Database Collaboration, INSDC)

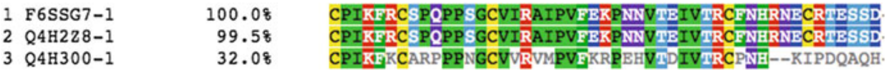
comparative analyses of many sequences. The differences in sequences due to mutations over time can be used to infer relationships between the sequences and their common ancestor.

Sequences evolve over time due to an accumulation of mutations causing them to diverge from one another (Alberts et al. 2002). Despite this divergence, biological sequences can still maintain enough similarity, which can be used to compare them and map their evolutionary history.

In this chapter, we will go over terminology, parts of and various ways to perform sequence comparison (sequence alignment), as well as how to find the optimal alignment for two or multiple sequences.

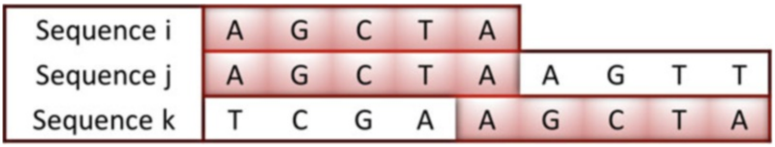
3.2 Sequence Alignment

The first step in annotating a new gene is to sequence it and then infer its function by finding similarities with genes of known function. For that, sequence alignment is performed, which involves arranging two or more sequences into rows, with characters aligned in successive columns (Fig. 3.5). Each element in an alignment is a match, a mismatch, or a gap. When a residue is aligned to an identical (or similar) residue, it is a match. Mismatches represent substitutions. If a position (column) is conserved (in multiple sequences) or has only conservative substitutions, it strongly suggests a functional or structural role of that position or region.



```
1 F6SSG7-1         100.0%     CPIKFRCSPQPPSGCVIRAIPVFEKPNNVTEIVTRCFNHRNECRTESSD
2 Q4H2Z8-1          99.5%     CPIKFRCSPQPPSGCVIRAIPVFEKPNNVTEIVTRCFNHRNECRTESSD
3 Q4H300-1          32.0%     CPIKFKCARPPPNGCVVRVMPVFKRPEHVTDIVTRCPNH--KIPDQAQH
```

Fig. 3.5 A sequence alignment is the arrangement of the biological sequences to identify regions of similarity that could be of functional, structural, or evolutionary importance. Shown above is a domain of the sequence of tumor suppressor protein p53 (F6SSG7–1) aligned to two of its homologs and visualized using the MView program (EMBL). Color coding, in this case, is based on identity and property of the residue and can be changed depending on the alignment program used



Sequence i	A	G	C	T	A				
Sequence j	A	G	C	T	A	A	G	T	T
Sequence k	T	C	G	A	A	G	C	T	A

Fig. 3.6 Sequence identity takes into consideration the smaller of the two sequences compared, which could be unreliable. On the left are three hypothetical nucleotide sequences, *i*, *j*, and *k*. If *i* is identical to *j* and *i* is identical to *k*, it is not necessary that *j* is identical to *k*. In this case, identity (*i*, *j*) = 100%, identity (*i*, *k*) = 100%, but identity (*j*, *k*) ≈ 71%

3.2.1 Sequence Identity, Sequence Similarity, and Sequence Homology

Sequence identity (percent identity) is the number of identical bases (DNA) or residues (amino acids) in an alignment at the same positions, relative to the length of the sequence:

$$\text{Sequence identity} = \frac{\text{Number of identical bases or residues}}{\min(\text{length i},\ \ \text{length j})} \times 100 \qquad (3.1)$$

Identity is not a very sensitive and reliable measure (Pearson 2014) because identity calculations do not include gaps. Another reason is that the shorter of the two sequences is used to measure identity, and, therefore, it is not transitive, i.e., if $A = B$ and $B = C$, then A does not necessarily equal C (Fig. 3.6).

Sequence similarity refers to similar residues/amino acids at corresponding positions (column of an alignment). In nucleotide sequences, sequence identity and sequence similarity mean the same. In proteins, amino acids can be more or less similar based on their physical, biochemical, functional, and structural properties. Similar substitutions (Table 3.1) in a protein may not affect the functional and structural properties of the protein. The protein tolerates such substitutions well. In an alignment, similar substitutions are not penalized as strictly when scoring the alignment. Sequence similarity is a measure of evolutionary distance and is often confused with sequence homology.

Sequence homology refers to having a common origin. Homology is an absolute qualitative term, i.e., it is either present or absent. Just as a bird's wing is homologous to a human hand (and not highly homologous), there is no gradient or percent or

Table 3.1 Substitution to a similar amino acid is not penalized as strictly as a mismatch

Side chain	Property	Similar amino acids
Hydrophilic	Positive	Lysine, arginine
	Negative	Aspartic acid, glutamic acid
Hydrophobic	Aliphatic	Isoleucine, leucine, valine, alanine
	Aromatic	Phenylalanine, tryptophan

Tools for exploring different properties of amino acids as well as commonly occurring substitutions are available via the NCBI http://www.ncbi.nlm.nih.gov/Class/Structure/aa/aa_explorer.cgi

degree of homology. The term homology is often incorrectly used quantitatively as X% homology or high homology. Since homology refers to common ancestry, two species cannot have a 50% common ancestor. This, however, needs to be distinguished from the specific case of large proteins where some domains of the protein share homology with a common ancestor and some do not.

Additionally, all inferred homology is just that inferred. It is a speculation. There is no way to confirm the common ancestor two sequences share. But we can use sequence identity or similarity to infer homology with higher confidence. Usually, a high sequence identity or similarity indicates that the two sequences did not originate independently of each other and, therefore, are possibly homologous (Koonin and Galperin 2003). Homologous sequences can be orthologs (genes or proteins from the same ancestor separated by a speciation event) or paralogs (genes/proteins within the same species separated by a duplication event).

3.2.2 Parts of a Sequence Alignment

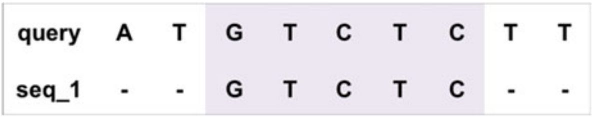
Query The sequence of interest, which is used to fetch and compare one or more homologous sequences. In any sequence analysis, identification of homologous sequences is usually the first step. A query (and its homologs) could be a complete sequence or short substring of a long sequence.

Sequence(s) Two (pairwise) or more (multiple) nucleotide or protein sequences

Substitution Scores Sum of scores for aligned pairs of characters. For substitution scores (for mismatches), we refer to substitution matrices.

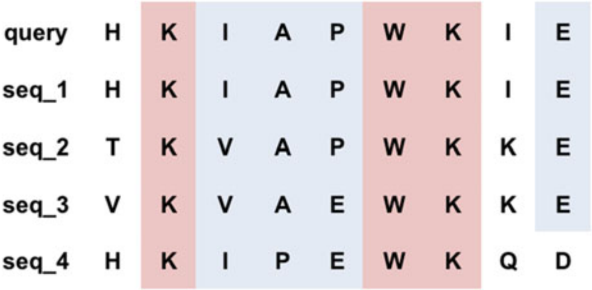
Gaps Gaps are introduced to maximize the matches in any column and obtain an *optimal* alignment. Gaps are a result of indels in sequences introduced over evolutionary time. On deleting a residue from one sequence, its absence generates a gap in the row for that sequence. When a residue gets inserted in one sequence, its presence generates a gap in the opposite sequence(s) Fig. 3.7.

A gap is represented by one or more "–" characters and can be placed in any sequence (query or homolog) to make the alignment optimal.



query	A	T	G	T	C	T	C	T	T
seq_1	-	-	G	T	C	T	C	-	-

Fig. 3.7 Introducing gaps can improve an alignment thereby revealing otherwise buried domains of similarity



query	H	K	I	A	P	W	K	I	E
seq_1	H	K	I	A	P	W	K	I	E
seq_2	T	K	V	A	P	W	K	K	E
seq_3	V	K	V	A	E	W	K	K	E
seq_4	H	K	I	P	E	W	K	Q	D

Fig. 3.8 In this example of a multiple sequence alignment, most evolutionarily conserved residue columns are highlighted red, and ones with similar residue substitutions are highlighted blue

Gap Penalty Sum of gap opening and gap extension scores (See Sect. 3.2.2).

3.2.3 Significance of Sequence Alignment

Sequence alignment is a useful tool to identify functional, structural, and evolutionary information from biological sequences. An alignment can reveal levels of similarity between sequences and a possible common ancestor (*homology*). Using sequence alignment, one can identify similar regions or motifs (Fig. 3.8) and similar functions or predict the 3D structure of proteins. Other applications include genome analysis, RNA secondary structure prediction, and database searching.

3.3 Achieving an Optimal Alignment

To achieve an optimal sequence alignment suited for our biological problem, we should answer the following questions leaving little room for assumptions as possible:

What do we want to align?
What is the best scoring strategy?
Which method to use?

3.3.1 What Kind of Sequences Do We Want to Align?

It is important to understand the problem at hand as it will help us decide whether we are looking to align entire sequences (*global*) or simply substrings/subsequences/motifs of the sequences (*local*).

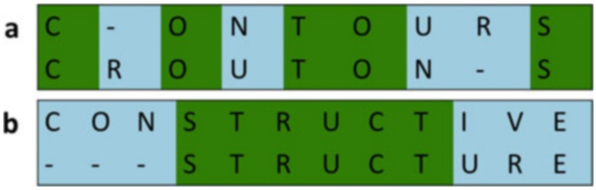


Fig. 3.9 Global alignments (**a**) find similarity over the whole length of the sequences, while local alignments (**b**) focus on identifying domains of similarity

Global sequence alignment (Fig. 3.9) covers the entire length of sequences involved when the sequences are roughly the same length and are reasonably similar. They are used to identify similar DNA and protein sequences (homologs) that suggest a similar biological role as well as, in the case of proteins, to identify similar 3D structures.

Local sequence alignments cover parts of the sequence and are used to compare small segments of all possible lengths when sequences have domains or regions of similarity and have different overall lengths. Local alignments are useful for identification of common motifs or domains in DNA and protein sequences that globally appear different.

3.3.2 What Is the Best Scoring Strategy?

Sequence alignment score is the column-by-column sum of the aligned letters, which could be matches or mismatches, using scoring or *substitution matrices.* Gaps in an alignment incur a *gap penalty score* subtracted from the total.

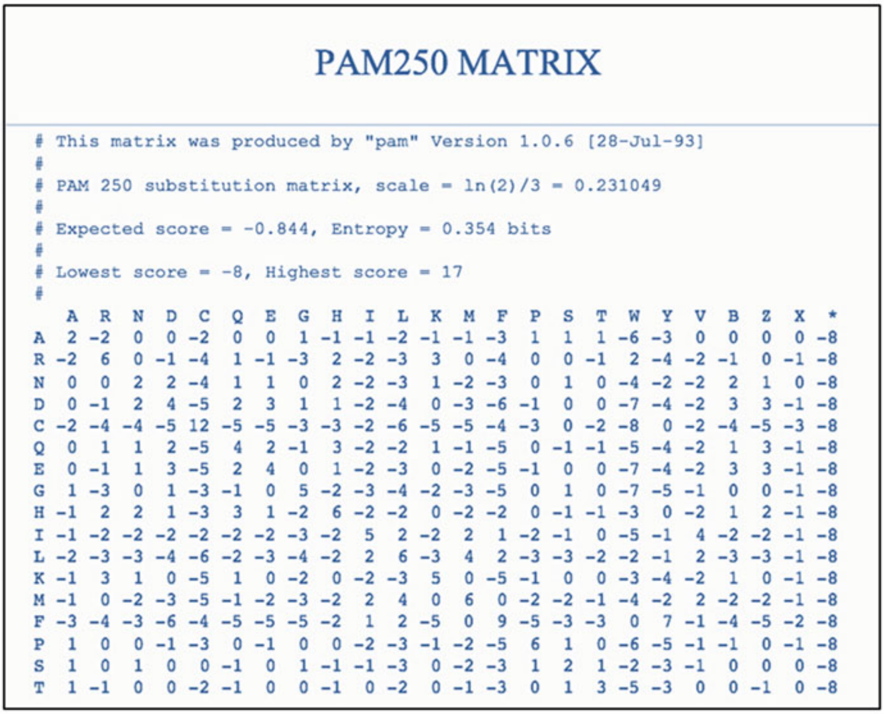
3.3.2.1 Substitution Matrices

Substitution matrices serve as lookup tables with substitution scores for aligning a pair of residues. A substitution matrix consists of rates or probabilities at which one amino acid (or nucleotide) substitutes or mutates into another.

For nucleotides, the likelihood of substitution between bases varies, e.g., transitions (between two purines or two pyrimidines) are more frequent than transversions (between a purine and a pyrimidine). Nucleotide scoring matrices comprise of two parameters – mismatch penalty and gap (or indel) penalty.

Amino acid substitution matrices are probability ratios, usually written in the form of a log-likelihood ratio (or log odds ratio), defined as the likelihood of one amino acid substituting to another (observed over the expected) based on available mutation data:

$$\text{Log odds ratio} = \log_2\left(\frac{\text{observed}}{\text{expected}}\right) \tag{3.2}$$



PAM250 MATRIX

```
# This matrix was produced by "pam" Version 1.0.6 [28-Jul-93]
#
# PAM 250 substitution matrix, scale = ln(2)/3 = 0.231049
#
# Expected score = -0.844, Entropy = 0.354 bits
#
# Lowest score = -8, Highest score = 17
#
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  2 -2  0  0 -2  0  0  1 -1 -1 -2 -1 -1 -3  1  1  1 -6 -3  0  0  0  0 -8
R -2  6  0 -1 -4  1 -1 -3  2 -2 -3  3  0 -4  0  0 -1  2 -4 -2 -1  0 -1 -8
N  0  0  2  2 -4  1  1  0  2 -2 -3  1 -2 -3  0  1  0 -4 -2 -2  2  1  0 -8
D  0 -1  2  4 -5  2  3  1  1 -2 -4  0 -3 -6 -1  0  0 -7 -4 -2  3  3 -1 -8
C -2 -4 -4 -5 12 -5 -5 -3 -3 -2 -6 -5 -5 -4 -3  0 -2 -8  0 -2 -4 -5 -3 -8
Q  0  1  1  2 -5  4  2 -1  3 -2 -2  1 -1 -5  0 -1 -1 -5 -4 -2  1  3 -1 -8
E  0 -1  1  3 -5  2  4  0  1 -2 -3  0 -2 -5 -1  0  0 -7 -4 -2  3  3 -1 -8
G  1 -3  0  1 -3 -1  0  5 -2 -3 -4 -2 -3 -5  0  1  0 -7 -5 -1  0  0 -1 -8
H -1  2  2  1 -3  3  1 -2  6 -2 -2  0 -2 -2  0 -1 -1 -3  0 -2  1  2 -1 -8
I -1 -2 -2 -2 -2 -2 -2 -3 -2  5  2 -2  2  1 -2 -1  0 -5 -1  4 -2 -2 -1 -8
L -2 -3 -3 -4 -6 -2 -3 -4 -2  2  6 -3  4  2 -3 -3 -2 -2 -1  2 -3 -3 -1 -8
K -1  3  1  0 -5  1  0 -2  0 -2 -3  5  0 -5 -1  0  0 -3 -4 -2  1  0 -1 -8
M -1  0 -2 -3 -5 -1 -2 -3 -2  2  4  0  6  0 -2 -2 -1 -4 -2  2 -2 -2 -1 -8
F -3 -4 -3 -6 -4 -5 -5 -5 -2  1  2 -5  0  9 -5 -3 -3  0  7 -1 -4 -5 -2 -8
P  1  0  0 -1 -3  0 -1  0  0 -2 -3 -1 -2 -5  6  1  0 -6 -5 -1 -1  0 -1 -8
S  1  0  1  0  0 -1  0  1 -1 -1 -3  0 -2 -3  1  2  1 -2 -3 -1  0  0  0 -8
T  1 -1  0  0 -2 -1  0  0 -1  0 -2  0 -1 -3  0  1  3 -5 -3  0  0 -1  0 -8
```

Fig. 3.10 A PAM matrix has 20 rows and 20 columns – each representing 1 of the 20 amino acids. Each cell at position (*i*, *j*) in a PAM matrix is the probability of the amino acid in the row, *i*, to be substituted by the amino acid in column *j*, over a given evolutionary time. PAM can be treated as a unit of evolution, defined as 1 accepted point mutation per 100 amino acids, after a particular evolutionary interval/distance. Based on this evolutionary distance, there can be different PAM matrices. Usually this distance is denoted by a number associated with the acronym PAM, e.g., PAM1 refers to a substitution matrix of proteins with an evolutionary distance of 1% mutation/position. Most commonly, PAM250 is used as the default matrix in similarity search programs (Dayhoff and Schwartz 1978)

In a substitution matrix (Fig. 3.10), amino acids are written across the top and along the side, and each cell in the matrix is a probability score of how one amino acid (row) would substitute to the other (column). The likelihood information about which amino acid substitutions are most and least common allows us to construct such matrices and score alignments. The two most common amino acid substitution matrices are the point accepted mutation (PAM) and blocks substitution matrix (BLOSUM).

Point Accepted Mutation or PAM Matrices Developed using 71 families of closely related proteins containing ~1500 observed mutations (Dayhoff and Schwartz 1978). These were used to calculate mutation rates, which in turn were used to model evolutionary relationships. The proteins were selected based on high sequence identity (85%) to ensure that the mutation observed was one mutation only and not the result of successive mutations. The term *point accepted mutation* refers to a point mutation of one amino acid to another that does not change the function of the protein significantly and is accepted by natural selection. For scoring protein sequence alignments, the amino acid substitution probabilities are reported in a normalized, logarithmic form. Therefore, PAM matrices are log odds matrices (Fig. 3.10).

Blocks Substitution Matrix or BLOSUM Based on local multiple alignments of evolutionarily distant proteins, Henikoff and Henikoff (1992) developed a series of blocks substitution matrices using 2000 blocks of aligned sequence segments characterizing more than 500 groups of related proteins.

To generate a BLOSUM matrix, multiple alignments of short regions of related sequences are first grouped. In each of these alignments, the sequences similar at some threshold percent identity are grouped and averaged. Such a group of sequences is called a block. This grouping is done to reduce the bias produced by using closely related sequences. After that, substitution frequencies for all pairs of amino acids are calculated, thereby generating log odds for BLOSUM. Instead of looking at the whole, the focus is on blocks of conserved sequences. These blocks have functional or structural importance. Usually wriiten as BLOSUMX, where the X represents the threshold percent identity of sequences clustered in that block. For example, in the BLOSUM62 matrix, the sequences grouped in a block are 62% identical. BLOSUM62 is the default matrix to score the alignment in most similarity search tools. Selection of scoring matrix depends on the nature of sequences to be aligned (Fig. 3.11).

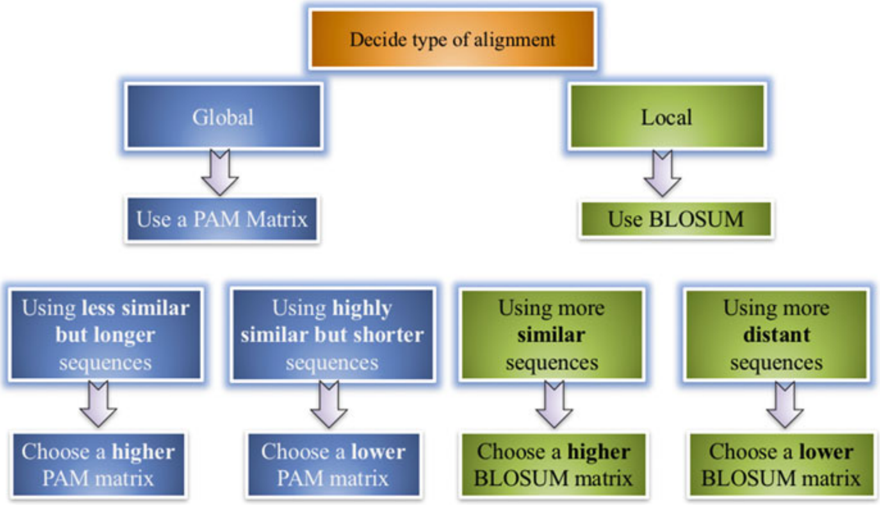

Decide type of alignment
Global
Local
Use a PAM Matrix
Use BLOSUM
Using less similar but longer sequences
Using highly similar but shorter sequences
Using more similar sequences
Using more distant sequences
Choose a higher PAM matrix
Choose a lower PAM matrix
Choose a higher BLOSUM matrix
Choose a lower BLOSUM matrix

Fig. 3.11 How to choose a substitution matrix?

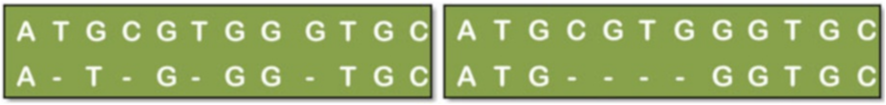


Fig. 3.12 In this example, it is more likely that an insertion of CGTG occurred once in the sequence on top, rather than four individual insertions. That is why a gap extension penalty is less than a gap opening penalty. It also encourages algorithms to introduce gaps of more than one length, if that improves the alignment

3.3.2.2 Gap Penalties

For best alignments, possible insertions and deletions must be considered. Since indels are expected to occur less often than substitutions, an introduction of a gap in a sequence incurs a penalty, known as the gap penalty. This penalty is to ensure that the total number of gaps does not get out of hand. Gap penalties are subtracted from the alignment score. When a new gap is opened, a gap opening penalty is incurred. If the gap is extended, a gap extension penalty is considered (Fig. 3.12). Smith and Waterman (1981) argued that a single mutational event creating an insertion (and, therefore, a gap) of x number of adjacent residues is more likely than many nonadjacent mutations.

The following equation represents the simplest implementation of gap scores:

$$G_{\text{total}} = G_{\text{o}} + G_{\text{e}}.L \tag{3.3}$$

G_{total} = Total gap penalty, G_{o} = Gap opening penalty, G_{e} = Gap extension penalty, L = total number of gaps – 1

3.3.3 Which Alignment Method to Use?

The choice of alignment method to use for sequence analysis depends on the dataset, i.e., whether we want to align two or multiple sequences.

3.3.3.1 Pairwise Sequence Alignment

It refers to the alignment of two sequences. The top three techniques for generating pairwise alignments are dot-matrix methods, dynamic programming, and word methods.

Dot Matrix (Gibbs and McIntyre 1970)

A dot-matrix plot is a 2D matrix with each of the two sequences written along the top row and left column (Fig. 3.13). A match between two characters is shown by a dot (hence the name). A line along the diagonal reflects that the pair of sequences has high similarity.

A dot-matrix plot can be used to identify regions of similarity, indels, repeats, and inverted repeats. Moreover, a dot plot can also be used to detect self-complementary

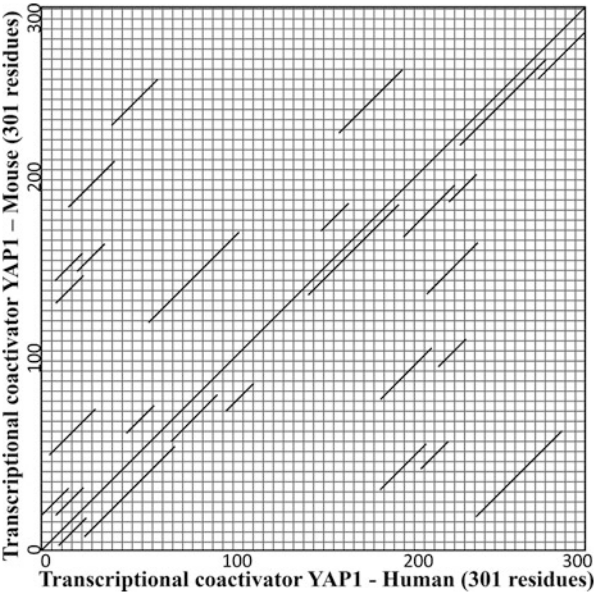


Fig. 3.13 A dot matrix provides a useful visualization tool in the form of a 2D matrix, for comparing a sequence against itself (for self-complementary regions) or two sequences against each other (for similarity, indels, repeats, and inverted repeats). In this example, the matrix compares yes-associated protein 1 from two species – mouse and human. The prominent diagonal reflects high sequence similarity between the two sequences (www.bioinformatics.nl)

regions in an RNA sequence. However, dot plots are limited to only two sequences, and there is an inherent noise which can be reduced by using a sliding window.

Dynamic Programming

Dynamic programming, a very popular bioinformatics optimization method, involves dividing a large problem into smaller subproblems and then using the individual solutions to reconstruct the solution for the larger problem. Dynamic programming algorithms also find use in gene recognition, RNA structure prediction, etc.

Needleman-Wunsch Dynamic Programming Algorithm for Global Alignment

Needleman and Wunsch (1970) applied dynamic programming for the first time to solve a biological problem, specifically, biological sequence alignment. The algorithm consists of three main steps:

Matrix Initialization

Consider two sequences AGACTAGT and CGAGACGT, and create a 2D matrix or grid as shown with each of the two sequences written across the top starting at the third column and down along the left side starting at the third row. In order to initialize the matrix, C, a scoring system needs to be established for the matches, mismatches, and penalty for gaps (Fig. 3.14). One such scoring system could be matches = 1, mismatches = −1, and gaps = −1.

To initialize, place a 0 in the cell at second row, second column or C (2,2).

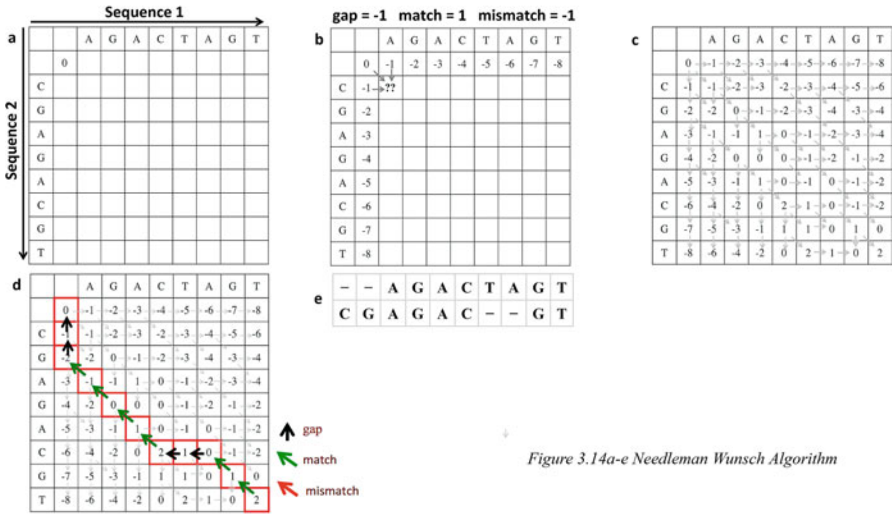


Fig. 3.14 (**a**–**e**) Needleman-Wunsch algorithm

Filling the Matrix

Next, move through the matrix row by row filling the cells. For the second row and second column (after the sequences), follow this formula:

$$C\ (i,j) = \max\left\{\begin{matrix} C(i-1,j-1)+S(a_i,b_j), \\ C(i-1,j)-g \\ C\ (i,j-1)-g \end{matrix}\right\}, \tag{3.4}$$

where

$i =$ row, $j =$ column
$S(a_i, b_j) =$ match/mismatch score between column i of sequence a and row j of sequence b
$g =$ gap penalty

For $i = 2, j = 2$ since there is no $i{-}1, j$ or $i, j{-}1$, take C (2,2) as 0, and fill the second row and second column as shown. Add $a - 1$ for every shift to the right for the second row and $a - 1$ for every down in the second column (Fig. 3.14b, c).

Traceback

Trace back the steps from the last cell (lowermost right) of the matrix back to the 0 at the origin, following the arrows. If two steps can be taken, both can be considered to generate different alignments. A diagonal arrow is a match (or mismatch), an arrow going up is a gap in the sequence on the top, and an arrow going left will introduce a gap in the sequence on the left (Fig. 3.14d). Based on the traceback, the optimal alignment generated is shown in Fig. 3.14e.

Smith-Waterman Dynamic Programming Algorithm for Local Alignment
Smith and Waterman (1981) developed a dynamic programming approach for local alignments with a few differences from the Needleman-Wunsch method:

All elements of the first column and the first row are set to 0.
Traceback starts at the cell with the highest score.

All negative cells are set to 0, which is how best aligned local segments show up. In order to do this, the formula for calculation of each $C\ (i, j)$ value includes 0 as one of the options:

$$C\ (i, j) = \max \left\{ \begin{array}{c} 0, \\ C(i-1, j-1) + S\left(a_i, b_j\right), \\ C(i-1, j) - g \\ C\ (i, j-1) - g \end{array} \right\}, \tag{3.5}$$

Word Based

These methods are also known as k-tup or k-tuple methods, designed to be more efficient than dot matrix or dynamic programming methods but do not necessarily find the most optimal alignments. Therefore, these methods find use in large database searching. As the name suggests, word-based methods use words or windows of short (k-length), nonoverlapping subsequences in the query sequence, which are matched to each of the sequence in the database. Two of the most commonly used implementations of the word-based methods are FASTA and BLAST.

FASTA David J. Lipman and William R. Pearson developed FASTA (pronounced FAST-“A”; Lipman and Pearson 1985) as a sequence alignment software. The FASTA format (where the file description starts with a “>”), which is now used significantly in bioinformatics, was first defined in this software.

FASTA takes a DNA or protein sequence as input and searches sequence databases using local alignment to find a match. To increase the speed of searching, FASTA employs a word search to narrow down segments of hits (located close to one another). *K-tup* specifies the size of the word and controls the sensitivity and speed of the search. A larger word size or k-tup will generate fewer hits (or matches), and a smaller k-tup will lead to a more sensitive search, e.g., a k-tup of 3 for BIOLOGY would generate the following words:

BIO
IOL
OLO
LOG
OGY

The program then looks for segments with clusters of nearby hits and scrutinizes them for a complete match. After a word-by-word search, FASTA goes on to perform a more detailed and optimized search based on a local alignment and Smith-Waterman algorithm for a query sequence to every sequence in the database.

FASTA Programs

Initially developed for protein sequence similarity search, the *FASTA programs* (collective term for the suite) are now used to find regions of local or global similarity between proteins or DNA along with the statistical significance of matches.

A complete list of FASTA programs and a web-based server can be found at http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml.

BLAST Basic Local Alignment Search Tool is also an algorithm for comparing nucleotide and/or protein sequences developed by Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David J. Lipman, and their publication is one of most cited (over 64,000 citations) in the field of bioinformatics (Altschul et al. 1990). BLAST was developed to be faster than FASTA without compromising sensitivity. However, unlike Smith-Waterman, BLAST may not always yield the most optimal alignments. Like other methods, BLAST compares a query to a large database of sequences to detect homologs or a set of sequences that have a certain level of similarity with the query. A large number of databases are available to customize searches; for example, one can choose to restrict the search to a single type of organism. To make the right choice, the user should compare methodologies and software available and choose the one best suited to answer their biological questions within the constraints of the assumptions and limitations of the method used.

Input (or query) is a FASTA-formatted sequence, accession number, or GI number. The different tabs on the top – blastn, blastp, blastx, tblastn, and tblastx – refer to various types of BLAST searches. Once complete, a BLAST report is generated. One or more sequences can be downloaded in different formats, to perform multiple sequence alignments or generate a phylogenetic tree.

Types of Scores in BLAST

Score, S: Describes the overall quality of the alignment between the query and the hit. Higher scores correspond to a higher-quality alignment and can be used to infer similarity.

Bit-Score: It is the log scale normalized version of raw score *S*, reported in units called *bits*:

$$S' = \frac{\lambda S - \ln K}{\ln 2} \tag{3.6}$$

Important Difference Between Score and Bit-Score

Raw scores are dependent on the size of the search space; therefore because of scale variations, raw scores of different alignments cannot be compared. On the other hand, a bit-score is normalized for the scoring system and, therefore, does not depend on the size of the search space. Bit-scores, therefore, can be used to compare different alignments from various searches.

E-Value The number of BLAST hits expected by chance.

A higher score means a sequence is less likely to have been picked up randomly; therefore, there is a higher possibility of biologically meaningful relationships between the sequences.

The E-value is calculated as follows:

$$E = Kmne^{-\lambda S} \tag{3.7}$$

where m, n = sequence lengths

K, λ = parameters that act as scaling depends on the scoring system (substitution matrix), gap penalties, and the size of the search space. The most widely used substitution matrix is BLOSUM62.

E-value is used to find whether an alignment is meaningful. E-value, however, is not homology. BLAST does not predict homology, but can be used to infer homology. E-value depends on the query and the size of the database, so no specific E-value could be used to determine homology or significance of hits. There is no gold standard cutoff or universal threshold. Every search is different, and user must decide which E-value to consider best based on their *query*, goals, and the biological question asked. It is why two E-values cannot be compared when searching different databases, i.e., an E-value from one query against a certain database cannot be compared to or used to draw similar inferences when using a different database. If the E-value is less than 1e-179, it is usually reported as 0.0.

Relationship between score and E-value: The higher the score S, the lower is the E-value.

Advantages of BLAST

It detects local regions as well as global alignments.

It can be used to infer homology.

It can provide insights into the function of new/uncharacterized proteins.

It is faster than FASTA without losing sensitivity.

It is worth stating here that the underlying algorithms for FASTA and BLAST have a long history of development and use. New variants continue to be added to the FASTA and BLAST suites. These methods are constantly evolving with time and continue to find wider applications in biology and bioinformatics.

It is important to keep in mind that the results of any searches made using the above tools will depend upon the query, databases used, and the goal of research. Different inputs will result in different results.

3.3.3.2 Multiple Sequence Alignment

Multiple sequence alignment (MSA) is an alignment of more than two sequences and is most commonly used to study similarity/differences between several homologs. MSA methods attempt to find the best alignment between all sequences, thereby detecting conserved regions, which could have critical structural and functional importance including catalytic sites of enzymes.

MSAs are used as inputs for the construction of phylogenetic trees. For pairwise sequence alignment, coming up with a scoring system with matches, mismatches, and penalties for gaps is pretty straightforward. However, scoring gets complicated in MSA. One way to score is to find the sum of pairwise alignment scores of pairs of columns. MSA algorithms have two most important characteristics that should be considered while making a choice of which method to use: accuracy and computational complexity (running time as well as space requirements).

Applications

Multiple sequence alignments reveal evolutionary history by allowing us to find biological importance in a set of sequences, which are not necessarily close.

Detect regions of similarity or variability between members of a protein family.

Critical residue identification and making significant structural and functional inferences.

Serve as the first step in most phylogenetic reconstructions.

The two approaches to generating multiple sequence alignments include progressive and iterative MSAs. Progressive MSA starts with a single sequence and progressively adds the others to the alignment, while iterative MSA realigns the sequences during multiple iterations of the process.

Progressive Methods (Feng and Doolittle 1987)

These methods have the underlying assumption that a high level of similarity between sequences indicates evolutionary relatedness. They are also sometimes known as *hierarchical* or *tree-based* methods.

Steps

Using Needleman-Wunsch algorithm for global alignment, calculate evolutionary distance between every pair of sequences.

Generate a reference phylogenetic tree or a guide tree.

Starting from the two closest branches of this tree, generate a consensus sequence, which is then used as a proxy for the pair of branches.

Progressively (hence the name), this is repeated for the next closest pair of branches until all sequences of the query set have been added to the consensus. Usually computed between a query and a subject, pairwise alignments can also include a query-query, query-consensus, or consensus-consensus pairs of sequences.

The choice of the two closest branches early on in the consensus calculation has a higher weight in the overall alignment quality and, therefore, should be made very carefully. If distant branches are included before all close branches have been included, the overall quality of the alignment will go down drastically. The reference phylogenetic tree, therefore, serves as a good way to guarantee that this mistake is not made.

Advantages: Faster and more efficient than dynamic programming methods
Disadvantages: Heuristic and needs high accuracy

Error propagation: Since, this method is progressive in nature, errors can get propagated.

Most popular software that employ progressive methods of multiple sequence alignment include Clustal suite, MUSCLE, T-Coffee, and MAFFT.

Clustal (Higgins and Sharp 1988) The Clustal suite is the most commonly used method for MSA. It includes several versions of Clustal which is the original version for progressive alignment based on a reference guide tree. Some incarnations include:

Clustal V – serves to combine different phases of progressive alignment.

Clustal W (Thompson 1994) – in addition to what Clustal V does, this adds other features like sequence weighting, position-specific gap penalties, and choice of weight matrix to be used. In Clustal W, sequences with high similarity are placed close on the reference/guide tree and, therefore, get added to the alignment and consensus sequences much earlier than the more divergent sequences.

Higgins and Sharp noticed that adding too many similar sequences early on can lead to a bias in the way the reference tree is generated. In an attempt to correct this, sequence weighting was introduced. Groups of similar sequences are given lower weights as compared to groups of more different/divergent sequences. A user has the choice of substitution matrices PAM or BLOSUM as well as the freedom to adjust gap penalties (based on position, content, and length of sequences).

Input NBRF/PIR, FASTA, EMBL/Swiss-Prot, Clustal, GCC/MSF, etc.

Steps

Perform progressive pairwise alignment.

Create a guide/reference tree.

Use the tree to create a multiple sequence alignment.

Output Format Clustal, NBRF/PIR, GCG/MSF, PHYLIP, GDE, or NEXUS.

Available at http://www.ebi.ac.uk/Tools/msa/clustalw2/

Clustal Omega (Sievers et al. 2011) is a new high-quality aligner that uses seeded guide trees and hidden Markov model profile-profile techniques to generate alignments for multiple sequences. Input can be GCG, FASTA, EMBL, GenBank, PHYLIP, etc., and output is an alignment in Clustal, FASTA, MSF, NEXUS, etc. It is available at http://www.ebi.ac.uk/Tools/msa/clustalo/.

The European Bioinformatics Institute (EBI) provides access to a large number of databases and analysis tools. Detailed overview of all EMBL-EBI services are presented in Li et al. (2015) and McWilliam et al. (2013).

MUSCLE (Edgar 2004) MUSCLE stands for MUltiple Sequence Comparison by Log-Expectation. MUSCLE is used to obtain better alignments than Clustal, which depends on the options chosen. For larger alignments, MUSCLE is a slightly faster method. It uses two distance measures for a pair of sequences: a kmer distance (for an unaligned pair) and the Kimura distance (for an aligned pair).

Input FASTA format

Steps

Draft Progressive Stage: It produces a rough draft of a multiple alignment from a guide tree with more stress on speed. The kmer distance is computed for each pair of input sequences.

Improved Progressive Stage: The tree is re-estimated using Kimura distance to avoid any errors stemming from approximated k-mer distance measure. Kimura distance is more accurate but requires an alignment and produces a better tree.

Refinement Stage: Refines the alignment produced above.

Output FASTA, Clustal W, MSF, and HTML formats.

Available at http://www.drive5.com/muscle.

T-Coffee (Notredame et al. 2000) T-Coffee stands for Tree-based Consistency Objective Function for alignment Evaluation. Progressive alignment methods tend to suffer from errors because of the greedy approach of including all sequences one by one, progressively. T-Coffee was developed as an implementation of the progressive alignment method, which could correct the propagation of errors, and on average produces more accurate alignments than the other methods. First, it creates a whole library of global as well as local pairwise sequences to guide the multiple sequence alignment. Next is an optimization step that is used to find the best fitting multiple sequence alignment. However, in an attempt to correct for errors, this method sacrifices speed and, therefore, is not considered suitable for larger datasets.

Input FASTA and PIR are supported.

Output As a default, the output is generated in the *aln format* (Clustal) but also produces other formats.

Available at http://www.tcoffee.org/

MAFFT (Katoh et al. 2002) Acronym for Multiple Alignment using Fast Fourier Transform. MAFFT is a multiple sequence alignment program for DNA and protein sequences.

Input FASTA format

MAFFT identifies homologous regions by using fast Fourier transform where an amino acid sequence is changed to a sequence consisting of each amino acid's volume and polarity values. This method uses a simplified scoring system to reduce CPU time and works well with distant sequences or those with large indels. This method continues to go through improvements in accuracy, speed, and applications (Katoh and Standley 2013).

Available at: http://mafft.cbrc.jp/alignment/software/

Iterative Methods

Iterative methods use multiple iterations to realign the sequence several times. They start with a pairwise realignment of sequences within subgroups and then realign the subgroups. Subgroups are chosen based on sequence relationships as seen on the guide tree or randomly. Iterative methods try to correct for the overdependence of progressive methods on the accuracy of the seed pairwise alignment. Most conventional approaches to iterative multiple sequence alignment use machine learning algorithms such as genetic algorithms and hidden Markov models (Churchill 1989; Baldi et al. 1994; Krogh et al. 1994) and have been extensively reviewed (Thompson et al. 2011).

3.3.3.3 Whole Genome Alignment and Visualization

Whole genome alignment (WGA) is the alignment of two or more genomes at DNA level. It is an amalgamation of linear sequence alignment and gene ortholog predictions. However, rapidly growing databases of whole genome sequences, the sheer size, and complexity of whole genomes make WGA a challenging analysis to perform. Another difficulty in alignment of genomes arises from the fact that all genomes undergo complex rearrangements and structural changes, such as duplications (Dewey 2012). Despite all these facts, WGAs are a powerful method of alignment since they allow for both large- and small-scale study of molecular evolution. On a large scale, these alignments can be used to locate rearrangements and duplications along with their frequency. On a small scale, like the alignments discussed before, WGAs can be used to analyze indels and single or multiple substitutions across the entire genome (Fig. 3.15). In addition to phylogenetic inference, WGAs are also used for genome annotation and function prediction of genes.

Evaluation: Just like any other tool, assessing the accuracy of whole genome alignments is important. However, a challenge is presented by the fact that on several occasions the true phylogeny of a set of genomes is not known. The following are four main approaches for a robust evaluation of WGAs: (1) simulation, (2) descriptive statistics, (3) comparison with other tools, and (4) confidence analysis of alignments to annotated regions. In the recent past, Alignathon, a comprehensive comparison of different WGA methods and their evaluation, was organized (Earl et al. 2014). OmicTools.com is a great compilation of several tools that perform WGA. A few notable examples from OMICTools include MUMmer4 (v4, released 2017) that rapidly aligns genomes, complete, draft form, or incomplete; FLAK, an

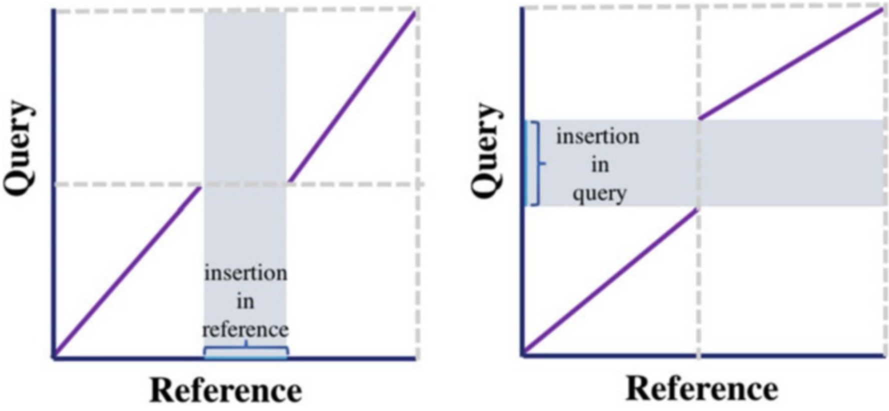


Fig. 3.15 Visualizing whole genome alignment: pairwise whole genome alignments can effectively reveal structural rearrangements, for example, in the form of insertions/deletions when a query genome is compared to a known reference genome

ultrafast fuzzy whole genome alignment; and visualization system that has a built-in native mechanism for approximate sequence matching (Healy 2016). Another family of tools for comparative genomics, WGA and visualization are VISTA (Frazer et al. 2004) and the newer version GenomeVISTA (Poliakov et al. 2014) – a comprehensive suite of programs and databases for analysis of multiple genomes.

As a concluding remark, it is imperative to say that sequence alignment reveals the amount and pattern of divergence among a set of sequences. Based on how conserved or variable a region is, between two or more sequences, much can be said about the importance of the region for functional and structural integrity. For instance, conserved regions could be responsible for binding site specificity in a protein-protein interaction, while a relatively variable region could be more promiscuous. Additionally, sequence alignment is the first step in any phylogenetic analysis. Regions of high similarity could be a consequence of evolutionary relationships, i.e., shared ancestry, and can be uncovered by using sequence alignment. The choice of alignment strategy to use must be directed by formulating relevant hypotheses to discover information guided by the problem at hand, including structure prediction, identification of a specific regulatory element, domain, motif, and family.

Things to Think About When Choosing Your Alignment Strategy
Is the length of the sequences the same?
Is it possible that only a small region in the sequences matches?
Are partial matches allowed?
Which substitution matrix is the best option for a given dataset?
Have there been indels from a common ancestor?

Acknowledgments This work is supported by a grant from the NIH Research Project Grant Program (2R01GM079656). The authors are grateful to Dr. David C. Marciano, Dr. Angela Wilkins, and Dr. Rhonald C. Lua for their helpful comments.

References

Alberts B, Johnson A, Lewis J et al (2002) Molecular biology of the cell, 4th edn. Garland Science, New York

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Baldi P, Chauvin Y, Hunkapiller T, McClure MA (1994) Hidden Markov models of biological primary sequence information. P Natl Acad Sci USA 91(3):1059–1063

Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. Bull Math Biol 51 (1):79–94

Correns C (1950) G. Mendel's law concerning the behavior of progeny of varietal hybrids. Genetics 35:33–41

Darwin C (1859.) On the origin of species by means of natural selection

Dayhoff M, Schwartz R (1978) A model of evolutionary change in proteins. Atlas Pro Seq Struct:345–352 10.1.1.145.4315

de Vries H (1900–1903) The mutation theory

Dewey CN (2012) Whole-genome alignment. In: Evolutionary genomics. Humana Press, Totowa, pp 237–257

Earl D, Nguyen N, Hickey G, Harris R (2014) Alignathon: a competitive assessment of whole genome alignment methods. bioRxiv:1–30. https://doi.org/10.1101/003285

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

Feng DF, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25:351–360

Forterre P (2015) The universal tree of life: an update. Front Microbiol 6:1–18. https://doi.org/10.3389/fmicb.2015.00717

Gibbs JA, McIntyre AG (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. Eur J Biochem 16:1–11. https://doi.org/10.1111/J.1432-1033.1970.Tb01046.X

Hagen JB (2000) The origins of bioinformatics. Nat Rev Genet 1:231–236. https://doi.org/10.1038/35042090

Healy J (2016) FLAK: ultra-fast fuzzy whole genome alignment. Advances in intelligent systems and computing, vol 477. Springer

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. P Natl Acad Sci USA 89:10915–10919

Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73:237–244. https://doi.org/10.1016/0378-1119(88)90330-7

Katoh K, Stanley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010

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 30:3059–3066. https://doi.org/10.1093/nar/gkf436

Koonin EV, Galperin MY (2003) Sequence – evolution – function: computational approaches in comparative genomics. Kluwer Academic, Boston

Krogh A, Mian IS, Haussler D (1994) A hidden Markov model that finds genes in *E. coli* DNA. Nucleic Acids Res 22(22):4768–4778

Li W, Cowley A, Uludag M et al (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res 43: W580–W580–4. https://doi.org/10.1093/nar/gkv279

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441

McWilliam H, Li W, Uludag M et al (2013) Analysis tool web services from the EMBL-EBI. Nucleic Acids Res 41:597–600. https://doi.org/10.1093/nar/gkt376

Mendel GJ (1865) Experiments on plant hybridization. Read at the meetings of the Brünn Natural History Society

Needleman SB, Wunsch CD (1970) General method applicable to search for similarities in amino acid sequence of 2 proteins. J Mol Biol 48:443

Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217. https://doi.org/10.1006/jmbi.2000.4042

Pearson WR (2014) An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinforma:1–9. https://doi.org/10.1002/0471250953.bi0301s42.An

Poliakov A, Foong J, Brudno M, Dubchak I (2014) GenomeVISTA-an integrated software package for whole-genome alignment and visualization. Bioinformatics 30:2654–2655. https://doi.org/10.1093/bioinformatics/btu355

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. P Natl Acad Sci USA 74:5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sievers F, Wilm A, Dineen D et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75

Smith T, Waterman M (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Thompson JD, Linard B, Lecompte O, Poch O (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One 6:e18093. https://doi.org/10.1371/journal.pone.0018093

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains archaea, Bacteria, and Eucarya. P Natl Acad Sci USA 87:4576–4579. https://doi.org/10.1073/pnas.87.12.4576 Webpage references: https://omictools.com/whole-genome-alignment-category

Zuckerland E, Pauling L (1965) History of evolutionary molecules as documents. J Theor Biol:357–366

Understanding Genomic Variations in the Context of Health and Disease: Annotation, Interpretation, and Challenges

4

Ankita Narang, Aniket Bhattacharya, Mitali Mukerji, and Debasis Dash

4.1 Introduction

The completion of the *Human Genome Project (HGP)* marked an important episode in the history of human genetics (Lander et al. 2001). It was the culmination of concerted efforts over a decade's time that empowered researchers and clinicians with the exact sequence of the 3 billion nucleotides that constitute the human genome. However, no two genomes are exactly alike; two unrelated individuals differ from each other at more than a million genomic loci (Consortium 2015b).

A. Narang (✉)
G.N. Ramachandran Knowledge Centre for Genome Informatics, Council of Scientific and Industrial Research – Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India

Epigenetics Lab, Dr. B.R. Ambedkar Center for Biomedical Research, University of Delhi (North Campus), Delhi, India

A. Bhattacharya
Genomics and Molecular Medicine, Council of Scientific and Industrial Research – Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India

Academy of Scientific and Innovative Research (AcSIR), Delhi, India

M. Mukerji
G.N. Ramachandran Knowledge Centre for Genome Informatics, Council of Scientific and Industrial Research – Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India

Genomics and Molecular Medicine, Council of Scientific and Industrial Research – Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India

Academy of Scientific and Innovative Research (AcSIR), Delhi, India

D. Dash
G.N. Ramachandran Knowledge Centre for Genome Informatics, Council of Scientific and Industrial Research – Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India

Academy of Scientific and Innovative Research (AcSIR), Delhi, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_4

Subsequently, there have been many international efforts with the aim to systematically catalog the variation data from different ethnic populations across the world. Such initiatives have now enabled population geneticists to understand variants in the context of population history, demography, environment, and disease. With the technological advancements propelled by next-generation sequencing (NGS), variants can now be identified at the single nucleotide level rather than estimating across few kilobases, using high-resolution genomic data. Today, while the number of repositories housing data from different variation studies is plenty (some of them well curated, regularly updated), there still remains a significant challenge in deciphering the functional roles of these variations. Assigning consequentiality to a variation is further complicated by its contextual nature; not all variations are meaningful in every context, and this might contribute, in part, to the paradox of missing heritability. In post-*ENCODE* (Encyclopedia of DNA Elements) era of genomics, when most of the genome is assigned functionality (Consortium 2012a), it becomes more important to annotate the non-coding variations (which have not been studied so extensively as their exonic counterparts, but hold immense importance with respect to the regulatory network dynamics in a highly context-specific fashion). This chapter highlights how to sieve "meaningful" variations from the "noisy" background and discusses how to annotate and ascribe functionality to these variants in the light of existing genomic knowledge.

4.2 Understanding the Variability in the Human Genome

There exists an extensive variability between the genomes of two different individuals. Variability in terms of the number of nucleotide changes ranges from single base pair (single nucleotide polymorphisms, SNPs) or a few base pairs (insertion-deletions, indels) to more complex events where an entire stretch of DNA gets altered due to deletion, duplication, or inversion. Figure 4.1 illustrates the major classes of genomic variations. SNPs and structural variations (SVs) occupy the two extreme ends of the *polymorphism* spectrum and differ substantially both in terms of their magnitude and effect (Zhang et al. 2009). The genomic distribution of SNPs is more uniform and widespread than that of the SVs. Recent estimates suggest that a typical human genome differs from the *reference genome* by 4.1–5 million base pairs, where more than 99.9% of the variants are SNPs and indels, while SVs are estimated to be in the range of 2100–2500 (Consortium 2015b). Despite being fewer in number than SNPs, SVs have a more pronounced effect because they encompass a larger fraction of the genome.

Variability among genomes can be defined at different hierarchies of organization (Table 4.1). Variants, with the exception of de novo and somatic variations, are genetically transmitted and, consequently, inherited from parents. De novo and somatic variations bring sporadic changes to an individual's genome; in a de novo event, all the cells of that individual contain the novel genetic change, while somatic variations lead to differences in the genomic content within a specific group of cells in that individual (Freed et al. 2014). In this chapter, we primarily discuss the methods to infer consequentiality to heritable human genetic variations, the

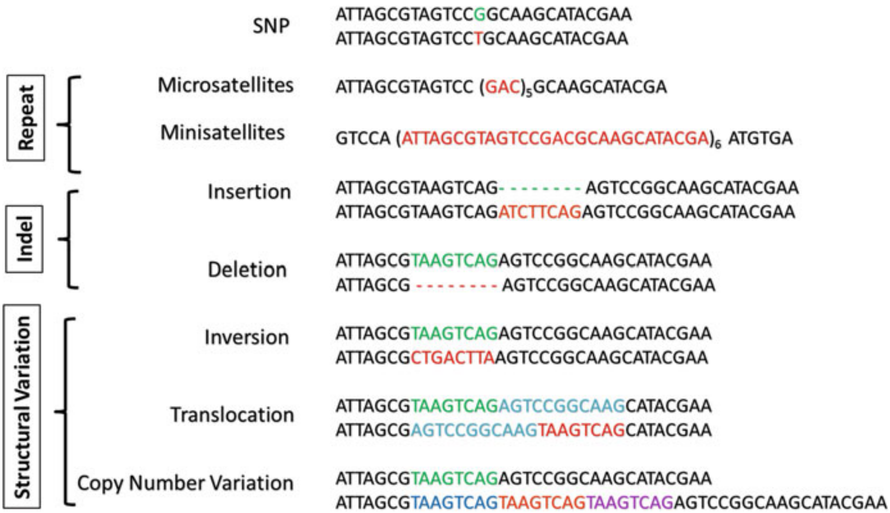

SNP
ATTAGCGTAGTCCGGCAAGCATACGAA
ATTAGCGTAGTCCTGCAAGCATACGAA
Repeat
Microsatellites
ATTAGCGTAGTCC (GAC)5GCAAGCATACGA
Minisatellites
GTCCA (ATTAGCGTAGTCCGACGCAAGCATACGA)6 ATGTGA
Indel
Insertion
ATTAGCGTAAGTCAG- - - - - - - - AGTCCGGCAAGCATACGAA
ATTAGCGTAAGTCAGATCTTCAGAGTCCGGCAAGCATACGAA
Deletion
ATTAGCGTAAGTCAGAGTCCGGCAAGCATACGAA
ATTAGCG - - - - - - - - AGTCCGGCAAGCATACGAA
Structural Variation
Inversion
ATTAGCGTAAGTCAGAGTCCGGCAAGCATACGAA
ATTAGCGCTGACTTAAGTCCGGCAAGCATACGAA
Translocation
ATTAGCGTAAGTCAGAGTCCGGCAAGCATACGAA
ATTAGCGAGTCCGGCAAGTAAGTCAGCATACGAA
Copy Number Variation
ATTAGCGTAAGTCAGAGTCCGGCAAGCATACGAA
ATTAGCGTAAGTCAGTAAGTCAGTAAGTCAGAGTCCGGCAAGCATACGAA

Fig. 4.1 Major classes of genomic variations (Adapted from Frazer et al. 2009)
SNPs, repeats, and structural variations are the three major classes of genomic variations. SNPs are point mutations, while CNVs are the more complex structural variations that encompass a larger fraction of the genome. Depending on their type, they have the ability to alter regulation and expression of genes, protein structure, and thus function

Table 4.1 Broad categorization of variations based on different characteristics

Broad categorization of genomic variants	
Type	Genetic, de novo, and somatic
Class	SNPs and structural variations
Frequency	Common, rare, and private
Consequence or functional impact	Coding (synonymous, non-synonymous) and non-coding

resources and bioinformatic tools that are currently available, and some of the pivotal studies dealing with the functionality of these variants. We mention the problems commonly encountered in these kinds of research endeavors and conclude with the way forward. The methods and resources related to the annotation of de novo and somatic variants have not been discussed in detail in this chapter.

4.2.1 Classification of Variations

Genomic variations can broadly be classified on the basis of their frequency and their function.

4.2.1.1 Frequency

On the basis of their frequency in the population(s), variations are majorly classified into two types – common and rare. Those having a frequency of ≥5% in population

(s) are referred to as common variations, while the frequency of rare variations ranges between 0.5% and 5% across different population(s). Private mutations are present only in the proband or close/immediate relatives (Cirulli and Goldstein 2010). Common variants are usually shared among different global populations, while rare variants are more restricted to one population or a continental group. Almost 86% of the rare variants are restricted to a certain continental group and, thus, have higher frequency in that population (although these are rare when their frequency is compared globally). The distribution of common and rare variations may have a potential effect on their phenotypic variability and disease risk (Consortium 2015b).

4.2.1.2 Function

On the basis of their functional consequences, variations are divided into different categories – *synonymous*, *non-synonymous*, loss-of-functional (LOF) variants (*frameshift* indels in coding regions, *stop gains*, and disruptions to essential splice sites), and regulatory regions (non-coding RNA, transcription factor binding sites, motifs). Most of the variations are either common or rare; however, estimates suggest an enrichment of rare and deleterious non-synonymous variations in the functional genomic regions. It was estimated that on an average, 76–190 rare non-synonymous changes and nearly 20 LOF variants contribute to disease risk and are important pathological candidates. Such variations, however, never reach to high frequencies in the population gene pool because of the constraint of a strong *purifying selection*. In the regulatory regions, there are around 18–69 variants that disrupt the transcription factor binding sites and may affect the regulation of the genome (Consortium 2012b). The trend of findings was consistent across various studies, though exact numbers may vary. Sequencing artifacts and annotation errors can bias the estimates. MacArthur et al. (2012) reported that healthy genome(s) have ~100 LOF variations that are majorly maintained in the heterozygous state and may cause recessive Mendelian disorders in the homozygous state. Thus presence of LOF variants confers differential disease risk among individuals. Distribution of common LOF variants is biased toward the nonessential genes with a potential to affect the phenotype (MacArthur et al. 2012).

De novo and somatic events are less studied in comparison to the genetic variants. Phenotypic consequence of these variants is culmination of effect from multiple factors in both healthy and diseased individuals, e.g., gene expression and epigenetic changes owing to such variants in combination with environmental factors (De 2011). Trio-based whole genome sequencing suggests presence of approximately 74 de novo single nucleotide variants in the genome of an individual, while the frequency of other complex de novo events is not exactly known (Veltman and Brunner 2012). De novo variants are known to be the primary cause of sporadic genetic disorders. Moreover, the number of these events is observed to be higher in the affected individuals (since these variants are rare and can be potentially deleterious to the gene functionality) (Veltman and Brunner 2012). Similarly, somatic variants have implications in the process of aging and their role in different forms of cancer has been well established (De 2011).

4.3 Factors Affecting the Frequency of Variation (s) in the Population

According to the principle of *Hardy-Weinberg equilibrium (HWE)*, frequency of variation in a population will remain unchanged/constant from generation to generation in the absence of any external evolutionary forces (Hardy 1908). However, in the real-world scenario, conditions defined for HWE are rarely met. It is more of a hypothetical concept because dynamic demographic events and changes in environmental factors can alter the frequency of an allelic variant in a population which leads to deviation from HWE. *Genetic drift*, *gene flow*, and selection are such mechanisms that can cause deviation in the frequency of an allele (Andrews 2010). Genetic drift is a random change in *allelic frequency* due to chance events. For example, *population bottleneck* shrinks the population size and decreases the overall genomic *heterozygosity* by complete loss of some alleles from the population (Griffiths et al. 1999). Gene flow or migration among different populations can also lead to change in the allele frequency by introduction of new alleles in a population (Andrews 2010; Lenormand 2002). These events are non directional, ultimately culminating in random consequences which needs to be examined on a case-to-case basis (Griffiths et al. 1999; Hartl and Clark 1997).

However, changes in the frequency of variant(s) in response to environmental or geographical factors that triggers the adaptability of an organism is termed as "selection" (Andrews 2010). In contrast to drift, selection doesn't affect the genome randomly and has locus-specific effects (Biswas and Akey 2006). Differences between genetic drift and selection are briefly described in Table 4.2.

In 1968, Motoo Kimura proposed the theory of neutral evolution, which provides a completely different view about selection (He 1994; Kimura 1984). According to this theory, all genomic loci evolve neutrally, and changes in the frequency of an allele can be considered to be a stochastic event which doesn't necessarily impact adaptation or fitness. With the availability of genomic data and development of better statistical methods, there are numerous convincing evidences of selection found in literature. Genomic selection can operate in three ways – "*Positive selection*" leads to increase in the frequency or fixation of an allele that favors the fitness of an individual. One of the most important features of positive selection is "genomic hitchhiking" which tends to increase the frequency of nearby linked loci also and sweeps the frequency of the entire genomic region to fixation or near fixation. On the contrary, "negative selection" or "*purifying selection*" purges out deleterious alleles (that decrease the fitness or potential for adaptation) from the population, while "balancing selection," a more general form of selection, favors the presence of heterozygotes over homozygotes (Biswas and Akey 2006; Nielsen 2005).

Table 4.2 Features of genetic drift and selection

Parameter	Genetic drift	Selection
1. Occurrence	Random	Non random
2. Mode	Non directional	Directional
3. Effects	Genome-wide	Locus-specific

4.4 High-Throughput Methods to Identify Genetic Variations

Gel-based methods were used for the identification of first-generation genetic markers such as restriction fragment length polymorphisms (*RFLP*s), *minisatellites*, random amplified polymorphic DNAs (RAPDs), and amplified fragment length polymorphism (*AFLP*) (Griffiths et al. 1999; Pattemore 2011). These methods had a strong impact on the emerging field of molecular biology, but their applicability was limited because of inherent drawbacks like high cost, labor intensiveness, greater time consumption, and inability to yield high-throughput results (Pattemore 2011). These methods become obsolete, and the present scenario is dominated by high-throughput technologies like arrays and sequencing for detection of SNPs and structural variations.

In the present scenario, two major approaches to identify genetic variations are genotyping arrays and NGS platforms.

4.4.1 Genotyping Arrays

Genotyping arrays were originally developed to detect SNPs but their utility was later extended to detect copy number variations (CNVs) also. Consequently, these arrays are now routinely being used to genotype both common SNPs and CNVs. Fragmented sample DNA is hybridized to unique set of thousands of *probes*. Each probe harbors SNP sites and there are replicates for both alleles of a SNP for a particular probe (probe set). The intensity of probe hybridization with its complementary target DNA determines the sample genotype for a particular SNP (Syvänen 2001). Computational algorithms are required to convert raw intensity values into genotype data (LaFramboise 2009). Affymetrix and Illumina are the two major genotyping platforms available till date. These differ from each other in terms of their chemistry but the principle of decoding genotype from intensity value remains the same. Initial versions of both platforms could detect fewer SNPs, but now latest versions (Affymetrix 6.0 and Illumina's HumanHap1M) of both platforms probe against approximately 1 million SNPs (LaFramboise 2009). In the initial versions of genotyping arrays, there were no probes for CNVs. Therefore, deviations from expected hybridization intensity values were used to find CNV regions from SNP arrays (LaFramboise 2009). These versions of arrays had some limitations due to scarcity of probes near duplicated and repeat-rich regions of genome that are known to harbor CNVs. To improve CNV detection, latest versions of Affymetrix and Illumina have now included probes for copy number regions which are known to be variable among populations. Customized genotyping arrays allow researchers to focus on specific regions, pathways, or genes of their interest. *Exome* arrays for rare variants, metabochip for metabolic and cardiovascular disorders, and drug-metabolizing enzymes and transporters (DMET) microarray for *pharmacogenomics* are few of the celebrated examples for the targeted research questions.

4.4.2 Next-Generation Sequencing-Based Methods

All probe-based methods require a priori knowledge of the genome sequence of an organism for designing probes to capture variations. But the applicability of NGS-based approaches is not limited to known organisms; whole genome sequencing of both novel and known organisms provides unbiased approach to detect all variations that exist in an organism (Ekblom and Galindo 2011; Mardis 2013). Thus sequencing-based methods are free from the bias of genotyping chips as both common and rare/private variations are captured (Cirulli and Goldstein 2010; Frazer et al. 2009). These methods allow identification of all forms of complex variations like inversions, translocations, and breakpoints of copy number variations at single nucleotide resolution which traditional genotyping-based methods cannot confidently resolve (Alkan et al. 2011; Medvedev et al. 2009). The major platforms that are used for sequencing are Roche's 454, Illumina, ABI's SOLiD, and recently introduced Ion Torrent, PacBio, and Nanopore. These platforms use different chemistries for sequencing. Efficiency of these sequencing methods can be compared on the basis of a number of parameters, viz., read length and number, base calling accuracy, and efficiency in capturing heterochromatin regions. Each has its own pros and cons (Mardis 2013; Shendure and Ji 2008). Targeted sequencing, another variant of NGS, allows us to sequence specific areas of interest in the genome. Exome sequencing is one of the most celebrated examples of targeted capture of exonic regions of the genome. It requires an additional step of exome enrichment, pulling out known exonic regions/fragments selectively from the fragmented DNA (Clark et al. 2011). Exome sequencing has its major strength in the identification of causal variation/gene in single-gene disorders (Gilissen et al. 2011; Ng et al. 2010). Variation calling from sequencing-based methods suffers from instrument-generated errors, so the major challenge is to accurately distinguish a real variant from a sequencing artifact (Nielsen et al. 2011).

From the above discussion, it is clear that inherent drawbacks and biases are associated with technologies that are used to identify variations. Irrespective of this, many discoveries have been made in the field of genomics by using genome-wide variation data from both genotyping arrays and sequencing platforms. Mapping of known and novel genetic variations helps in the identification of genes implicated in diseases. Genome-wide variation data have also contributed a lot in understanding evolution and population history. Initial studies based on a single or a few marker datasets didn't have enough resolution to provide complete understanding of demographic events and population history.

Advancements in sequencing technologies also allow us to study ancient DNA. For example, genome sequencing of extinct hominins – Neanderthals and Denisovans – provides clues about population divergence events (Stoneking and Krause 2011).

4.5 Catalogs of Basal Variation Data from Global Populations

One of the major goals post HGP is to capture the genetic diversity of different global populations so as to decipher the underlying genetic basis of their disease susceptibility and phenotypic variability. There are numerous concerted efforts worldwide to make these genetic resources as informative as possible.

dbSNP (Sherry et al. 1999, 2001; Smigielski et al. 2000) was one of the major projects initiated by the National Center for Biotechnology Information (NCBI) and National Human Genome Research Institute (NHGRI) in 1999 to catalog variations submitted by laboratories, industries, genome sequencing centers, etc. It is a central repository which gathers data from all major projects and databases and is regularly updated. It also provides a standard nomenclature to variations that allows users to navigate variation through specific identifiers. These unique identifiers are known as RefSNP IDs or rsIDs. With the completion of HGP in 2003, the HapMap (Haplotype Map) project was initiated (Gibbs et al. 2003; Thorisson et al. 2005). Initial phase of HapMap genotyped approximately 1.3 million common SNPs in 269 individuals from four major world populations – (i) Yoruba in Ibadan, Nigeria (abbreviation YRI); (ii) Utah, USA, from the *Centre d'Etude du Polymorphisme Humain* collection (abbreviation CEU); (iii) Han Chinese in Beijing, China (abbreviation CHB), and (iv) Japanese in Tokyo, Japan (abbreviation JPT). In 2007 and 2010, HapMap released data for the next two phases. The major aim of this project is to build an informative genetic resource of common variations among individuals or populations. There were other initiatives like the Human Genome Diversity Project (HGDP) that expanded the horizon of human genetic variation by analyzing variations in 52 worldwide populations (Li et al. 2008). Apart from SNPs, catalogs for more complex variations like CNVs were also developed. The Database of Genomic Variants houses data for structural variations in healthy individuals from published studies (MacDonald et al. 2014). But the genetic diversity of Indian populations was not captured by any of these major global initiatives. The Indian Genome Variation Consortium (IGVC) project (Consortium 2005, 2008; Narang et al. 2010), an initiative of the Council of Scientific and Industrial Research (CSIR), had cataloged variations across diverse ethnic groups in India. Initial phases of this project had markers from candidate gene-based studies. In the subsequent phases, genome-wide markers for 26 Indian populations were genotyped. These populations were a subset of previous study. In 2009, the Human Genome Organization (HUGO) Pan-Asian SNP consortium (Abdulla et al. 2009) mapped the diversity of Asian populations which are underrepresented in other genomic surveys. PanSNPdb (Ngamphiw et al. 2011), an initiative of HUGO Pan-Asian SNP consortium, houses both SNP and CNV data from 71 Southeast Asian populations. Another breakthrough in genomics research was marked by the 1000 Genomes Project. In 2010, data from pilot phase of this project was released, and variations (SNPs, CNVs, indels) having frequency of 1% or higher in populations were captured which represents 95% of the human genetic diversity (Consortium 2010). This phase

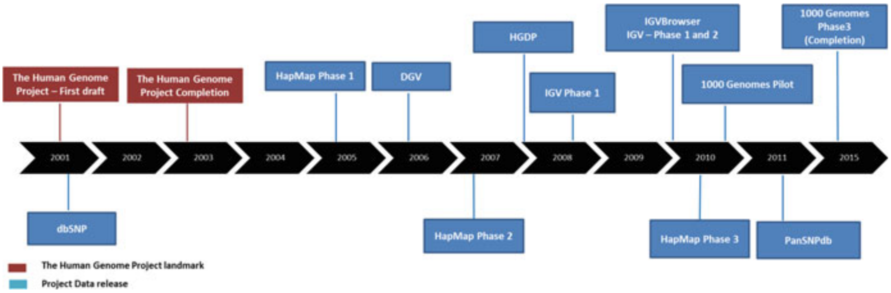

The Human Genome Project – First draft
The Human Genome Project Completion
HapMap Phase 1
DGV
HGDP
IGV Phase 1
IGVBrowser IGV – Phase 1 and 2
1000 Genomes Pilot
1000 Genomes Phase3 (Completion)
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2015
dbSNP
HapMap Phase 2
HapMap Phase 3
PanSNPdb
The Human Genome Project landmark
Project Data release

Fig. 4.2 Timeline for the development of major human variation catalogs
The figure depicts the chronological development of human variation catalogs from 2001 through 2012. Before HGP, dbSNP was the only major catalog for human variations. But, after the completion of the HGP in 2003, efforts toward the development of such catalogs were accelerated. HapMap project was initiated after the completion of the Human Genome Project with four worldwide populations. In spite of the worldwide coverage, Indian populations were missed by the HapMap project. A major attempt toward cataloging Indian variation data was carried out by the IGVC after its announcement in 2005 using 55 ethnic Indian populations in the first phase released in 2008. Efforts for cataloging the structural variation data were also initiated and thus DGV was developed in 2006. As a part of HUGO Pan-Asian SNP consortium, PanSNPdb was developed in 2011 that catalogs variation data from 71 Southeast Asian populations. In 2008, 1000 Genomes Project made the first announcement to catalog variations from sequencing data. The pilot project was released in 2010, followed by its completion in 2015

included sequencing of 179 complete genomes from 4 different ancestries and 679 exomes of 697 individuals from 7 different world populations. The ultimate aim of the project is to sequence ~2500 genomes from 26 global populations to identify 99% of the variants (with frequency > 1%) across different ancestries. This aim was finally accomplished in 2015, after the release of the intermediate phase 1 in 2012 (Consortium 2012b). Cumulatively, there are nearly 84.7 million SNPs, 3.6 million indel polymorphisms, and 60,000 structural variants reported till September 2015 by the 1000 Genomes Project (Consortium 2015b). Timeline of development of major human variation catalogs is shown in Fig. 4.2. Apart from providing an extensive catalog of variants, genome analysis tools were also developed and provided as open-source software for the benefit of the research community.

Other than these global initiatives, population-specific genome sequencing projects (like UK10K project) demonstrated the need of population-/continent-specific high-quality reference panels for detection and imputation of low-frequency or rare variants, which is not feasible with the representation in global projects, i.e., small sample size would not allow detection of these rare variants and, consequently, would be missed (Huang et al. 2015). Many such initiatives are in progress with a promise to aid development of more accurate reference panels and individualized diagnostics.

4.6 Applications of Cataloging Population Genomic Data

Variations are static but their dynamic behavior in response to environmental cues contributes to an immense genomic and phenomic diversity in populations. Examples that elucidate the potential applications of population-level variation data are discussed below.

4.6.1 Genomic Signatures of Selection and Adaptation

Migration and the subsequent exposure of modern humans to diverse niches forced them to adapt selectively in response to different climatic zones, diet, pathogens, and other factors (Coop et al. 2009). The allele that increased the fitness of a population had selectively increased in frequency and sometimes had even reached fixation for better adaptability (to the physical parameter which drove selection). There are many examples that have shown association between selection and adaptation in response to diverse conditions. Few case studies are discussed below.

4.6.1.1 Selection in Response to Geographical and Climatic Changes

One of the classical examples for understanding the phenotypic variability and adaptation in response to climatic changes is the spectrum of skin color in diverse human populations. The quantity and the distribution of the pigment melanin in melanosomes are the major determinants of skin color in humans (Quillen and Shriver 2011). The gradient of human skin tone across diverse world populations is well correlated with latitudes (Hancock et al. 2008, 2011; Jablonski and Chaplin 2000, 2010). This observation at the phenotypic level is also in line with an important biological balance which needs to be maintained between Vitamin D synthesis and folate metabolism. People living in the equatorial regions have darker skin, while people residing at higher latitudes have a comparatively lighter skin tone. The melanin content of the skin is directly proportional to the amount of UV radiation absorbed by it. The fine balance between Vitamin D synthesis and folate degradation thus depends on exposure to sunlight (the environmental trigger driving selection) (Jablonski and Chaplin 2000, 2010). Darker skin prevents folate degradation from direct UV radiation in the tropics. Lighter skin pigmentation is an advantage for people at high latitudes as it helps in absorption of UV radiation and thus promotes vitamin D synthesis. This clearly indicates that the difference in skin melanin levels across latitudes acts as an adaptive trait at different geographical clines. Variations linked to this adaptive trait have different frequencies across diverse populations. One of the landmark discoveries in this context was elucidating the role of *SLC24A5* gene in golden phenotype (lighter skin color) of zebra fish mutants (Lamason et al. 2005). This gene encodes NCKX5 protein, a putative cation exchanger. There exists a human ortholog for this gene and to study its role in human pigmentation; researchers have looked for variations in it. A coding variation rs1426654 (A/G) in the third exon of this gene (which leads to substitution of alanine by threonine at position 111 in the encoded protein) was the only coding

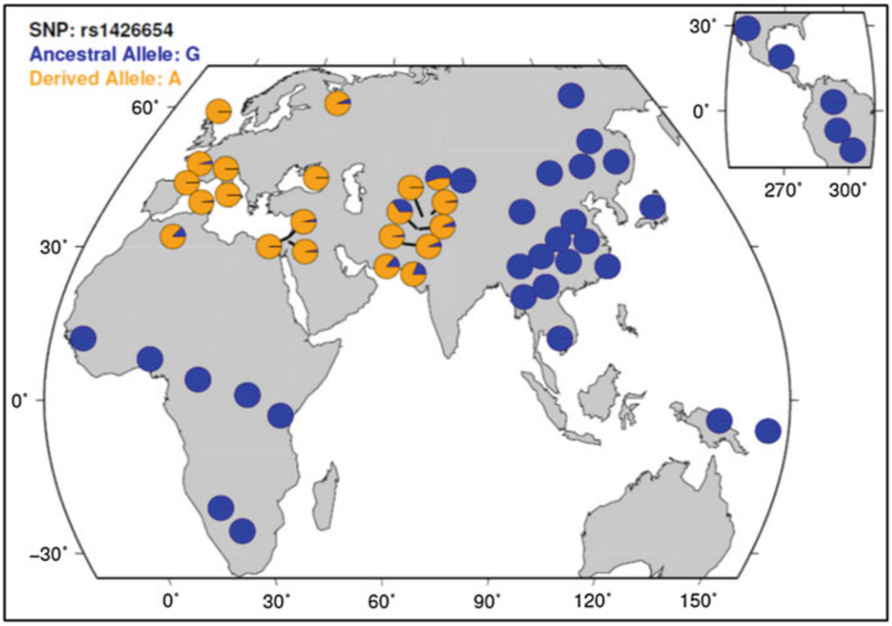


Fig. 4.3 Geographical distribution of the allele frequencies for rs1426654 across the world. (Source: HGDP Selection Browser)
rs1426654 is highly differentiated among Europeans and West Africans. Ancestral allele "G" and derived allele "A" are almost fixed in the Europeans and West Africans, respectively. This polymorphism in the *SLC24A5* gene has been found to play a major role in the skin pigmentation

SNP found in the HapMap project at that time. Interestingly, it is known to be one of the highly differentiated SNPs between Europeans and West Africans. Ancestral allele and derived allele of this SNP are almost fixed in Western Africans and Europeans, respectively (Fig. 4.3). Significant association between melanin index and genotypes of *SLC24A5* gene in admixed populations also explained the pigmentation differences. However, it is important to note that the ancestral allele of this SNP is almost fixed in East Asians as well, but they have a pale skin tone which suggests that genes involved in skin pigmentation have evolved independently and are not shared across global populations. Many other candidate genes related to skin pigmentation including *SLC45A2*, *OCA2*, *TYR*, *KITL*, and *MITF* are also under selection in different populations around the globe (Sturm 2009).

4.6.1.2 Selection Against Pathogen Load

Malaria resistance is one of the well-known examples of human adaptation to pathogen load in malaria-endemic regions like parts of Africa. Such endemic geographical zones have a selective advantage; populations residing there experience a strong selection pressure and harbor protective variations which reduce the risk of infection. Genes that are expressed in red blood cells, such as human leukocyte antigen (*HLA*), glucose-6-phosphate dehydrogenase (*G6PD*), and Duffy factor (*FY*), accumulate variations that provide advantage against malaria (Tishkoff et al. 2001). Another example is the variation in apolipoprotein L1 (*APOL1*) gene that provides resistance against human African trypanosomiasis (HAT) or sleeping sickness but with the susceptibility to chronic kidney disease (CKD) in African-Americans (Ko et al. 2013). There are numerous such examples of balancing selection where resistance against one pathogen is counterbalanced by susceptibility to some other disease.

4.6.1.3 Adaptation to Dietary Shifts

Humans have adapted to different diets in response to agriculture and this had resulted in lifestyle shifts during the process of evolution (Laland et al. 2010). Several studies have shown selection for genomic variations that favor adaptation with changes in diet patterns across different populations. Apart from single nucleotide changes, Perry et al. (2007) demonstrated copy number variation in amylase gene *AMY1* (which varies across different populations in response to their starch intake). The concentration of amylase present in the saliva is directly proportional to the copy number of *AMY1*. Populations that have a high starch intake also have more copies of *AMY1* than populations that are accustomed to a low-starch diet. For example, Japanese, European-Americans, and Hadza hunter-gatherers of Africa consume high-starch diets and also have higher copies of *AMY1* gene in comparison to populations like Biaka, Mbuti, Datog pastoralists, and Yakuts who consume less starch. This positive correlation between number of copies of *AMY1* gene and starch intake is independent of geography. Lactose tolerance in European populations is another celebrated example of dietary adaptation.

4.6.2 Pharmacogenomics

Genetic variations in drug-metabolizing enzymes, receptors, and transporters are known to be responsible for differences in drug responses among individuals and hence, populations (Zhou et al. 2008). Variations in genes related to drug response can be important determinants of drug efficacy and toxicity. One of the major aims of personalized genomics is to prescribe medicine or specify dosage on the basis of an individual's genetic architecture (Ginsburg and Willard 2009; Zhou et al. 2008). The concept of personalized medicine has important implications in translational and therapeutic research.

4.6.3 Designing Disease Association Studies

Genome-wide association studies (GWAS) are performed to decipher the genetic basis of common diseases including diabetes, cardiovascular diseases, autoimmune diseases, and psychiatric disorders (Bush and Moore 2012; McCarthy et al. 2008), where common variants influence disease predisposition. Since the variation is common in the population, its effect size is smaller than that of rare variants. In

GWAS, disease risk is explained by the cumulative effect of multiple common alleles of small effect size (Bush and Moore 2012; Manolio et al. 2009). GWAS involves scanning of common variations to identify the loci where frequency of an allele (or genotype) is higher in cases than in controls from the same population background; such loci stand a greater chance of being associated with disease(s). *Linkage disequilibrium (LD)* is another important factor or parameter that determines the number of markers required for GWAS. In general, populations that have undergone higher number of *recombination* events have decreased LD, while a population with less number of recombination events has extended LD. Association between alleles in a region decreases with increasing age of the population, and thus variants that act as proxies or tags for each other will no longer be correlated. Hence, populations with older histories like African populations require genotyping of a large number of markers than populations with comparatively recent histories like America, Europe, etc. (Ardlie et al. 2002; Slatkin 2008). Homogenous population background in such studies is a prerequisite to avoid spurious results (which is a reflection of the genetic variability among cases and controls). Thus, the representation of different ethnic populations or groups in genomic databases may aid in the selection of appropriate controls.

4.6.4 Reference Panels for Genome Imputation

Before the advent of high-throughput sequencing technology, missing common variants (or genotypes) in GWAS meta-analysis were imputed using HapMap reference panels, which has cataloged variation data from different ethnic backgrounds (genotyped on chip). Many novel associations using GWAS meta-analyses of many complex disorders like diabetes, vitiligo, cardiovascular, and neurological disorders were reported, while many known associations could not be replicated. Accuracy of such references for imputing common variants was quite good. However, the imputation of low-frequency and rare variants was a limitation at that time. With the easy accessibility to genome and exome sequencing data, imputation of low-frequency/rare variants has become quiet feasible by sequencing sufficient number of samples required for rare variants' imputation. Another cost-effective alternative is the use of customized exome chips for rare variants to impute genotypes (Auer et al. 2012; Consortium 2015b; Huang et al. 2015; Spencer et al. 2009; Zheng et al. 2015a). Such approaches allow the discovery of association of novel and rare/low-frequency variants with many complex traits or phenotypes. A study reported the association of rare coding variants with blood cell traits, where exome sequencing data from 761 African-American individuals was used to impute novel genotypes in 13,000 individuals with the same ancestry (Auer et al. 2012). UK10K project evaluated the association of rare/low-frequency variants with 31 core traits common in 2 cohorts of European ancestry. They reported two novel associations of non-coding rare and low-frequency variants with triglycerides, adiponectin, and low-density lipoprotein cholesterol levels (Consortium 2015c). In another study from the same project, they reported novel association of non-coding genetic variants with bone mineral density (BMD) and fracture (Zheng et al. 2015b).

4.7 In Silico Strategies for Prioritizing Genomic Variants Using Genomics and Annotation Databases: An Approach to Reduce the Search Space

One of the daunting challenges in population genetics today is to make sense of the huge repertoire of genetic variations and thousands of associations from GWAS and genome-wide scans of selection. This issue can be addressed by prioritization of functionally important variation(s) that can affect phenotype and disease. Annotation resources and prediction servers gather functional information from computational predictions and/or experimental evidences to predict the effect of variation on phenotype. Assessment of the functional potential of a variant is dependent on its type and nature. Functional interpretation of the coding variations is quiet straightforward compared to that of the non-coding variations because of the assumption that the former might have a direct effect on protein structure and function, while the latter can alter regulation in the genome in a number of ways (not all of which are properly understood).

Recently, exome and genome sequencing have been used as powerful tools to identify rare protein-coding variation(s) in extreme phenotypes like monogenic or Mendelian disorders (Ng et al. 2010). Many prioritization strategies and pipelines have been developed to narrow down the search space by identifying few putative candidates out of several thousands (Biesecker 2010; Stitziel et al. 2011). The variants in Mendelian or rare diseases are first filtered on the basis of their frequency reported in the population genomic databases. Presence of ancestry-specific variants (markers) becomes more valuable when there is a lack of background control data for the population in which the study was conducted. For filtering out the exonic variants, large exome repositories were developed wherein thousands of individuals across different ancestries were sequenced. Exome Aggregation Consortium (ExAC) and the National Heart, Lung, and Blood Institute (NHLBI) Grand Opportunity Exome Sequencing Project (ESP) are the landmark efforts in this direction. Another filtering criterion used for prioritization of coding variants includes evolutionary approaches that predict tolerance to amino-acid substitutions (substitutions at the evolutionarily conserved sites are more damaging than the non-conserved ones) and the choice of disease model in pedigree-based case studies. Foo et al. (2012) described a simplistic approach for the prioritization of coding variants in Mendelian or rare diseases with large effects. In cases of consanguinity, homozygosity mapping can be used to find stretches of DNA in affected siblings, and these regions can further be used for targeted resequencing which minimizes the cost of sequencing multiple genes (Smith et al. 2011).

The coding genome (~1.5%) is very small in comparison to the non-coding portion that encompasses a large number of variations which, in turn, modulate the genome dynamics by regulating gene expression (Brown 2002). Non-coding variations can exert their effect either by changing the binding potential of transcription factors or other regulatory elements (Boyle et al. 2012). It further helps in elucidating the molecular mechanism of variation that may explain phenotype. Most of the variants identified in genome scans of selection or GWAS are intronic, and in some cases multiple variants in the same gene have been reported to be associated with phenotype or disease. Prioritization of coding variant is quiet straightforward in case of highly penetrant or Mendelian diseases, but this is not the case with non-coding variants. Narrowing down the list of associations into one or few regulatory causal candidates requires integration of variation data with multiple and diverse functional information like conservation patterns in the genome, environmental variables, epigenomics, expression, and regulatory signature. In another case, where variants from different genes have been associated with the same phenotype, differences in gene expression patterns relevant to phenotype/disease can be measured in different tissues.

Many studies have integrated population genomics along with classical molecular biology experiments to elucidate the functionality of non-coding variants. For example, adaptive variation in serum- and glucocorticoid-regulated kinase1 (*SGK1*) gene regulates its expression in stress responses (Luca et al. 2009), function of adaptive variation in ectodysplasin A receptor (*EDAR*) is associated with hair thickness in the Han Chinese population and is identified as a biological target of adaptive variation (Kamberov et al. 2013), cis-acting sequence variation in the promoter of *VNN1* gene is found to be associated with cardiovascular disease (Kaskow et al. 2013), etc.

In the current-day scenario, the availability of high-throughput functional data from collaborative efforts like ENCODE (Consortium 2012a), Fantom5 (Consortium 2014), Roadmap Epigenomics (Bernstein et al. 2010), and Genotype-Tissue Expression (GTEx; Consortium 2015a) allows integration of multiple omics datasets to annotate and understand the functionality of candidate genomic regions or variations. Target variants may have different impacts on the regulation and/or expression based on their genomic position. Regulatory potential of a variant in the intra-/intergenic region and untranslated region (UTR) can be assessed by mapping DNaseI hypersensitivity (HS) data (a marker of active regulatory elements) along with histone modifications, transcription factor, and RNA Pol II occupancy datasets. The presence of DNaseI HS site is a proxy for open DNA and may suggest the presence of enhancer or transcription factor along with the assessment of signatures of associated chromatin modifications (Kellis et al. 2014). Therefore, it might be possible that variant allele disturbs or favors that potential binding. Further, the expression of this variant in different cell types and/or tissues can be evaluated which helps in the selection of appropriate or impacted cell lines for the experimental validation (Fig. 4.4). For example, rs12203592 in *IRF4* gene is one of the GWAS candidates that has been shown to be associated with pigmentation traits. Overlapping this intronic region with ENCODE data suggest the presence of DNaseI HS as well as its expression in melanocyte-derived cell lines. This prediction aided in conducting the directed experiments for validation (Praetorius et al. 2013). Variations in the 3’UTR of genes which overlap with miRNA target sites (especially in the seed region) may have important downstream ramifications. If both the

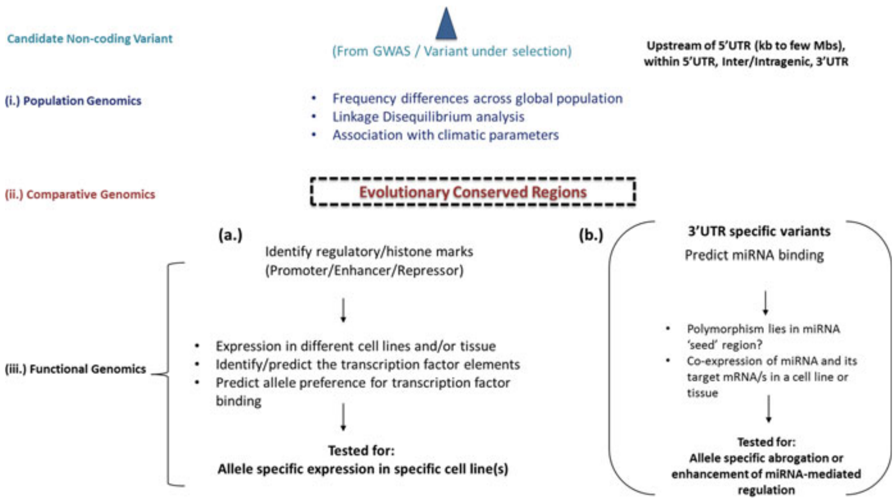

Candidate Non-coding Variant
(From GWAS / Variant under selection)
Upstream of 5'UTR (kb to few Mbs), within 5'UTR, Inter/Intragenic, 3'UTR
(i.) Population Genomics
Frequency differences across global population
Linkage Disequilibrium analysis
Association with climatic parameters
(ii.) Comparative Genomics
Evolutionary Conserved Regions
(iii.) Functional Genomics
(a.)
Identify regulatory/histone marks (Promoter/Enhancer/Repressor)
Expression in different cell lines and/or tissue
Identify/predict the transcription factor elements
Predict allele preference for transcription factor binding
Tested for:
Allele specific expression in specific cell line(s)
(b.)
3'UTR specific variants
Predict miRNA binding
Polymorphism lies in miRNA 'seed' region?
Co-expression of miRNA and its target mRNA/s in a cell line or tissue
Tested for:
Allele specific abrogation or enhancement of miRNA-mediated regulation

Fig. 4.4 Brief outline for annotating the non-coding variants
In silico integrated omics approach is used to annotate the candidate non-coding variations either from GWAS or genome-wide selection studies. (i) Using population genomics, estimates of global allele frequency differences and linkage disequilibrium (LD) were retrieved or computed. In addition to this, impact of environmental selection on allele frequency can be assessed – association of the variant with different climatic parameters can be inferred using resources like dbCLINE. (ii) Comparative genomics helps to delineate the conserved genomic regions across species. Evolutionary conservation is a parameter which suggests regions of importance; however, it is not much informative about functionality of genome. (iii) Mapping/overlaying candidate genomic regions over the functional elements using functional genomics resources like ENCODE, GTEx, etc. provides insights about the regulation and/or expression. Variants in 3ÚTR can further be scanned to understand the impact of different allelic variants on the formation of miRNA-mRNA duplex

miRNA and its target transcript are co-expressed in a particular tissue, the miRNA-mRNA duplex formation is often favored for one allele over the other. 3'UTR SNPs overlapping with miRNA binding sites can both create and abolish miRNA-mediated gene regulation. TCF21, a transcription factor pivotal to vascular development, is differentially regulated by miR-224 depending on whether an individual has C or G at rs12190287. The risk allele C facilitates RNA structure formation which can allow easy access to the miRNA to ensure efficient binding and subsequent degradation of the transcript in coronary heart disease (Miller et al. 2014). These variations can add and alter players in miRNA networks. For example, in case of rs2168518 (G > A), a seed region variant of miR-4513, regulation of one of its target genes *GOSR2* is completely altered (Ghanbari et al. 2014).

Another aspect is association of genetic variants with the methylation of CpG islands in the promoter regions. Variation in the methylation levels at CpG blocks contributes to phenotypic variability, and it has also been linked to genetic variants, which are known as methylation quantitative trait loci (meQTLs). Correlation between GWAS candidate SNPs and methylation has been tested, and many significant associations were reported (Bjornsson et al. 2004; Hidalgo et al. 2014; Rushton et al. 2015).

Schaub et al. (2012) have assessed the functionality of variation using a scoring scheme (Regulomedb scores: http://regulomedb.org/help). Score assignment is based on known and predicted functional evidences from multiple resources including ENCODE, Roadmap Epigenome Consortium, GEO (Gene Expression Omnibus), and published literature. Utility of this scoring scheme has been demonstrated by assigning functional role to GWAS SNPs associated with different traits/ phenotypes (Schaub et al. 2012). Enrichment of GWAS SNPs (lead SNPs) or SNPs that are in LD with GWAS SNPs (functional SNPs) in experimentally identified functional regions supported their likely role in regulation.

GTEx is an important addition to the armory of resources for prioritization of functional variants. It houses data for 175 healthy individuals who have been densely genotyped and their RNA-Seq data is also available across 43 tissue types (Consortium 2015a). Taking gene expression as a quantitative trait, the consortium has cataloged a list of SNPs whose allelic states are linked with the expression of certain genes in particular tissues (along with its effect size). These SNPs are known as expression quantitative trait loci (eQTLs) and serve as an anchor point to connect the static information of genetic variation with dynamic gene expression. GWAS SNPs can be queried using these data to ascertain their associated functional consequences in different tissue types. Regulatory eQTLs shared among tissues can also be detected. A celebrated example of the use of such paired variation-expression data is the discovery of sun-exposure eQTLs in an European population (Kita and Fraser 2016).

4.7.1 Resources for Functional Annotation of Variation Data

Availability of millions of variations from sequencing studies has further increased the complexity of annotation and analysis as the size of the search space has now increased by several orders of magnitude and this has generated the need for computational resources and automated pipelines that are fast and can bypass manual errors and increase the sensitivity. There were numerous (and continuous) efforts by different groups to make databases, web servers, and pipelines for annotating variations that can identify the underlying genotype to phenotype correlations. Table 4.3 provides a list of all major databases, web tools, servers, and pipelines that are categorized into different sections on the basis of the type of

Table 4.3 Catalogs and pipelines for functional annotation of variation data

Type	Name	URL
Annotation resources – databases and web servers		
Disease association databases	GWAS Catalog	http://www.genome.gov/gwastudies/
	GWAS Central	http://www.gwascentral.org/
	GAD (Genetic Association Database)	http://geneticassociationdb.nih.gov/
	OMIM (Online Mendelian Inheritance in Man)	http://www.omim.org/
Genomic selection	1000 Genomes Selection Browser	http://hsb.upf.edu/
	dbPSHP	http://jjwanglab.org/dbpshp
Environment-related variables	dbCLINE	http://genapps2.uchicago.edu:8081/dbcline/main.jsp
Pharmacogenomics	PharmGKB	https://www.pharmgkb.org/
CNV annotation	DECIPHER	http://decipher.sanger.ac.uk/
	CNVD	http://202.97.205.78/CNVD/
eQTLs, gene expression (cell type and tissue specific), epigenomics and regulation	Genotype-Tissue Expression (GTEx)	http://www.gtexportal.org/home/
	eQTL browser	http://eqtl.uchicago.edu/Home.html
	SNPExpress	http://compute1.lsrc.duke.edu/softwares/SNPExpress/1_database.php
	HaploReg	http://www.broadinstitute.org/mammals/haploreg/haploreg.php
	ENCODE	https://www.encodeproject.org/
	Roadmap Epigenomics Project	http://www.roadmapepigenomics.org/
Annotation from multiple resources		
Multi-annotation tools	SNPnexus	http://www.snp-nexus.org/
	BioMart - Ensembl	http://www.ensembl.org/biomart/martview
	ANNOVAR, wANNOVAR	http://annovar.openbioinformatics.org/en/latest/, http://wannovar.usc.edu/
	SeattleSeq Variant Annotation	http://snp.gs.washington.edu/SeattleSeqAnnotation141/
	UCSC Genome Browser (visualization), UCSC table browser (data retrieval)	https://genome.ucsc.edu/, https://genome.ucsc.edu/cgi-bin/hgTables
	SCAN	http://www.scandb.org/newinterface/about.html

(continued)

Table 4.3 (continued)

Type	Name	URL
Variation prioritization		
Variation prioritization	Coding variations	
	PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/
	SIFT	http://sift.jcvi.org/
	VnD	http://vnd.kobic.re.kr:8080/VnD/
	Var-MD	http://research.nhgri.nih.gov/software/Var-MD/
	MetaRanker	http://www.cbs.dtu.dk/services/MetaRanker-2.0/
	Coding and non-coding	
	VAAST 2	http://www.yandell-lab.org/software/vaast.html
	RegulomeDB	http://www.regulomedb.org/
	Ingenuity Variant Analysis	https://www.ingenuity.com/products/variant-analysis
	Combined Annotation-Dependent Depletion (CADD)	http://cadd.gs.washington.edu/
	Variant Effect Predictor (VEP)	http://asia.ensembl.org/info/docs/tools/vep/index.html?redirect=no

annotation and prioritization method. It can be a valuable resource that can be used to annotate and prioritize variations based on their effect and function.

4.8 Challenges in Inferring Genotype to Phenotype Associations

To address pertinent biological questions from a large repertoire of variation catalogs and to narrow down huge search space to a few genomic leads, we need to understand the challenges associated before making these inferences.

4.8.1 Genetic Architecture of the Population

Population variation data from different genetic backgrounds is a rich resource and has an enormous applicability if it is used in a systematic and contextual manner. Context or the meaning of a particular variation depends on the genetic architecture of population in question. In simple words, a variation which is beneficial in one population may confer susceptibility/risk in another. Therefore variation is a

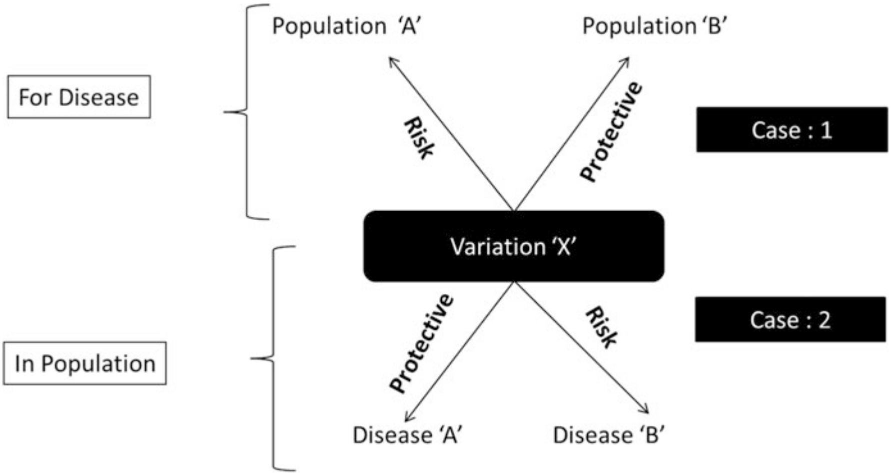


Fig. 4.5 Contextual nature of variations
This figure depicts differential inference of the same variation in a particular disease or in a population. In one scenario, this variation can be a risk factor in one population but protective in another. However, in a different scenario, this variation within a population can confer risk for some disease but protect against others

multifaceted variable. There can be different scenarios (Fig. 4.5) that can explain this property of variation.

4.8.2 Presence of Modifiers or Buffering Variants

Apart from the identification of rare variants which may predispose an individual to a disease or a specific phenotype, many large national and international sequencing projects are now focusing on identifying the modifier or buffering mutations which protect an individual from exhibiting disease phenotype in spite of the presence of the disease variant(s). Such individuals are referred to as "human knockouts," and this also paves the way for the new approach for drug target identification. Apart from rich genomic data, well-curated phenotype data is a prerequisite to conduct such studies. In today's scenario, where we are struggling to make sense of "big data," sequencing a small number of individuals with extensive phenomic data is suggested as a better approach rather than sequencing genomes without any associated phenotypic information (Mich and Glenwoo 2014).

4.8.3 Estimation of Missing Heritability

Put in simple terms, missing heritability means where the associated or implicated genetic markers largely fail to explain their contribution or role in complex disorders because of their small *effect size*. Complex disorders are the outcome of some genetic and environmental interactions; however, contribution of environment toward disease is not measured in association studies. Thus, additive contribution of genetic markers explains small proportion of heritability which implies that there could be other factors such as environment, genetic interactions (epistasis), epigenetics, and other complex structural variants that may contribute to disease heritability (Eichler et al. 2010). Since initial efforts were completely focused on the common disease-common variant (CD-CV) hypothesis, it has been postulated that analyzing association of both rare and common variants could be a better model to explain the heritability. Designing such studies could be expensive as they require larger sample size (Consortium 2015c; Zuk et al. 2014) Integrative omics could be an alternative approach for delineating the etiology of complex disorders. Integrating genetic information with different tiers of regulation such as gene expression, regulatory signatures, epigenomics, and microbiomics via network/system biology approaches provides insights about perturbations in underlying molecular networks in diseased state (Björkegren et al. 2015). Construction of cell-type networks is more informative in understanding the pathology of disease rather than generic networks.

4.8.4 Data Sharing and Interoperability

Linkage-based studies, GWAS, and high-throughput genomic studies have implicated large number of associations in both rare and complex disorders. There are certain challenges which limit the application of data integration for system-wide analyses. Data sharing among the scientific community, data curation (reliability of reported associations), and interoperability (the use of standard nomenclatures, database architecture) are some prerequisites to make informative data repositories or knowledge base (Flannick and Florez 2016). Developments of open-source softwares such as locus-specific database (LSDB) system and genome browsers are some of the successful initiatives toward data formatting, integration, and interoperability (Fokkema et al. 2005; Stein et al. 2002). It allows data submission in standard formats that is scientifically acceptable. Such resources can also be used by users with limited informatics knowledge. These efforts are much needed for the translation of genomics applications to clinical diagnostics and treatment.

With the availability of whole genome data from different populations, the dimension of genomic search space has now been increasing exponentially. Combinatorial possibilities of number, frequency, and type of variations can lead to many outcomes that can explain the phenotypic consequences. Apart from finding meaningful variations from a set of millions, there are computational challenges associated with processing and analysis of “big data.” Every variation is not informative in every context. To churn out a meaningful set of variations, one needs to follow a “Divide and conquer” approach, i.e., minimizing large set of variations into smaller sets in a stepwise fashion.

Aided by the recent advancements in genome editing technologies (like CRISPR-Cas9), variations are no longer static sacrosanct genomic entities. SNPs can now be "engineered" with exemplary precision, and this allows studying their role under different experimental conditions in systems ranging from human embryonic cells to mouse and fish disease models. While the last decade and a half has been spent in comprehensively curating and annotating variation data, the thrust of tomorrow's research should primarily be based upon deciphering their biological functions. With technologies such as CRISPR around the corner, the ground is all set to move beyond association studies to actual perturbation experiments. This would improve the functional annotation of variations in their proper genomic perspective and hence provide a more holistic understanding of them.

Acknowledgments Project funding from the Council of Scientific and Industrial Research (CMM-0016 and MLP-901), DST Inspire and D.S. Kothari Postdoctoral Fellowship to AN, and CSIR-Senior Research Fellowship to AB are duly acknowledged. We acknowledge the efforts of Uma Anwardekar for fruitful comments and discussions.

References

Abdulla MA et al (2009) Mapping human genetic diversity in Asia. Science (New York, NY) 326:1541–1545. https://doi.org/10.1126/science.1177074

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363–376

Andrews CA (2010) Natural selection, genetic drift, and gene flow do not act in isolation in natural populations. Nat Educ Knowl 3:5

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299–309

Auer PL et al (2012) Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO exome sequencing project. Am J Hum Genet 91:794–808

Bernstein BE et al (2010) The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 28:1045–1048

Biesecker LG (2010) Exome sequencing makes medical genomics a reality. Nat Genet 42:13

Biswas S, Akey JM (2006) Genomic insights into positive selection. Trends Genet 22:437–446

Björkegren JL, Kovacic JC, Dudley JT, Schadt EE (2015) Genome-wide significant loci: how important are they?: systems genetics to understand heritability of coronary artery disease and other common complex disorders. J Am Coll Cardiol 65:830–845

Bjornsson HT, Fallin MD, Feinberg AP (2004) An integrated epigenetic and genetic approach to common human disease. Trends Genet 20:350–358

Boyle AP et al (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22:1790–1797

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

Bush WS, Moore JH (2012) Genome-wide association studies. PLoS Comput Biol 8:e1002822

Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11:415–425

Clark MJ et al (2011) Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29:908–914

Consortium EP (2012a) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Consortium GP (2012b) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

Consortium G (2015a) The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348:648–660

Consortium GP (2015b) A global reference for human genetic variation. Nature 526:68–74

Consortium UK (2015c) The UK10K project identifies rare variants in health and disease. Nature 526:82–90

Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

Consortium IGV (2005) The Indian Genome Variation database (IGVdb): a project overview. Hum Genet 118:1–11. https://doi.org/10.1007/s00439-005-0009-9

Consortium IGV (2008) Genetic landscape of the people of India: a canvas for disease gene exploration. J Genet 87:3–20

Consortium TF (2014) A promoter-level mammalian expression atlas. Nature 507:462–470

Coop G et al (2009) The role of geography in human adaptation. Plos Genet 5:e1000500

De S (2011) Somatic mosaicism in healthy human tissues. Trends Genet 27:217–223

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15

Flannick J, Florez JC (2016) Type 2 diabetes: genetic data sharing to advance complex disease research. Nat Rev Genet 17:535

Fokkema IF, den Dunnen JT, Taschner PE (2005) LOVD: easy creation of a locus-specific sequence variation database using an “LSDB-in-a-box” approach. Hum Mutat 26:63–68

Foo J-N, Liu J-J, Tan E-K (2012) Whole-genome and whole-exome sequencing in neurological diseases. Nat Rev Neurol 8:508–517

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251

Freed D, Stevens EL, Pevsner J (2014) Somatic mosaicism in the human genome. Genes 5:1064–1094

Ghanbari M et al (2014) A genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure, and coronary artery disease. Hum Mutat 35:1524–1531

Gibbs RA et al (2003) The international HapMap project. Nature 426:789–796

Gilissen C, Hoischen A, Brunner HG, Veltman JA (2011) Unlocking Mendelian disease using exome sequencing genome. Genome Biol 11:64

Ginsburg GS, Willard HF (2009) Genomic and personalized medicine: foundations and applications. Transl Res 154:277–287

Griffiths AJ, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM (1999) Modern genetic analysis. Freeman, New York

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A (2008) Adaptations to climate in candidate genes for common metabolic disorders. Plos Genet 4:e32

Hancock AM et al (2011) Adaptations to climate-mediated selective pressures in humans. Plos Genet 7:e1001375

Hardy GH (1908) Mendelian proportions in a mixed population. Science 28:49–50

Hartl DL, Clark AG (1997) Principles of population genetics, vol 116. Sinauer associates, Sunderland

He T (1994) Anecdotal, historical and critical commentaries on genetics. Genetics 136:423–426

Hidalgo B et al (2014) Epigenome-wide association study of fasting measures of glucose, insulin, and HOMA-IR in the genetics of lipid lowering drugs and diet network study. Diabetes 63:801–807

Huang J et al (2015) Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun 6:8111

Jablonski NG, Chaplin G (2000) The evolution of human skin coloration. J Hum Evol 39:57–106

Jablonski NG, Chaplin G (2010) Human skin pigmentation as an adaptation to UV radiation. Proc Natl Acad Sci 107:8962–8968

Kamberov YG et al (2013) Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152:691–702

Kaskow BJ et al (2013) Molecular prioritization strategies to identify functional genetic variants in the cardiovascular disease-associated expression QTL Vanin-1. Eur J Hum Genet 22(5): 688–695

Kellis M et al (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci 111:6131–6138

Kimura M (1984) The neutral theory of molecular evolution. Cambridge University Press, New York

Kita R, Fraser HB (2016) Local adaptation of sun-exposure-dependent gene expression regulation in human skin. PLoS Genet 12:e1006382

Ko W-Y et al (2013) Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. Am J Hum Genet 93:54–66

LaFramboise T (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res 37:4181–4193

Laland KN, Odling-Smee J, Myles S (2010) How culture shaped the human genome: bringing genetics and the human sciences together. Nat Rev Genet 11:137–148

Lamason RL et al (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310:1782–1786

Lander ES et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lenormand T (2002) Gene flow and the limits to natural selection. Trends Ecol Evol 17:183–189

Li JZ et al (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104

Luca F, Kashyap S, Southard C, Zou M, Witonsky D, Di Rienzo A, Conzen SD (2009) Adaptive variation regulates the expression of the human SGK1 gene in response to stress. PLoS Genet 5: e1000489

MacArthur DG et al (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in. Nucleic Acids Res 42(Database issue): D986–D992

Manolio TA et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Mardis ER (2013) Next-generation sequencing platforms. Annu Rev Anal Chem 6:287–303

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369

Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6:S13–S20

Mich TB, Glenwoo A (2014) The hunt for missing genes. https://epubs.scu.edu.au/theses/262/

Miller CL et al (2014) Coronary heart disease-associated variation in TCF21 disrupts a miR-224 binding site and miRNA-mediated regulation. PLoS Genet 10:e1004263

Narang A, Roy RD, Chaurasia A, Mukhopadhyay A, Mukerji M, Dash D, Indian Genome Variation Consortium (2010) IGVBrowser–a genomic variation resource from diverse Indian populations. Database: J Biol Database Curation 2010:baq022. https://doi.org/10.1093/database/baq022

Ng SB et al (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30–35

Ngamphiw C et al (2011) PanSNPdb: the Pan-Asian SNP genotyping database. Plos One 6:e21451

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443–451

Pattemore JA (2011) Single nucleotide polymorphism (SNP) discovery and analysis for barley genotyping. https://epubs.scu.edu.au/theses/262/

Perry GH et al (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256–1260

Praetorius C et al (2013) A polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway. Cell 155:1022–1033

Quillen EE, Shriver MD (2011) Unpacking human evolution to find the genetic determinants of human skin pigmentation. J Invest Dermatol 131(E1):E5–E7

Rushton MD et al (2015) Methylation quantitative trait locus analysis of osteoarthritis links epigenetics with genetic risk. Hum Mol Genet 24:7432–7444

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. Genome Res 22:1748–1759

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Sherry ST, Ward M, Sirotkin K (1999) dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9:677–679

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485

Smigielski EM, Sirotkin K, Ward M, Sherry ST (2000) dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 28:352–355

Smith KR et al (2011) Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol 12:R85

Spencer CC, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5:e1000477

Stein LD et al (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12:1599–1610

Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 12:227

Stoneking M, Krause J (2011) Learning about human population history from ancient and modern genomes. Nat Rev Genet 12:603–614

Sturm RA (2009) Molecular genetics of human pigmentation diversity. Hum Mol Genet 18:R9–R17

Syvänen A-C (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930–942

Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The international HapMap project web site. Genome Res 15:1592–1593

Tishkoff SA et al (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293:455–462

Veltman JA, Brunner HG (2012) De novo mutations in human genetic disease. Nat Rev Genetics 13:565–575

Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451–481

Zheng H-F, Rong J-J, Liu M, Han F, Zhang X-W, Richards JB, Wang L (2015a) Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. PLoS One 10: e0116487

Zheng HF et al (2015b) Whole [hyphen] genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526:112–117

Zhou S-F et al (2008) Clinical pharmacogenetics and potential application in personalized medicine. Curr Drug Metab 9:738–784

Zuk O et al (2014) Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci 111:E455–E464

Metagenomics: Focusing on the Haystack 5

Indu Khatri and Meenakshi Anurag

5.1 Introduction

The term “metagenomics” was first used to describe composite genomes of cultured soil microorganisms (Handelsman et al. 1998). In the context of the environmental studies, metagenomics, also known as “community genomics” or “ecogenomics” or “environmental genomics,” is the study of composite genetic material in an environmental sample. There are a large number of microbes that are considered as uncultured in different environments, be it air, soil, water, mines, and animals, and are considered inaccessible for study with traditional approaches. Humans are constantly exposed to a large and diverse pool of microorganism, which can reside in, on, and around our bodies. These microbiotas and their genomes, collectively called as the microbiome, are being characterized by “metagenomics” approaches that integrate next-generation sequencing (NGS) technologies and bioinformatics analysis. The primary focus is on the assembly of 16S ribosomal RNA hypervariable region called as targeted sequencing or whole-genome shotgun DNA sequencing reads. Such studies have been possible because of advances made in the field of genomics and its constant growth in terms of sequencing technology. Apart from this, assembly algorithms and annotation pipelines have provided key opportunities to be exploited by the scientific community. Advances in single-cell genomics, transcriptomics, and metagenomics have revolutionized studies related to cancer genomics, gene expression, metabolic pathway studies, cellular analysis, environmental analysis, and many more areas. There has been tremendous growth in terms

I. Khatri
Leiden University Medical Center, Leiden University, Leiden, The Netherlands

M. Anurag (✉)
Lester & Sue Smith Breast Center & Department of Medicine, Baylor College of Medicine, Houston, TX, USA
e-mail: anurag@bcm.edu

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_5

of sequencing, assembly, and annotation at the genomics level. However, for metagenomics there is a critical need to develop new technologies and in-depth analytical approaches. Here, we present a generalized methodology that can be used for sampling and analysis of metagenomics samples acquired from any environmental location.

5.2 Metagenomics: General Methodology

Metagenomics projects utilize various methodologies which depend on the aim, and a standard metagenomics analysis protocol is depicted in Fig. 5.1. The basic steps in metagenomics analysis including sampling, sequencing, metagenome assembly, binning, annotation of metagenomes, experimental procedures, statistical analysis, and data storage and sharing are discussed.

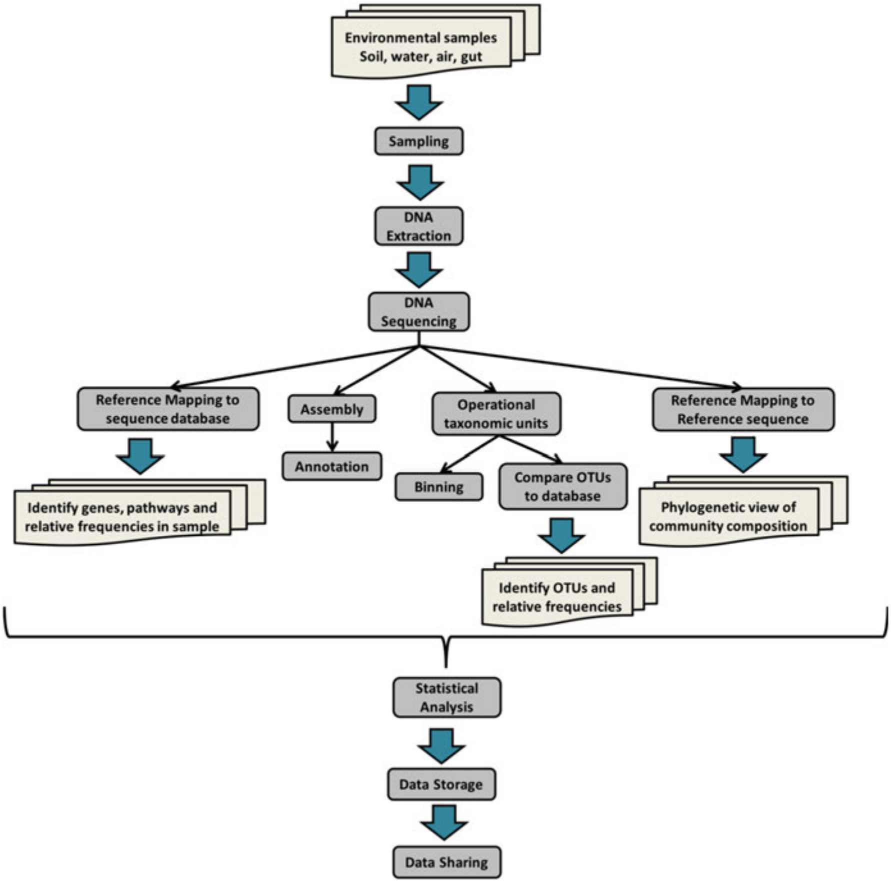

Environmental samples
Soil, water, air, gut
Sampling
DNA Extraction
DNA Sequencing
Reference Mapping to sequence database
Assembly
Operational taxonomic units
Reference Mapping to Reference sequence
Identify genes, pathways and relative frequencies in sample
Annotation
Binning
Compare OTUs to database
Phylogenetic view of community composition
Identify OTUs and relative frequencies
Statistical Analysis
Data Storage
Data Sharing

Fig. 5.1 Flow diagram of a typical metagenomic experiment

5.2.1 Sampling and DNA Extraction

The first and most crucial step is sample acquisition, which critically depends on the sample source. Collection of environmental samples from specific sites across various time points is analyzed in relative metagenomics studies which provide significant insight into both temporal and spatial characteristics of microflora. Another important step in a metagenomics data analysis is the processing of the samples efficiently and to ensure that DNA extracted from the sample represents all the cells present in the sample. In addition, special considerations should be given for sampling and DNA extraction which depends specifically on the sample source. For example, in a soil sample, physical separation and isolation of cells are important for maximizing DNA yield or avoid the co-extraction of enzymatic inhibitors which may interfere further in subsequent sample processing (Delmont et al. 2011). Samples from biopsies or groundwater often yield very small amounts of DNA (Singleton et al. 2011); therefore multiple displacement amplification can be performed (Lasken 2009) to amplify femtograms of DNA to micrograms.

Handling of metagenomics data with precision is a challenge for the scientific community due to large data volume leading to storage issues. Metagenomics data can be exploited for various purposes; therefore, strict and comprehensive guidelines are needed to make data publicly available with a proper format known as metadata. The metadata is known as "data about the data" that contains the when, where, and under what conditions the samples were collected. Metadata is as important as sequence data (Wooley et al. 2010), and minimum information about metagenome sequence (MIMS) contains standard formats that minimally describe the environmental and experimental data. The Genomic Standards Consortium (http://gensc.org/), an international group, has standardized the description, the exchange of genomes and metagenomes, and the rules for the associated metadata.

5.2.2 DNA Sequencing

Sequencing technologies revolutionized the genomics and metagenomics field with high-throughput sequencing. Big, dream projects consisting of sequencing genomes have become a relatively routine task owing to advances in NGS, multiplexing, reduced sequencing cost, and improved algorithms. Metagenomics samples are sequenced in the same manner; however, these samples contain both culturable and non-culturable organisms and also many such genera that have not been exploited yet by the field of genomics. The assignment of taxa to a larger percentage of the metagenome data is still a challenge. Currently, the majority of metagenomics analysis deals with sequencing of the 16S rRNA of the microbial community or a particular gene to trace the community composition which is not typical metagenomics and is referred as metagenetics or metabarcoding. In contrast, whole-genome sequencing is performed on metagenomics samples instead of sequencing a single gene. Of the various NGS sequencing technologies, 454/Roche and Illumina/Solexa have been used extensively for sequencing metagenomics samples. 454/Roche generates longer reads facilitating the assignment of a read to a particular operational taxonomic unit (OTU) which is more reliable as compared to very short reads generated through Illumina high-throughput sequencing.

5.2.3 Assembly and Annotation

The majority of current assemblers have been designed to assemble single, clonal genomes, and their utility for assembling and resolution of large number of complex organisms has to be evaluated critically. Standard assembly methods and algorithms such as de novo assembly and reference mapping are employed in metagenomics data analysis; however, to tackle the significant variation in strain and at the species level, metagenomics assemblers have been designed with the "clonal assumption" that does not allow contig formation for some heterogeneous taxa. Out of various assemblers, *de Bruijn* graph-based assemblers like MetaVelvet (Namiki et al. 2012) and Meta-IDBA (Peng et al. 2011) deal explicitly with non-clonality in sequencing data and try to identify a subgraph that connects related genomes. The meta-assemblers are still in development, and their accuracy assessment is still a major goal of developers as no complete reference exists to which the interpretations can be compared. Assembly is more efficient for genome reconstruction when reference genomes of closely related species are available and in low complex samples (Luo et al. 2013; Teeling and Glockner 2012). However, low read coverage, high frequency of polymorphism, and repetitive regions can hamper the process (De Filippo et al. 2012).

The assembled contigs with minimal length of 30,000 bp or longer can be annotated through existing genome annotation pipelines, such as rapid annotation using subsystem technology (RAST; Aziz et al. 2008) or integrated microbial genomes (IMG; Markowitz et al. 2007, 2009). For the annotation of the entire communities, the standard genome annotation tools are less significant, and a two-step annotation is preferentially followed. First, genes or features of interest are identified, and second, functional assignments are performed by assigning gene functions and taxonomic neighbors (Thomas et al. 2012). FragGeneScan (Rho et al. 2010), MetaGeneMark (McHardy et al. 2007), MetaGeneAnnotator (Noguchi et al. 2008), and Orphelia (Hoff et al. 2009) are the metagenome annotation tools used for defining gene features, e.g., codon usage to find the coding regions. Also, nonprotein-coding genes such as tRNAs (Gardner et al. 2009; Lowe and Eddy 1997), signal peptides (Bendtsen et al. 2004), or clustered regularly interspaced short palindromic repeats (CRISPRs; Bland et al. 2007; Grissa et al. 2007) can be identified but might require long contiguous sequences and vast computational resources.

The functional annotations are provided as gene features via gene or protein mapping to existing nonredundant (NR) protein sequence database. The sequence that cannot be mapped to the known sequence space is termed as ORFans which represents the novel gene contents. ORFans could be erroneous coding sequence (CDS) calls or may be biochemically uncharacterized *bona fide* genes or have no sequence but structural homology to the existing protein families or folds. The reference databases including Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al. 2004), eggnog (Muller et al. 2010), cluster of orthologous groups/eukaryotic orthologous groups (COG/KOG; Tatusov et al. 2003), PFAM (Finn et al. 2014), and TIGRFAM (Selengut et al. 2007) are used to provide functional context to metagenomics CDS. Three prominent systems Metagenome-RAST (MG-RAST; Glass et al. 2010), integrated microbial genomes and microbiomes (IMG/M; Markowitz et al. 2007), and CAMERA (Sun et al. 2011) perform quality control, feature prediction, and functional annotation through standardized protocols and also serve as large repositories of metagenomics datasets. These web servers have a graphical user-friendly interface that assists users to perform taxonomical and functional analysis of metagenomes, which, unfortunately, might be saturated and not customizable at times. Earlier it was reported that the standard metagenome annotation tools can only annotate 20–50% of the metagenomics sequences (Gilbert et al. 2010) and requires further refinement in the annotation algorithms, where sequence and structural homology can be taken into account altogether which is the major computational challenge.

Pathway reconstruction, one of the annotation goals, could be achieved reliably if there is robust functional annotation. To reconstruct a pathway, every gene should be in an apt metabolic context, missing enzymes should be filled in the pathways, and optimal metabolic states should be found. MinPath (Ye and Doak 2009) and MetaPath (Liu and Pop 2011) use KEGG (Kanehisa et al. 2004) and MetaCyc (Caspi et al. 2014) repositories for building networks. Most of the current platforms are not able to reconstruct variant metabolic pathways (de Crécy-Lagard 2014), since pathways and enzymes are not conserved among different environment and the inhabiting species. A web service implementation by KEGG, GhostKOALA (Kanehisa Laboratories www.kegg.jp/ghostkoala/), relates taxonomic origin of the metagenomes with their respective functional annotation, and the metabolic pathways from different taxa can be visualized in a composite map. Metabolic pathways can be constructed using gene-function interactions, synteny, and copy number of annotated genes and integrating them with the metabolic potential of metagenome consortium.

5.2.4 Taxonomic Classification and Binning

Binning, as name suggests, is to group the sequencing reads representing an individual genome or genomes of closely related organisms. The algorithms employed in grouping related sequences act either as supervised classifiers or unsupervised classifiers. Binning can be performed based on either sequence similarity/alignment or compositional features or both. Another strategy employed by tools is compositional binning that bins the genomes based on the property of conserved nucleotide composition that carry weak but detectable phylogenetic signals, e.g., GC content or particular K-mer (tetramer or hexamer) abundance distribution (Pride et al. 2003), or based on similarity-based binning where the unknown DNA fragments are binned according to the known genes in the reference database. Compositional-based binning algorithms have been exploited in PhyloPythia (McHardy et al. 2007) and PCAHIER (Zheng and Wu 2010), whereas a similarity-based binning algorithm was employed in IMG/M (Markowitz et al. 2007), MG-RAST (Glass et al. 2010), MEtaGenome ANalyzer (MEGAN; Huson et al. 2016), CARMA (Krause et al. 2008), MetaPhyler (Liu et al. 2010), and many more. Some programs such as PhymmBL (Brady and Salzberg 2009) and MetaCluster (Leung et al. 2011) employ both compositional- and similarity-based algorithms. All these tools employ either an unsupervised or supervised approach to define the bins. The compositional-based binning is not reliable for short reads of approximately 100 bp length, but if reference data is available, then with supervised similarity-based method, the taxonomic assignment of the read can be made (McHardy et al. 2007). The bins obtained will be assigned taxonomy at the phylum level which is very high and results in chimeric bins composed of two or more genomes that belong to the same phylum. The similarity-based binning algorithm if improved to assignments at lower taxonomic levels may help in creating accurate bins for a specific organism at least to a species level. Such binned reads can be assembled to obtain partial genomes of yet-uncultured or unknown organisms. The binning of reads before assembling reduces the complexity of assembly efforts and computational requirements.

The metabolic potential of the metagenome can be deciphered after the microbial diversity is known. Whole-metagenome approach where whole DNA of the community is sequenced can be used to obtain the complete information of a microbial community. The choice of sequencing platform will influence the computational resources and selection of available software to process the sequencing results. These choices in turn will be reflected in taxonomic species/genus/family level classification. Novel microorganisms identified from the analysis can potentially establish new genes with novel functions.

Taxonomic annotation can be made better by using more than one phylogenetic marker. Metagenome shotgun sequencing allows for the identification of single copy marker genes among various databases. Parallel-META (Su et al. 2014) can be used to extract ribosomal marker genes from metagenomics sequences to conduct taxonomic annotations. Single copy marker genes can be extracted using MOCAT (Kultima et al. 2012) that uses the RefMG database (Ciccarelli et al. 2006), a collection of 40 single copy universal marker genes, and “a pipeline for AutoMated PHylogenOmic infeRence” (AMPHORA; Wu and Eisen 2008), a database with 31 single copy marker genes. This pipeline, distinct from identification of marker genes, performs multiple sequence alignment, distance calculations, and clustering. The reference genomes were used to perform taxonomic annotation at a species-level resolution.

5.2.5 Statistical Analysis

The metagenomics data consists of large number of species, corresponding genes, and their functions as compared to the number of samples analyzed. Thus, multiple hypotheses are to be formed, tested, and implemented for comprehensive presentation of data. Various multivariate statistical visualization programs such as Metastats (White et al. 2009) and R packages, viz., ShotgunFunctionalizeR (Kristiansson et al. 2009), have been built to statistically analyze the metagenome data.

5.2.6 Data Storage and Sharing

Genome research has always been connected to sharing raw data, the final assemblies and annotations; however, to store metagenomics data, database management and storage system are required. All the data is stored at the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), and other metagenomics repositories. The digital form of data storage is generally preferred, and despite the decreasing cost of generating NGS data, storage costs may not decline (Weymann et al. 2017); therefore, acquiring data storage in a cost-effective manner is also important.

The microbial systems can be very dynamic at different time points, e.g., as in the human gut; therefore, temporal sampling has substantial impact on data analysis, interpretations, and results (Thomas et al. 2012). Due to the magnitude of variation in small-scale experiments (Prosser 2010), a sufficient number of replicates are needed. Samples should be collected from the same habitat and should be processed in a similar fashion. The experimental plan and interpretations, if done carefully, facilitate dataset integration into new or existing theories (Burke et al. 2011). The critical aim of metagenomics projects is to relate functional and phylogenetic information to the biological, chemical, and physical characteristics of that environment and ultimately achieve retrospective correlation analysis.

5.3 Species Diversity

The diversity of species in an environmental sample is a critical question where the vast majority of marker genes have been used to classify metagenomics reads. Species-specific gene markers such as 16S/18S ribosomal DNA (rDNA) sequences have been used to estimate the species diversity and coverage in most of the analyses. rDNA as a marker gene has limitations including horizontal transfers within microbes (Schouls et al. 2003) and the presence of multiple copies of the marker gene (DeSantis et al. 2006). Other housekeeping genes such as *rpoB* (Walsh et al. 2004) are strong candidates, and also *amoA*, *pmoA*, *nirS*, *nirK*, *nosZ*, and *pufM* (Case et al. 2007) have been exploited in different contexts as molecular markers.

Quantifying species diversity is not trivial due to the incorporation of species richness, evenness of species, or differential abundance (Simpson 1949). In comparison of two communities, if both the communities have the same number of species but their abundance varies, then the community with the shortest difference with "assumed even abundance" will be considered as more diverse.

The diversity indices of the species are measured as α-diversity, β-diversity, and γ-diversity in ecology and microbial ecology. The α-diversity is defined as the biodiversity in a defined habitat (i.e., a smaller ecosystem), whereas β-diversity compares species diversity between habitats (or between two ecosystems).The γ-diversity is considered as the total biodiversity over a large region containing several ecosystems (Wooley et al. 2010). Rarefraction curves are used to estimate the coverage obtained from sampling which tells whether the species in a particular habitat has been exhaustively sampled or not. All these indices are calculated in metagenomics data analysis by employing various software and tools including EstimateS (Colwell et al. 2004), Quantitative Insights Into Microbial Ecology (QIIME; Caporaso et al. 2010), and Kraken (Davis et al. 2013). Another method to calculate species diversity is through the use of statistical estimators, in particular nonparametric estimators. Simpson's index (Simpson 1949) is based on the probability of the same species taken randomly from the community and is used to assign two independent subjects. The Shannon–Wiener index *H´* (Shannon 1948) is an entropy measurement and is directly proportional to the number of species in the sample. These methods are used for heterogeneity measurements and differ primarily in calculating the taxa abundance to measure the final richness estimation (Escobar-Zepeda et al. 2015). Simpson and Shannon–Wiener indices prioritize more-frequent and rare species, respectively, in the sample (Krebs 2014).

The use of diversity indices which quantify and compare microbial diversity among samples is a better approach as compared to ones based on molecular markers. The species diversity analysis should be done carefully as it can be uninformative. The biases related to sampling should be reduced considering the criteria for species or OTU definition.

5.4 Comparative Metagenomics

The comparison between two or more metagenomes facilitates the understanding of genomic differences and how they are affected by the abiotic environment. Various sequence-based traits such as GC content (Yooseph et al. 2007), microbial genome size (Raes et al. 2007), taxonomy (von Mering et al. 2007), and functional content (Turnbaugh et al. 2006) have been compared to gather biological insights through comparison between two or more metagenomes. Statistical analysis is a necessity to analyze several metagenomics datasets, and principal component analysis (PCA) and nonmetric multidimensional scaling (NM-MDS) have been used to visualize the metagenomics data analysis and reveal major factors that affect the data most (Brulc et al. 2009).

5.5 Challenges in Metagenomics Analysis

Sequencing of a complex environmental community for metagenomics analysis often represents only a minute fraction of the vast number of culturable and unculturable microorganisms actually present (Desai et al. 2012). To obtain just onefold coverage of the entire community in a gram of soil requires hundreds of millions of reads without guarantee that every member of that community was sequenced. The unknown community composition and relative abundance of microorganisms limits our ability to calculate the coverage robustly. Even perfect 16S amplicon-based characterization of microbial species fails to distinguish between different strains (Desai et al. 2012). Furthermore, no tools are available that determine the availability of sufficient coverage to interpret data of a certain depth for a community. The low coverage data represents randomly subsampled genomic content of the community. Despite complete coverage with millions invested, the analysis of metagenomics data requires tools and protocol development comparable to genomic analysis. Moreover, if the approaches led to the identification of new microbial community members and discovery of new molecules, problems associated with cloning biases, sampling biases, misidentification of "decorating enzymes" and incorrect promoter sites in genomes, and dispersion of genes involved in secondary metabolite production (Escobar-Zepeda et al. 2015) should be considered.

Similarly, human metagenomic experiments and analysis also have associated limitations and pitfalls as they are sensitive to the environment including any particular condition or intervention (Kim et al. 2017). Various factors including diet, drugs, age, geography, and sex have all been reported to influence function and composition of the human microbiome (Blaser et al. 2013;Dave et al. 2012; Lozupone et al. 2012). Another challenge is the longitudinal stability. Unlike gut, the microbiome of other sites, like the human vagina, can vary in short periods without always indicating dysbiosis (Williams and Lin 1971). In animal experiments, the prime limitation is the cage effect, which is best studied in mice kept in the same cage and can share the same microbiome because of coprophagia (Campbell et al. 2012). When it comes to handling and analyzing samples, issues pertaining to low microbial biomass, environmental contamination, and presence of negative/positive control samples should be addressed. The major informatics challenges associated with human metagenome analysis, similar to other metagenomes, are the large volume and bulkiness of the data and the heterogeneous microbial community. One additional challenge has been the rapid identification of host sequences contaminating metagenomics datasets, which is time- and memory-extensive process and hence needs to be revisited. There have been efforts to overcome these challenges with tools like CS-SCORE (Haque et al. 2015); however, algorithm improvement is needed.

5.6 Applications of Metagenomics

5.6.1 Correlations Between Environmental Data and Metadata

Metagenomics studies aid in investigating genomic potential of the bacterial community and how it is affected by and is affecting its habitat. The correlation between sequence data, environment, and environmental attributes or their correlation among themselves reveals new biological insights. For example, a bivariate metagenome study in obese vs lean mouse reveals that obese individuals are enriched in carbohydrate-active enzymes (Turnbaugh et al. 2006). Multivariate correlation analysis in a nutrient poor ocean habitat revealed covariation in amino acid transport and cofactor synthesis molecules (Gianoulis et al. 2009).

5.6.2 Investigating Symbiosis

Symbiotic relationships occur when two or more organisms are symbionts which represent a small-scale metagenomics and can be analyzed in a similar fashion. The organisms in symbiotic relations are few, and their distance to each other phylogenetically eases the binning of the reads in separate bins and can be assembled separately. Wu and collegues (2006) exploited a similar method to bin the ESS data from bacterial symbionts living in the glassy-winged sharpshooter and inferred that one member of a symbiont synthesizes amino acids for the host insect, while the other produces cofactors and vitamins (Wu et al. 2006).

5.6.3 Gene Family Enrichment

The immense amount of genetic material has led to the possibility of associating new gene families with new members of existing gene families. The small bacterial eukaryotic protein kinase-like (ELK) gene family was enriched severalfold through the Global Ocean Sampling (GOS) metagenomics project (Wooley et al. 2010).

5.6.4 Human Microbiome

Symbiotic microbes have coevolved with humans for millions of years and play a critical role in health of the host. The focus of human microbiome research has been on the bacteria residing in the gut, which represents the most abundant and diverse part of the human microbiome (Consortium 2012). Colonization of these bacteria commences at birth, and the method of delivery (i.e., vaginal or cesarean section) influences the basal community (Dominguez-Bello et al. 2010). Early-life events, such as mode of delivery (Fig. 5.2 – adapted from Rutayisire et al. 2016), dietary transitions or restrictions (Bergstrom et al. 2014; Rutayisire et al. 2016), and antibiotic use (Cho et al. 2012), shape the dynamic microbiome of infants. This gradually

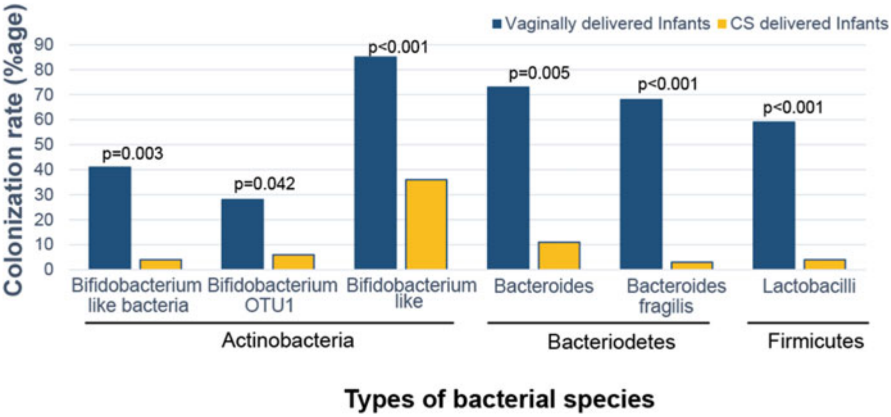


Fig. 5.2 Microbiota colonization pattern significantly associated with the mode of delivery during the first 7 days after birth. Bacterial species with quantified colonization rate has been shown. (Adapted from Rutayisire et al. 2016)

stabilizes with age and leads to adult gut microbiota, which is highly resilient to minor perturbations. This longitudinal stability, collectively with vast interpersonal diversity of the microbiome, allows identification of ~80% individuals by their distinct "microbial fingerprint" (Franzosa et al. 2015). The human microbiota communities contribute to various host biological processes, thus deeply influencing human health. Global initiatives have been taken to understand the healthy microbiome and its composition.

5.6.5 Metagenomics in Diseases

Recent findings have emphasized the effect of gut microbiome in human health and therapeutic response (Scarpellini et al. 2015). The gut microbiome, primarily, is composed of viruses and fungi and has been shown to be modulated in diet-associated insulin resistance in type 2 diabetic patients using a metagenome-wide association analysis (Qin et al. 2012). Gut microbiota has been established as a metformin action site, and metformin–microbiota interactions have been studied to show that altered gut microbiota mediates some of metformin's antidiabetic effects (Wu et al. 2017). The Human Pan-Microbe Communities (HPMC) database (http://www.hpmcd.org/) is an excellent source of highly curated, searchable, metagenomic resource focusing on facilitating the investigation of human gastrointestinal microbiota (Forster et al. 2016).

Historically, cancer has been associated with different forms of microorganisms. The metagenomics era has revolutionized microbiome profiling which helps to boost a number of studies exploring microbial linkage to cancer. Several studies on microbes and cancers have shown distinct associations between various viruses and different types of cancers. Human papilloma virus (HPV) causes the majority of cervical, anal, and oropharyngeal cancer (Chaturvedi et al. 2011; Daling et al. 2004; Gillison et al. 2008; Winer et al. 2006). Similarly, Epstein–Barr virus has been found to be responsible for nasopharyngeal carcinoma, Hodgkin's, Burkitt's lymphoma, etc. (Anagnostopoulos et al. 1989; Henle and Henle 1976; Leung et al. 2014).

5.6.6 Clinical Implications

In translating the role of microbiomes into clinical applications, Danino et al. (2015) engineered a probiotic *E. coli* to harbor specific gene circuits that produce signals allowing detection of tumor in urine, in case of liver metastases. This concept was based on the fact that metastasis leads to translocation of the probiotic *E. coli* to the liver. Metagenomics has also allowed physicians to probe complex phenotypes such as microbial dysbiosis with intestinal disorders (Antharam et al. 2013) and disruptions of the skin microbiome that may be associated with skin disorders (Weyrich et al. 2015). Recently, different bacterial profiles in the breast were observed between healthy women and breast cancer patients. Interestingly, higher abundances of DNA damage causing bacteria were detected in breast cancer patients, along with decrease in some lactic acid bacteria, known for their beneficial health effects (Urbaniak et al. 2016). Such studies raise important questions regarding the role of the mammary microbiome in risk assessment to develop breast cancer.

Metagenomics analytics is changing rapidly with evolutions of tools and analysis procedures in terms of scalability, sensitivity, and performance. The field allows us to discover new genes, proteins, and the genomes of non-cultivable organisms with better accuracy and less time as compared to classical microbiology or molecular methods. However, no standard tool or method is available that can answer all our questions in metagenomics. The lack of standards reduces reproducibility and is still a case by case study. The major problem associated with metagenomics study is also data management as most institutes lack computational infrastructure to deal with long-term storage of raw, intermediate data, and final analyzed datasets.

Comparison between different biomes and different environmental locations will provide insight into the microflora distribution and help understand the environment around us.

All the advances in the field of human metagenomics add up to the profound impact that the microbiome and their metagenomics have on human health in providing new diagnostic and therapeutic opportunities. However, existing therapeutic approaches for modulating microbiomes in the clinic remain relatively underdeveloped. More studies focused on metagenomics of different organs need to be performed, comparing the tissues from healthy versus affected individuals. Further exploration of additive, subtractive, or modulatory strategies affecting the human microbiota and its clinical implementation could potentially be the next big milestone in the field of translational and applied microbiology. The near future challenge is in the accurate manipulation and analysis of the vast amounts of data and to develop approaches to interpret data in a more integrative way that will reflect the biodiversity present in our world. The development of more bioinformatics tools for metagenomics analysis is necessary, but the expertise of scientific community to manipulate such tools and interpret their results is a critical parameter for successful metagenomics studies.

References

Anagnostopoulos I, Herbst H, Niedobitek G, Stein H (1989) Demonstration of monoclonal EBV genomes in Hodgkin's disease and Ki-1-positive anaplastic large cell lymphoma by combined Southern blot and in situ hybridization. Blood 74:810–816

Antharam VC, Li EC, Ishmael A, Sharma A, Mai V et al (2013) Intestinal dysbiosis and depletion of butyrogenic bacteria in Clostridium difficile infection and nosocomial diarrhea. J Clin Microbiol 51:2884–2892

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T et al (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Bergstrom A, Skov TH, Bahl MI, Roager HM, Christensen LB et al (2014) Establishment of intestinal microbiota during early life: a longitudinal, explorative study of a large cohort of Danish infants. Appl Environ Microbiol 80:2889–2900

Bland C, Ramsey TL, Sabree F, Lowe M, Brown K et al (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinf 8:209

Blaser M, Bork P, Fraser C, Knight R, Wang J (2013) The microbiome explored: recent insights and future challenges. Nat Rev Microbiol 11:213–217

Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods 6:673–676

Brulc JM, Antonopoulos DA, Miller MEB, Wilson MK, Yannarell AC et al (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc Natl Acad Sci U S A 106:1948–1953

Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T (2011) Bacterial community assembly based on functional genes rather than species. Proc Natl Acad Sci 108:14288–14293

Campbell JH, Foster CM, Vishnivetskaya T, Campbell AG, Yang ZK et al (2012) Host genetic and environmental effects on mouse intestinal microbiota. ISME J 6:2033–2044

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336

Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol 73:278–288

Caspi R, Altman T, Billington R, Dreher K, Foerster H et al (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 42:D459–D471

Chaturvedi AK, Engels EA, Pfeiffer RM, Hernandez BY, Xiao W et al (2011) Human papillomavirus and rising oropharyngeal cancer incidence in the United States. J Clin Oncol 29:4294–4301

Cho I, Yamanishi S, Cox L, Methe BA, Zavadil J et al (2012) Antibiotics in early life alter the murine colonic microbiome and adiposity. Nature 488:621–626

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287

Colwell RK, Mao CX, Chang J (2004) Interpolating, Extrapolating, and comparing incidence-based species accumulation curves. Ecology 85:2717–2727

Consortium THMP (2012) Structure, function and diversity of the healthy human microbiome. Nature 486:207–214

Daling JR, Madeleine MM, Johnson LG, Schwartz SM, Shera KA et al (2004) Human papillomavirus, smoking, and sexual practices in the etiology of anal cancer. Cancer 101:270–280

Danino T, Prindle A, Kwong GA, Skalak M, Li H et al (2015) Programmable probiotics for detection of cancer in urine. Sci Transl Med 7:289ra284

Dave M, Higgins PD, Middha S, Rioux KP (2012) The human gut microbiome: current knowledge, challenges, and future directions. Transl Res: J Lab Clin Med 160:246–257

Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ (2013) Kraken: A set of tools for quality control and analysis of high-throughput sequence data. Methods 63:41–49

de Crécy-Lagard V (2014) Variations in metabolic pathways create challenges for automated metabolic reconstructions: Examples from the tetrahydrofolate synthesis pathway. Comput Struct Biotechnol J 10:41–50

De Filippo C, Ramazzotti M, Fontana P, Cavalieri D (2012) Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Brief Bioinform 13:696–710

Delmont TO, Robe P, Clark I, Simonet P, Vogel TM (2011) Metagenomic comparison of direct and indirect soil DNA extraction approaches. J Microbiol Methods 86:397–400

Desai N, Antonopoulos D, Gilbert JA, Glass EM, Meyer F (2012) From genomics to metagenomics. Curr Opin Biotechnol 23:72–76

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL et al (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072

Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G et al (2010) Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A 107:11971–11975

Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A (2015) The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Front Genet 6:348

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Forster SC, Browne HP, Kumar N, Hunt M, Denise H et al (2016) HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. Nucleic Acids Res 44:D604–D609

Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP et al (2015) Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A 112:E2930–E2938

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL et al (2009) Rfam: updates to the RNA families database. Nucleic Acids Res 37:D136–D140

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO et al (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci U S A 106:1374–1379

Gilbert JA, Field D, Swift P, Thomas S, Cummings D et al (2010) The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. PLoS ONE 5:e15545

Gillison ML, Chaturvedi AK, Lowy DR (2008) HPV prophylactic vaccines and the potential prevention of noncervical cancers in both men and women. Cancer 113:3036–3046

Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb Protoc 2010: pdb.prot5368

Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35:W52–W57

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245–R249

Haque MM, Bose T, Dutta A, Reddy CV, Mande SS (2015) CS-SCORE: rapid identification and removal of human genome contaminants from metagenomic datasets. Genomics 106:116–121

Henle G, Henle W (1976) Epstein-Barr virus-specific IgA serum antibodies as an outstanding feature of nasopharyngeal carcinoma. Int J Cancer 17:1–7

Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res 37:W101–W105

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M et al (2016) MEGAN community edition – interactive exploration and analysis of large-scale microbiome sequencing data. PLOS Comput Biol 12:e1004957

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32:277D–280D

Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C et al (2017) Optimizing methods and dodging pitfalls in microbiome research. Microbiome 5:52

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW et al (2008) Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res 36:2230–2239

Krebs C (2014) Species diversity measures. In: Ecological methodology. Addison-Wesley Educational Publishers, Inc, Boston

Kristiansson E, Hugenholtz P, Dalevi D (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics 25:2737–2738

Kultima JR, Sunagawa S, Li J, Chen W, Chen H et al (2012) MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS ONE 7:e47656

Lasken RS (2009) Genomic DNA amplification by the multiple displacement amplification (MDA) method. Biochem Soc Trans 37:450–453

Leung HCM, Yiu SM, Yang B, Peng Y, Wang Y et al (2011) A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics 27:1489–1495

Leung SF, Chan KC, Ma BB, Hui EP, Mo F et al (2014) Plasma Epstein-Barr viral DNA load at midpoint of radiotherapy course predicts outcome in advanced-stage nasopharyngeal carcinoma. Ann Oncol 25:1204–1208

Liu B, Pop M (2011) MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. BMC Proc 5:S9

Liu B, Gibbons T, Ghodsi M, Pop M (2010) MetaPhyler: taxonomic profiling for metagenomic sequences. In: 2010 I.E. international conference on Bioinformatics and Biomedicine (BIBM). IEEE, Hong Kong, pp 95–100

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R (2012) Diversity, stability and resilience of the human gut microbiota. Nature 489:220–230

Luo C, Rodriguez-R LM, Konstantinidis KT (2013) A user's guide to quantitative and comparative analysis of metagenomic datasets. Methods Enzymol 531:525–547

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K et al (2007) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36:D534–D538

Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Chu K, Kyrpides NC (2009) IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics 25:2271–2278

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 4:63–72

Muller J, Szklarczyk D, Julien P, Letunic I, Roth A et al (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 38:D190–D195

Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40:e155–e155

Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15:387–396

Peng Y, Leung HCM, Yiu SM, Chin FYL (2011) Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27:i94–i101

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res 13:145–158

Prosser JI (2010) Replicate or lie. Environ Microbiol 12:1806–1810

Qin J, Li Y, Cai Z, Li S, Zhu J et al (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490:55–60

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. Genome Biol 8:R10

Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 38:e191–e191

Rutayisire E, Huang K, Liu Y, Tao F (2016) The mode of delivery affects the diversity and colonization pattern of the gut microbiota during the first year of infants' life: a systematic review. BMC Gastroenterol 16:86

Scarpellini E, Ianiro G, Attili F, Bassanelli C, De Santis A, Gasbarrini A (2015) The human gut microbiota and virome: Potential therapeutic implications. Dig Liver Dis 47:1007–1012

Schouls LM, Schot CS, Jacobs JA (2003) Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. J Bacteriol 185:7241–7246

Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M et al (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res 35:D260–D264

Shannon CE (1948) A mathematical theory of communication, Part I. Bell Syst Tech J 27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Simpson EH (1949) Measurement of diversity. Nature 163:688

Singleton DR, Richardson SD, Aitken MD (2011) Pyrosequence analysis of bacterial communities in aerobic bioreactors treating polycyclic aromatic hydrocarbon-contaminated soil. Biodegradation 22:1061–1073

Su X, Pan W, Song B, Xu J, Ning K (2014) Parallel-META 2.0: enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization. PLoS ONE 9:e89323

Sun S, Chen J, Li W, Altintas I, Lin A et al (2011) Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. Nucleic Acids Res 39:D546–D551

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinform 4:41

Teeling H, Glockner FO (2012) Current opportunities and challenges in microbial metagenome analysis – a bioinformatic perspective. Brief Bioinform 13:728–742

Thomas T, Gilbert J, Meyer F (2012) Metagenomics – a guide from sampling to data analysis. Microb Inf Exp 2:3

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1131

Urbaniak C, Gloor GB, Brackstone M, Scott L, Tangney M, Reid G (2016) The Microbiota of Breast Tissue and Its Association with Breast Cancer. Appl Environ Microbiol 82:5039–5048

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T et al (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315:1126–1130

Walsh DA, Bapteste E, Kamekura M, Doolittle WF (2004) Evolution of the RNA polymerase B′ subunit gene (rpoB′) in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. Mol Biol Evol 21:2340–2351

Weymann D, Laskin J, Roscoe R, Schrader KA, Chia S, Yip S, Cheung WY, Gelmon KA, Karsan A, Renouf DJ, Marra M, Regier DA (2017) The cost and cost trajectory of whole-genome analysis guiding treatment of patients with advanced cancers. Mol Genet Genomic Med 5:251–260

Weyrich LS, Dixit S, Farrer AG, Cooper AJ, Cooper AJ (2015) The skin microbiome: associations between altered microbial communities and disease. Aust J Dermatol 56:268–274

White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol 5:e1000352

Williams HR, Lin TY (1971) Methyl- 14 C-glycinated hemoglobin as a substrate for proteases. Biochim Biophys Acta 250:603–607

Winer RL, Hughes JP, Feng Q, O'Reilly S, Kiviat NB et al (2006) Condom use and the risk of genital human papillomavirus infection in young women. N Engl J Med 354:2645–2654

Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6: e1000667

Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9:R151

Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL et al (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol 4:e188

Wu H, Esteve E, Tremaroli V, Khan MT, Caesar R et al (2017) Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. Nat Med 23:850–858

Ye Y, Doak TG (2009) A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput Biol 5:e1000465

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ et al (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:e16

Zheng H, Wu H (2010) Short prokaryotic DNA fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. J Bioinform Comput Biol 8:995–1011

Computational Epigenomics and Its Application in Regulatory Genomics

6

Shalu Jhanwar

6.1 Introduction

A genome contains the entire genetic instructions essential to develop and direct the activities of an organism, thereby acting as a blueprint for gene regulation. However, the genome does not work in isolation. The regulation of the chromatin and gene expression is affected by tissues and cell types, developmental stages, aging, as well as diverse surrounding factors including chemical pollutants, dietary components, and temperature changes. Therefore, the interpretation of the instructions provided by the genome differs among the cell types irrespective of having the same genome in all the (nucleated) human cells. To describe the mechanisms of cell fate and lineage specification during animal development, Conrad Hal Waddington first introduced the term *epigenetic landscape* (Waddington 2012). Thus, epigenetics endeavored to bring together two important bio-streams, i.e., genetics and developmental biology, to unfold the genetic program for development. In contrast to the early epigenetics that was originated exclusively in embryology and development, the modern epigenetics emphasizes on defining mechanisms of transmission of information that are not encoded in DNA (Felsenfeld 2014). In a nutshell, *epigenetics* deals with the effects of changes in chromatin structure excluding any modification in the primary DNA sequence that subsequently leads to heritable alterations in gene expression (Wu and Morris 2001). The epigenetic mechanisms may result in either activation or repression of regulatory elements as well as genes by compacting (heterochromatin) and unfolding (euchromatin) the chromatin, respectively. Thus, linking genomics with epigenetics is crucial to determine the

S. Jhanwar (✉)
Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Universitat Pompeu Fabra (UPF), Barcelona, Spain

Developmental Genetics, Department of Biomedicine, University of Basel, Basel, Switzerland
e-mail: shalu.jhanwar@unibas.ch

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_6

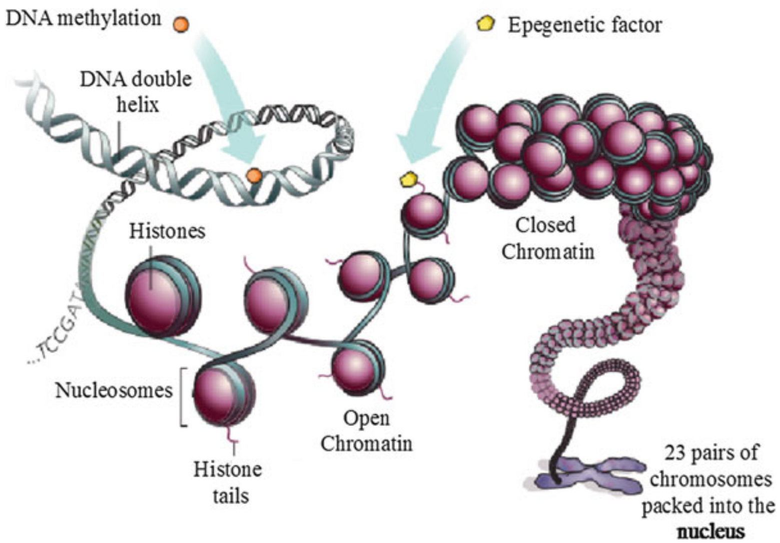


Fig. 6.1 The epigenetic code: an example of a higher order of chromatin folding, where DNA is packaged around nucleosomes. The DNA double helix wraps around histones that contain unstructured N- or C-terminal tails and a globular structure domain. (Adapted from Marx 2012)

dynamics of chromatin states that shapes underlying gene expression. In this chapter, an application of epigenetics in the field of regulatory genomics, specifically enhancer-mediated regulation, is explored. With the extensive usage of NGS, the paradigm has shifted from the experimental-based characterization to the omics-based characterization of enhancers. In particular, a number of sophisticated machine-learning-based enhancer predictors have been widely used to integrate information of genomic and epigenomic features. Thus, the catalog of enhancers and its annotation has increased rapidly.

The important concepts learned so far indicate that the modern epigenetics comprises of covalent modifications of DNA bases, wrapping of DNA around the nucleosomes, posttranslational modifications of histones, noncoding RNA-mediated mechanisms, and a higher order of chromatin folding (Fig. 6.1).

DNA methylation, first reported by Hotchkiss (1948), is a heritable epigenetic mark that directly modifies the bases in DNA. The predominant modifications in animal and plant DNA involve the covalent transfer of a methyl group to cytosine nucleotides through DNA methyltransferases (DNMTs), followed by adenine and guanine methylation. DNA methylation plays a fundamental role in the maintenance and regulation of crucial biological processes such as pluripotency, X-chromosome inactivation, embryonic development, genomic imprinting, regulation of chromatin structure, chromosome stability, and transcriptional activity. When DNA methylation is dysregulated, it contributes to diseases ranging from cancer to neurodegenerative and autoimmune disorders (Robertson 2005).

The DNA double helix wraps around tripartite proteins known as *histones* that contain unstructured N- or C-terminal tails and a structured globular domain (Fig. 6.1). The flexible tails may undergo several posttranslational modifications (PTMs) such as acetylation, methylation, phosphorylation, and ubiquitylation (tan et al. 2011). These PTMs either directly change the chromatin structure and dynamics (Choi 2013) or work via an indirect mechanism by communicating with "readers" like chromatin remodeling agents, histone acetyltransferases (HATs), and transcriptional coactivators (P300) (Strahl and Allis 2000). Moreover, recent studies have precisely demonstrated their role to directly shape nucleosome functions (Xu et al. 2005). The patterns of histone PTMs correlate with the distinct chromosomal states that regulate access to the open chromatin (DNA accessibility) for precise binding of DNA-binding proteins, thus, leading to the concept of the *histone-code hypothesis* (see details in Sect. 6.3).

Further, eukaryotic chromatin is made up of subunits called nucleosomes; each consists of histone octamer core wrapped around by 147 bp of DNA. The position of nucleosomes across a genome plays a significant regulatory function by modifying the accessibility of the binding sites of transcription factors (TFs) and the transcriptional machinery. This affects processes such as DNA repair, DNA transcription, and DNA replication. The regions possessing regulatory activity, i.e., regulatory elements (RE), generally reside in open or accessible genomic regions. A wide range of recently developed high-throughput sequencing-based approaches such as DNase-seq (Song and Crawford 2010), ATAC-seq (Buenrostro et al. 2013), FAIRE-seq (Giresi et al. 2007), and MNase-seq (Schones et al. 2008) provide a comprehensive view of nucleosome-depleted open accessible DNA at genome-wide level. When integrated with DNA-binding protein data, they enable to localize and delineate REs controlling cell fate.

A major quest to understand *what molecular mechanisms underlie gene regulation* lies in understanding the constraints of interactions between the genes and their surrounding REs. In particular, *how the 3D landscape of the genome confines enhancer-promoter interactions* is central to the scientific community (Plank and Dean 2014). Recently developed chromosome conformation capture (3C)-based methods have made great strides toward dissecting these chromatin interactions (Dekker et al. 2013). Present consensus supports the hypothesis that enhancers make direct physical contact with promoter regions through "looping" mechanism to form a compact topology (de Wit and de Laat 2012). It has become possible to capture "one to one," "one to all," and "many to many" forms of 3D interactions throughout the genome. A close spatial proximity may result in direct or indirect and specific or nonspecific random interaction between the pair of a locus (Dekker et al. 2013).

6.2 Computational Epigenomics

Over the past decades, a succession of robust technological advances in next-generation sequencing (NGS) has allowed us to resolve the genome-wide maps of epigenomic features with high accuracy and comprehensiveness.

A timeline of sequencing-based technologies for mapping of human epigenomes is presented in Fig. 6.2 (Martens and Stunnenberg 2013). These key technology platforms of DNA methylation and chromatin profiling could be easily integrated with expression profiling and existing genome data to study the underlying complex regulatory mechanisms guiding gene expression. The availability of large-scale epigenome data in public domain has created ample of opportunities and inspired computational efforts that aim to develop new algorithms to systematically analyze and integrate different types of genomic and epigenomic data.

6.2.1 DNA Methylation

Commonly implemented experimental methods involve three relevant approaches (Bock 2012): (1) genomic DNA digestion with methyl-sensitive restriction enzymes (MRE-seq) (Maunakea et al. 2010), (2) affinity-based enrichment of methylated DNA regions by either a methyl-binding domain or an antibody (MeDIP-seq) (Down et al. 2008), and (3) use of chemical conversion methods such as bisulfite modification of DNA (Frommer et al. 1992) either in a restricted region of the genome (RRBS) or at the whole-genome (WGBS) level. These techniques behave differently concerning CpG coverage, resolution, quantitative accuracy, efficiency, and sequencing cost (Fig. 6.3). The unique characteristics of each method provide an opportunity to choose an appropriate method best suited to answer a particular biological question while maintaining a trade-off between cost, coverage, resolution, cohort size, and a number of affordable replicates.

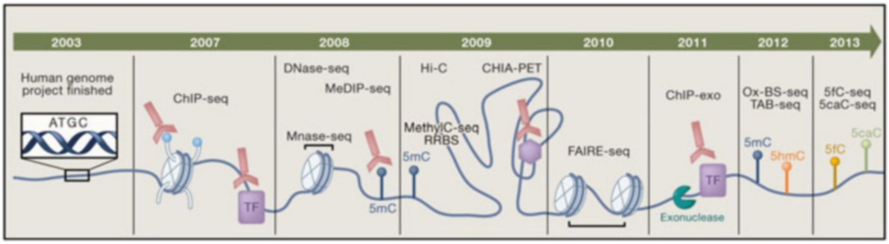


Fig. 6.2 Mapping of human epigenome: sequencing-based methods for mapping of human epigenome features in chronological order of release. It includes commonly used NGS techniques for DNA methylation, histone modification, DNA accessibility, nucleosome identification, and chromatin interaction. (Adapted from Rivera and Ren 2013)

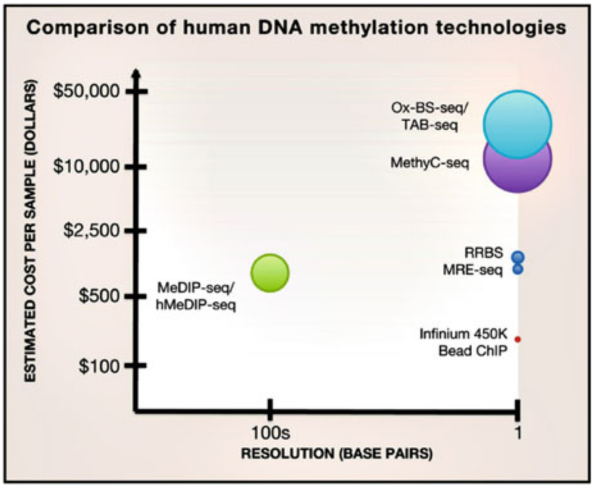


Fig. 6.3 DNA methylation techniques: an overview of existing DNA methylation methods concerning base pairs and estimated sequencing cost. (Adapted from Rivera and Ren 2013)

Despite the high sequencing cost, WGBS is essential to investigate DNA methylation changes at a single base-pair resolution systematically. A typical computational analysis pipeline of bisulfite sequencing (BS-seq) involves four basic steps: (1) quality control of the raw sequencing reads, (2) alignment against the reference genome, (3) methylation calling, and (4) identification of the differentially methylated regions. To obtain a correct alignment is a crucial and challenging step. As a result of bisulfite treatment, the complexity of the libraries is reduced, i.e., the GC content is reduced. Moreover, both strands of DNA from the reference genome must be considered separately because cytosine methylation might not be symmetric (non-CpG). Correct sequencing and mapping against the genome are of utmost importance to obtain accurate methylation state from a BS-seq experiment. A recommended workflow with the computational tools and details of each step is shown in Fig. 6.4 (Krueger et al. 2012).

Finally, differentially methylated regions (DMR) can either be annotated for the overlap with genomic regions (promoters, introns, exons, and intergenic regions) or could be associated to the gene ontology (GO) to determine enriched biological processes and molecular functions within DMRs. Computational tools (Table 6.1) identifying DMRs mainly differ regarding statistical tests used, ability to define differentially methylated regions, type of the data being analyzed, and support for covariate adjustments (Robinson et al. 2014).

6.2.2 Chromatin Immunoprecipitation

In the past decades, chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has emerged as the method of choice replacing the previously used microarray hybridization (ChIP-chip) technique to study the genome-wide binding of the proteins such as histone modifications, transcription factors (TFs),

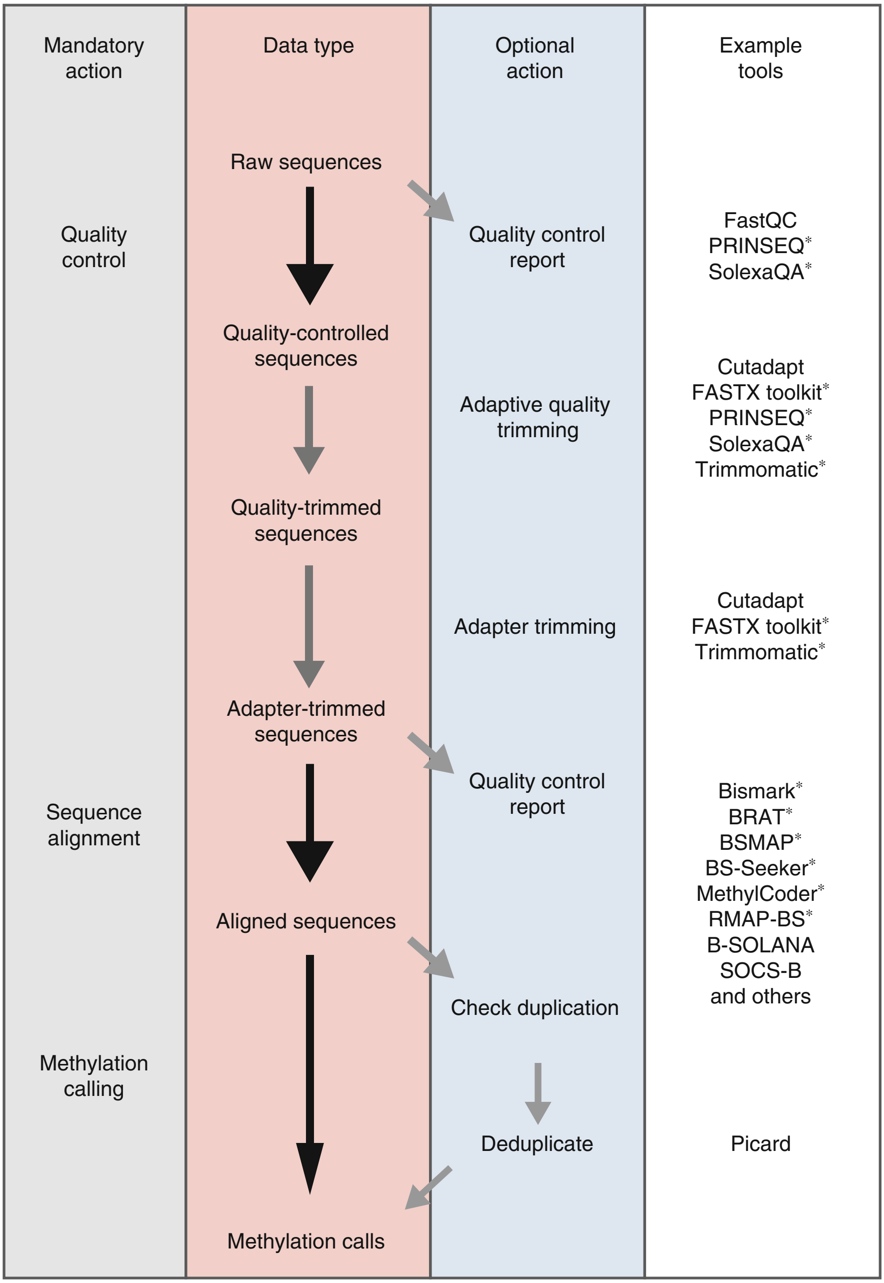


Fig. 6.4 A comprehensive workflow to analyze BS-seq data. Black and gray arrows represent necessary as well as optional steps, whereas * indicate tools dedicated to base-space data. (Adapted from Krueger et al. 2012)

Table 6.1 A list of commonly used tools to identify differentially methylated loci and regions

Method	Citation	Designed for	Determines regions or uses predefined	Accounts for covariates	Statistical elements used
Minfi	Aryee et al. (2014)	450k	Determines	Yes	Bump hunting
IMA	Wang et al. (2012)		Predefined	No	Wilcoxon
COHCAP	Warden et al. (2013)	450k or BS-seq	Predefined	Yes	FET, *t*-test, ANOVA
BSmooth	Hansen et al. (2012)	BS-seq	Determines	No	Bump hunting on smoothed t-like score
DSS	Feng et al. (2014)		Determines	No	Wald
MOABS	Sun et al. (2014)		Determines	No	Credible methylation difference
BiSeq	Hebestreit et al. (2013)		Determines	Yes	Wald
DMAP	Stockwell et al. (2014)		Predefined	Yes	ANOVA, χ^2, FET
methylKit	Akalin et al. (2012)		Predefined	Yes	Logistic regression
RADMeth	Dolzhenko and Smith (2014)		Determines	Yes	Likelihood ratio
methylSig	Park et al. (2014)		Predefined	No	Likelihood ratio
Bumphunter	Jaffe et al. (2012)	General	Determines	Yes	Permutation, smoothing
ABCD-DNA	Robinson et al. (2012)	MeDIP-seq	Predefined	Yes	Likelihood ratio
DiffBind	Ross-Innes et al. (2012)		Predefined	Yes	Likelihood ratio
M&M	Zhang et al. (2013)	MeDIP-seq + MRE-seq	Determines	No	Similar to FET

The table summarizes key features such as data type (450k or BS-seq), definition of methylated regions (predefined or dynamic), ability to perform covariate adjustment, and statistics explored by different computational methods (Robinson et al. 2014)

chromatin remodeling enzymes, and chromosomal positions of nucleosomes and polymerases (Ku et al. 2011) in a wide variety of organisms.

Briefly, the primary ChIP-seq computational pipeline includes quality control of raw sequencing reads, statistics-based enrichment of genomic regions, and functional annotation of the enriched regions as outlined in Fig. 6.5.

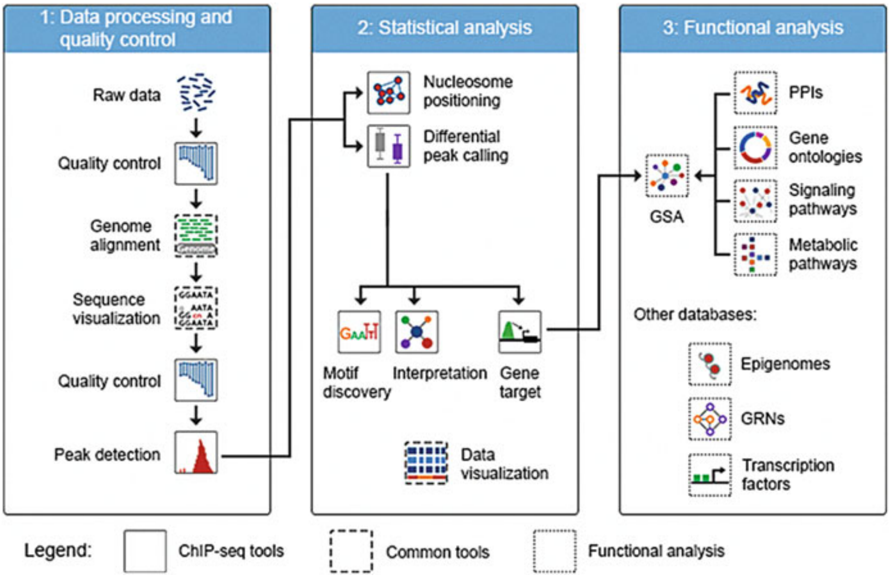


Fig. 6.5 Schematic of a typical ChIP-seq data analysis pipeline: The usual steps are divided into (*1*) data preprocessing and quality control, (*2*) statistical analysis, and (*3*) functional analysis. Each step comprises of underlying subparts as shown in the figure. (Adapted from Henry et al. 2014)

For an efficient ChIP-seq analysis, sufficient sequencing depth is essential that mainly depends on the size as well as the number of the protein binding sites, and the size of the genome itself. A minimum of 20 and 40 million uniquely mapped tags corresponding to a TF (point source) and histone (broad source) ChIP, respectively, is highly recommended in mammals. The fraction of nonredundant mapped reads (nonredundant fraction or NRF), i.e., the ratio between the number of genomic positions covered by uniquely mapped reads and the total number of uniquely mapped reads, is a useful measure of ChIP library complexity (Landt et al. 2012). Computation of the precise fragment length is crucial to determine the accurate binding sites of DNA-binding proteins. Fragment length can be calculated based on the distance either between the peaks of tag density on positive and negative strands (single-end reads) or between the pairs of the reads (paired-end sequencing).

After mapping, the location of the DNA-binding proteins (regions of ChIP enrichment) is determined using peak callers (Bailey et al. 2013) that considers different design strategies for signal smoothing and statistical background models, normalization methods (Table 6.2), and assessment of the peak quality to calculate *P*-values and false discovery rates (FDR). The final set of peaks obtained is highly relying on the algorithm used and parameter settings such that relaxed thresholds may lead to many false positives and are more likely to be noise. However, a global ChIP enrichment can be depicted using FRiP (fraction of all mapped reads within peaks) (Ji et al. 2008). A FRiP enrichment of 1% or more is recommended by ENCODE (Landt et al. 2012). Another useful metric to assess signal-to-noise ratio is

Table 6.2 A list of frequently used peak callers for the analysis of ChIP-seq

Software tool	Version	Availability	Point source (peaks)	Broad regions (domains)
BayesPeak	1.10.0	http://bioconductor.org/packages/release/bioc/html/BayesPeak.html	Yes	–
Beads[a]	1.1	http://beads.sourceforge.net/	Yes	Yes
CCAT	3	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperCCAT.htm	–	Yes
CisGenome	2	http://www.biostat.jhsph.edu/~hji/cisgenome/	Yes	–
CSAR	1.10.0	http://bioconductor.org/packages/release/bioc/html/CSAR.html	Yes	–
dPeak	0.9.9	http://www.stat.wisc.edu/~chungdon/dpeak/	Yes	–
GPS/GEM	1.3	http://cgs.csail.mit.edu/gps/	Yes	–
HPeak	2.1	http://www.sph.umich.edu/csg/qin/HPeak/	Yes	–
MACS	2.0.10	https://github.com/taoliu/MACS/	Yes	Yes
NarrowPeaks[a]	1.4.0	http://bioconductor.org/packages/release/bioc/html/NarrowPeaks.html	Yes	–
PeakAnalyzer/ PeakSplitter[a]	1.4	http://www.bioinformatics.org/peakanalyzer	Yes	–
PeakRanger	1.16	http://ranger.sourceforge.net/	Yes	Yes
PeakSeq	1.1	http://info.gersteinlab.org/PeakSeq	Yes	–
polyaPeak[a]	0.1	http://web1.sph.emory.edu/users/hwu30/polyaPeak.html	Yes	–
RSEG	0.6	http://smithlab.usc.edu/histone/rseg/	–	Yes
SICER	1.1	http://home.gwu.edu/~wpeng/Software.htm	–	Yes
SIPeS	2	http://gmdd.shgmo.org/Computational-Biology/ChIP-Seq/download/SIPeS	Yes	–
SISSRs	1.4	http://sissrs.rajajothi.com/	Yes	–
SPP	1.1	http://compbio.med.harvard.edu/Supplements/ChIP-seq/	Yes	Yes
Useq	8.5.1	http://sourceforge.net/projects/useq/	Yes	–
ZINBA	2.02.03	http://code.google.com/p/zinba/	Yes	Yes

The table lists tools for processing of enriched regions (peaks, domains, and mixed signals) identified using ChIP-seq. (Bailey et al. 2013)

[a] Only for post-processing

given by strand cross correlation, i.e., the Pearson linear correlation between the Crick and the Watson strand, after moving Watson by k base pairs (Kharchenko et al. 2008). The consistency of the peaks among the replicates is immensely useful as the most significant peaks are expected to measure same underlying biology across multiple biological replicates. Irreproducible discovery rate (IDR) is a widely accepted statistic to quantify the consistent and inconsistent groups of peaks present between the replicates (Li et al. 2011). Finally, ChIP-seq peaks are linked with functionally important genomic regions to obtain the biological inference and could be further investigated to find the consensus motifs (see details in Sect. 6.2.3).

Future challenges include the development of novel methods integrating ChIP-seq data with other NGS techniques from a statistical viewpoint to perform in-depth data analysis, for example, *how an integration of ChIP-seq and RNA-seq can deduce gene regulatory networks that might help to explain mechanisms of gene regulation.*

6.2.3 DNA Accessibility: Open Chromatin

Genome-wide maps of open chromatin sites are explored to study transcriptional regulation and footprinting of DNA-binding proteins using high-throughput sequencing: DNase-seq (Song and Crawford 2010), FAIRE-seq (Giresi et al. 2007), ATAC-seq (Buenrostro et al. 2013), and MNase-seq (Schones et al. 2008). Among all, ATAC-seq is the fastest procedure and provides sequencing libraries using the lowest amount of the cells. Moreover, a detailed application of ATAC-seq in determining TF footprinting, nucleosome positioning, as well as chromatin accessibility usually makes it a method of choice. Therefore, the computational analysis presented here is mainly concerned with ATAC-seq.

Both tag counts and tag length (currently varies between 36 and 300 bp) are the most instrumental parameters to access the sequencing quality (Fig. 6.6). Briefly, the analysis involves alignment of raw sequencing reads using aligners (Maq, RMAP, CloudBurst, SOAP, SHRiMP, BWA, and Bowtie) (Li and Homer 2010); removal of short, duplicated, or overrepresented reads using Picard (Li et al. 2009, http://broadinstitute.github.io/picard) and SAMtools (Li et al. 2009); and visualization of the signal on genome browsers (IGV; Thorvaldsdóttir et al. 2013) and UCSC genome browser (Fujita et al. 2011). A high percentage of the tags (~20–80% of the sequencing reads) usually maps to the mitochondria (Montefiori et al. 2017). Thus, it is essential to typically discard these mitochondrial reads from the downstream analysis as the open chromatin regions of interest are usually present in the nuclear DNA. From this point, the analysis can proceed further in different directions (Fig. 6.6) based on the question of interest such as determining nucleosome positioning (NucleoATAC, Schep et al. 2015) and identification of enriched open chromatin regions (F-seq, Boyle et al. 2008; Hotspot, John et al. 2011; MACS, Zhang et al. 2008; and ZINBA, Rashid et al. 2011) using peak callers.

A comprehensive set of noncoding enriched regions, obtained either using ATAC-seq or ChIP-seq, usually lacks annotation and might correspond to cis-regulatory regions such as promoters and enhancers (see below). To achieve the functional relevance, these cis-regulatory regions are commonly assigned biological meaning based on the annotation of nearby genes (GREAT; McLean et al. 2010). Thus, the peak-to-gene assignment is a highly crucial step (see Sect. 6.4). Moreover, these enriched regions (peaks) often harbor short sequences of

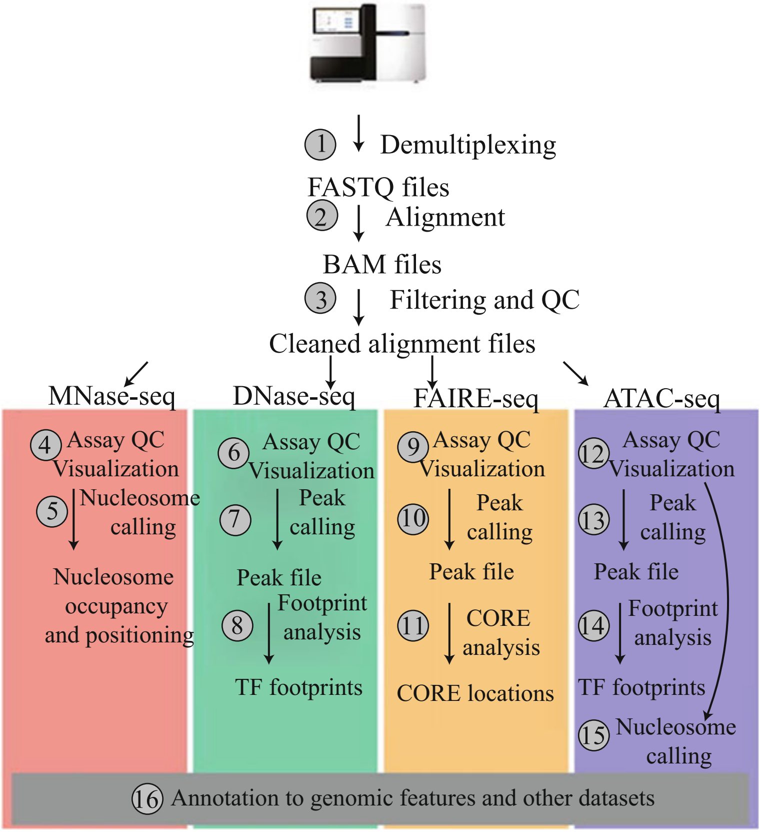


Fig. 6.6 Computational analysis of DNA accessibility: a comprehensive analytical workflow describing major steps involving the analysis of high-throughput chromatin accessibility sequencing methods. The basic alignment steps are common to all. The color boxes incorporate different steps dedicated to each method. (Adapted from Tsompana and Buck 2014)

actual binding sites of TFs that are usually found in the vicinity to the summit of enriched peaks. Identifying the conserved known/de novo motifs using MEME and FIMO (Bailey et al. 2009) and HOMER (Heinz et al. 2010) in these sequences is highly beneficial to gain fundamental insights of underlying biological mechanisms associated with these putative regulatory regions.

Taking advantage of the high sequencing depth of ATAC-seq, TF analysis can be further streamlined to identify TF footprints ((CENTIPEDE; Pique-Regi et al. 2011) and (PIQ; Sherwood et al. 2014)) that combine known information of positional weight matrices (TF motifs) with enriched open chromatin regions (DNase-seq/ ATAC-seq). The use of open chromatin data in expression QTL (eQTL) studies linking regulatory regions with disease phenotypes might help develop epigenetic and regulatory biomarkers of disease.

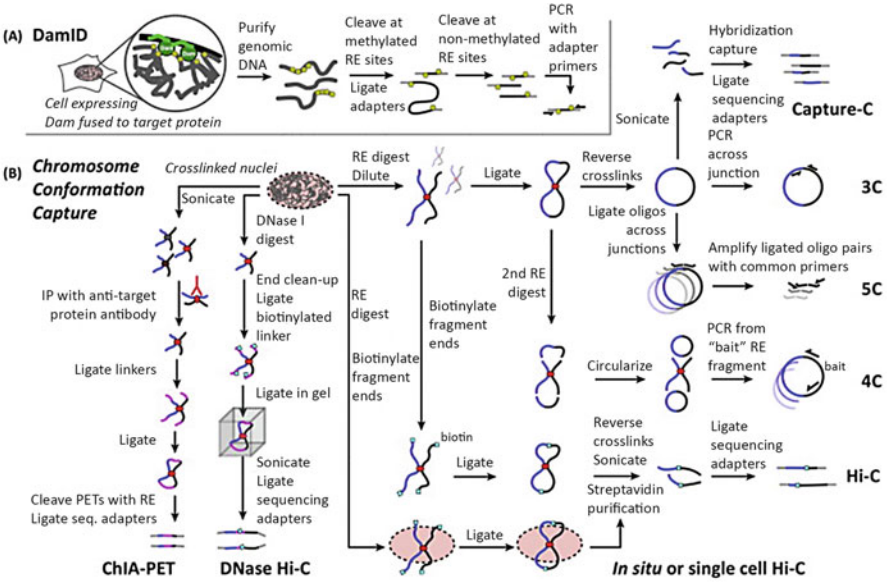

(A) DamID
Cell expressing Dam fused to target protein
Purify genomic DNA
Cleave at methylated RE sites
Ligate adapters
Cleave at non-methylated RE sites
PCR with adapter primers
(B) Chromosome Conformation Capture
Crosslinked nuclei
Sonicate
IP with anti-target protein antibody
Ligate linkers
Ligate
Cleave PETs with RE
Ligate seq. adapters
ChIA-PET
DNase I digest
End clean-up
Ligate biotinylated linker
Ligate in gel
Sonicate
Ligate sequencing adapters
DNase Hi-C
RE digest
Dilute
Ligate
RE digest
Biotinylate fragment ends
Biotinylate fragment ends
biotin
Ligate
Ligate
2nd RE digest
Reverse crosslinks
Sonicate
Hybridization capture
Ligate sequencing adapters
Capture-C
PCR across junction
3C
Ligate oligos across junctions
Amplify ligated oligo pairs with common primers
5C
Circularize
PCR from "bait" RE fragment
bait
4C
Reverse crosslinks
Sonicate
Streptavidin purification
Ligate sequencing adapters
Hi-C
In situ or single cell Hi-C

Fig. 6.7 Chromatin conformation capture methods: a summary of the experimental workflow of conformation capture methods for the analysis of 3D chromatin organization within a cell (Risca and Greenleaf 2015). These techniques have different approaches for fragmentation, enrichment, and detection of the ligation junctions, thereby giving rise to a diverse set of methods. (Adapted from Risca and Greenleaf 2015)

6.2.4 Chromatin Conformation

Chromosome conformation capture (3C)-based techniques have been explored to understand the physical wiring diagram of the genome (Fig. 6.7). The target-based methods (3C and 4C), though they provide interaction profiles corresponding to individual loci, are limited to the prior knowledge of the possible interactions with the targets, while ChIA-PET, 5C, and Hi-C provide a comprehensive, genome-wide, and unbiased portrait of the regulatory chromatin looping, thereby potentially revealing all interactions between promoters and enhancers (de Wit and de Laat 2012). However, a close spatial proximity might be due to a direct or indirect and specific or nonspecific (random) interaction between a pair of loci. Interestingly, a graphical representation of a high-resolution interaction map reveals the presence of highly self-interacting topologically associating domains (TAD; Dekker et al. 2013), where loci lying within the domains interact more often with limited interaction with regions in other domains.

Computational approaches to analyze chromatin interaction involve either chromatin interaction restraint-based modeling or polymer ensemble-based approach. Specifically, restraint-based modeling has proven informative and efficient for analyzing stable chromosomal looped domains. Instead, a polymer-based method is robust to determine statistical organizational features of chromosomal folding, their dynamics and variability across different cells, and specific interactions between chromosomes (Dekker et al. 2013). With the maturation of computational epigenomics, various methods have recently been implemented to process Hi-C data. A recent benchmark study of commonly used Hi-C methods including HiCCUPS, TADbit, HIPPIE, and HiCseg has discussed key features of each tool in detail that differs with respect to usability, computing resources required to finish the analysis, type of interactions given (*cis* or *trans*), and alignment, filtering, and normalization of contact matrix strategies (Forcato et al. 2017).

These 3D representations provide a way to leverage information about the higher-order chromosomal organization such as the formation of globular domains, chromosome territories, and mechanism involved in folding (Dekker et al. 2013) as well as interactions between REs. For instance, in a particular case of *Hox* genes, i.e., the primary determinants of the animal body plan, a landscape of the two topological domains (TADs) flanking a *HoxD* cluster illustrates the modular organization containing distinct regulatory capacities around the HoxD cluster (Andrey et al. 2013). The transition of *Hox* genes between separate regulatory landscapes is a useful paradigm to interpret the functional rationale underlying this organization as well as its evolutionary origin.

6.3 Chromatin Dynamics Mediated by Regulatory Elements

Apart from the well-understood 1.5% coding region, various attempts have been made to understand the function of the noncoding portion of the human genome over the years. The noncoding part of the genome harbors REs containing binding sites for various DNA-binding proteins such as TFs and cofactors that work in synchrony at these REs to regulate the transcription of the eukaryotic genes (Sheffield and Furey 2012). However, the accessibility of the binding site of TFs within REs might be affected because of the epigenetic factors such as chromatin remodelers, histone modifications, and DNA methylation. The dynamism of specific epigenomic factors at REs is instrumental in characterizing distinct regulatory states within the genome (see the Sect. 6.4), since gene expression is regulated by affecting either the accessibility or the affinity of TFs at the regulatory sites. Therefore, REs act as decision-making entities (Fig. 6.8) governing the cell's response to stimuli from surrounding cells and the environment (Sheffield and Furey 2012).

The distinct classes of REs typically include silencers, promoters, enhancers, insulators, and locus control regions (LCRs), each performing specific functions (Maston et al. 2006; Dunham et al. 2012; Heintzman et al. 2009). The landscape of promoters and enhancers is better characterized than the other classes. In contrast to the promoters that lie immediate upstream to the transcriptional start site of their gene targets, enhancers can regulate target gene expression irrespective of their location (either distal or proximal to the target genes) and orientation within the genome. These enhancers may interact with promoters through looping mechanism (Wang et al. 2005). A repertoire of all REs in the entire genome constitutes the *regulome* of an organism.

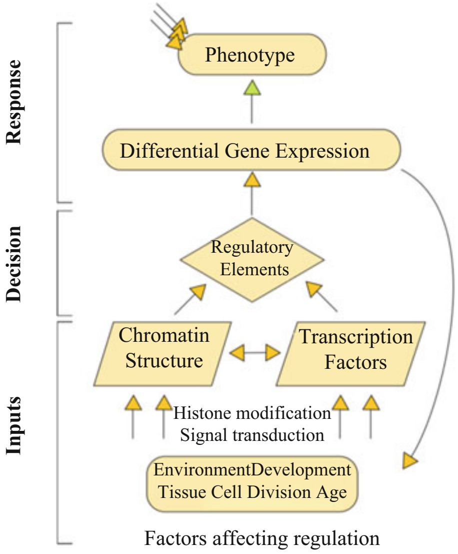


Fig. 6.8 Transcriptional regulation by regulatory elements: a flowchart showing transcriptional regulation by regulatory elements. Surrounding factors control the availability of regulatory elements via altering chromatin structure and transcription factor binding. RE modulates gene expression to yield distinct phenotype. (Adapted from Sheffield and Furey 2012)

6.4 Challenges and Opportunities to Study Regulatory Elements, Particularly Enhancers

Elucidating the functional meaning of noncoding part of the genome is challenging. Unlike a few thousand well-conserved protein-coding genes, the proposed number of REs has grown to millions and seems to lack high conservation even between genetically closely related species. Intriguingly, REs can modulate target gene expression by either a direct interaction or indirectly via action on TFs. Moreover, the regulome is dynamic and varies with factors such as age, genetic background, developmental time, cell/tissue type, and the environment. Apart from this, each class of RE harbors distinct functional states that may perform a variety of specific functions. There are certain histone modifications governing rules to determine operational states of enhancers. For instance, the most common histone acetylations (H3K27 and H3K9) and H3 monomethylated at lysine 4 (H3K4me1) enrich at *active* enhancers. Additionally, active H3K4me1 along with the repressive polycomb protein-associated H3K27me3 marks *closed or poised* state of the enhancer (Shlyueva et al. 2014) and so on. Overall, multiple layers of complexity in RE-mediated mechanisms pose an interesting challenge to fully understand the robust system of gene regulation mediated by REs.

In spite of the challenges mentioned above, a rapid advancement in DNA-protein interaction techniques and genome-wide epigenomic interaction studies has provided new functional insights into RE-mediated mechanisms. The unprecedented availability of the genome-wide epigenomic data has been instrumental in identifying regulatory regions, especially enhancers that are located anywhere in the genome unlike promoters lying immediate upstream to the target genes. Existing computational endeavors have typically explored the dynamic patterns of histone posttranslational modifications (histone-code hypothesis) integrating DNA accessible regions using sophisticated machine-learning (ML) approaches (Table 6.3) to identify either distinct regulatory states (ChromHMM) or putative enhancers (RFECS) in particular. Also, sequence-based features (Fig. 6.9) such as CpG islands, GC content, sequence conservation, and TF binding sites assist in identifying enhancers (Whitaker et al. 2015) such as in DEEP (Kleftogiannis et al. 2015). Moreover, transcription of enhancers in the form of noncoding enhancer RNA (eRNA) has increasingly exploited by in silico methods as an important feature to detect functionally active enhancers. As enhancers share common structural and functional characteristics with promoters, few epigenetic modifications coexist at different REs, thereby showing dynamic and ambiguous epigenomic patterns (Weingarten-Gabbay and Segal 2014).

Due to the peculiar characteristics of enhancers such as their distal location from the target genes, high cell-type/tissue specificity, expression in the form of eRNA, and ambiguous epigenetic modification patterns, computational methods are still struggling to achieve accurate enhancer predictions in a cell-type/tissue-specific manner. Further, due to the complex mechanism of the gene regulation mediated by enhancers, a precise mapping to their gene targets is not so straightforward. Beyond experimental approaches elucidating enhancer-target relationship, a couple of widely accepted enhancer-mediated gene regulation hypotheses include *distance-based* approach to associate enhancer state with the most proximal target gene expression or promoters and *chromatin structure-based* approach to map enhancer and promoter interactions through looping mechanism. Other factors such as eQTL information, eRNA co-expression, and TF co-expression might improve gene-enhancer association (Fishilevich et al. 2017).

6.5 Computational Advancements to Elucidate Effect of Regulatory Variants in Disease Pathogenicity

Both loss- and gain-of-function mutations associated with cis-regulatory regions have potential to generate transcriptional alterations, thereby causing a gradient of phenotypes. A huge portion of the disease associated variants (GWAS: 88%) falls into the noncoding regions that might be causal and contribute to either an altered or a diseased phenotype (Hindorff et al. 2009). The exact mechanisms that lead to these altered phenotypes are not yet elucidated in the majority of cases. Noncoding variations in active enhancers constitute one of the major groups responsible for functional interpretation of GWAS loci that might lead to the changes in the expression of target genes (Degner et al. 2012). The binding site of DNA-binding proteins can be considered to drive on/off switches for gene transcription, which, if mutated by noncoding variations, might be associated with human diseases and evolution (Fig. 6.10).

Unlike protein-coding variants, the amount and frequency of intergenic variants are high. Moreover, only a small fraction of the plethora of noncoding variants is

Table 6.3 A list of the commonly used machine-learning-based tools for the in silico identification of enhancers

	Supervised methods					Unsupervised methods	
Tool	CSI ANN	ChromaGenSVM	RFECS	ChroModule	EnhancerFinder	ChromHMM	Segway
Reference	Firpi et al. (2010)	Fernandez and Miranda-Saavedra (2012)	Rajagopalan et al. (2013)	Won et al. (2013)	Erwin et al. (2014)	Ernst and Kellis (2012)	Hoffman et al. (2012)
Method	Time delay NN	SVM	Random forest	HMM	Multiple kernel learning	HMM	Dynamic Bayesian network (DBN)
Features	Histone marks		P300, histone marks	Histone marks and chromatin accessibility	Epigenomic data, PhastCons scores and 4-mer occurrences	Histone marks, CBP, DHS	Histone marks
Validation	Computational: overlap with known RE genomic and epigenomic features					Computational: overlap with known RE genomic and epigenomic features Experimental: luciferase assay	

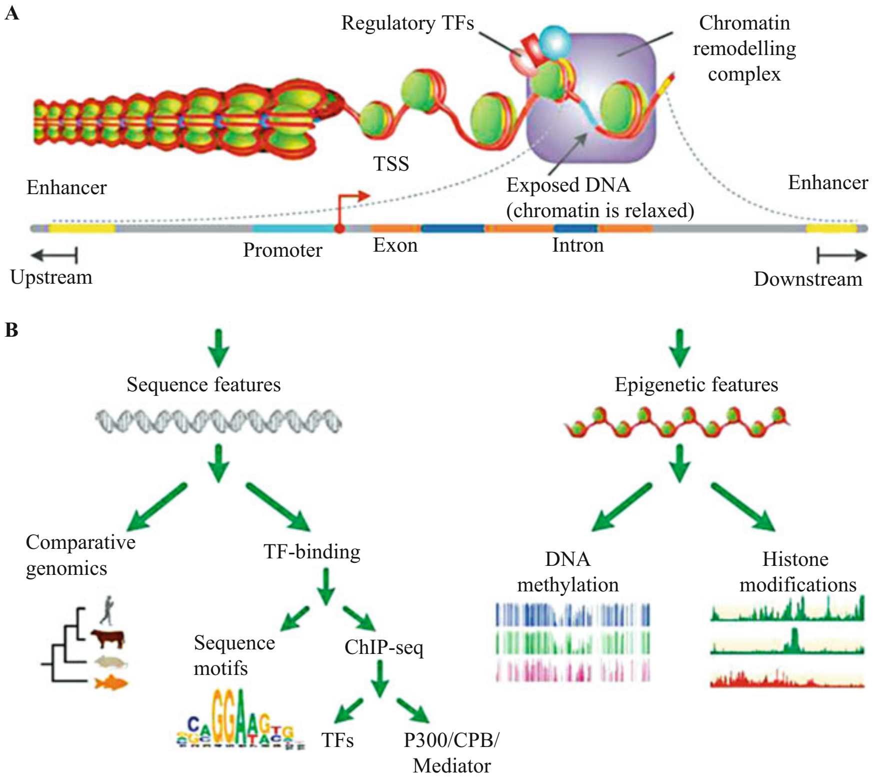


Fig. 6.9 Enhancers – working mechanism and relevant characteristics. (**a**) An overview of the transcriptional regulatory event that includes regulatory TFs, chromatin remodeling complex (coactivators), and histone acetyltransferases (HATs). (**b**) Epigenomic and genomic features relevant for computational prediction of cis-regulatory elements. Features associated with DNA sequence are pertinent to identify the TF binding and conservation across species, whereas epigenetic features determine chromatin structure modifications. Thus, the epigenomic along with the sequence-based signatures are very well exploited by unsupervised and supervised machine-learning (ML)-based approaches. (Adapted from Whitaker et al. 2015)

deleterious. Hence, one of the main challenges is to prioritize these causal cis-regulatory variations disrupting the functional regulatory machinery and subsequently uncover their molecular mechanisms in gene regulation. Several tools including GWAS3D (Li et al. 2013), FunSeq (Fu et al. 2014), regSNP (Teng et al. 2012), and RAVEN (Andersen et al. 2008), resources like RegulomeDB (Boyle et al. 2012), and scoring (GWAVA (Kircher et al. 2014) and CADD (Ritchie et al. 2014)) mechanisms have been developed to assess and prioritize deleterious effects of regulatory variants. To establish the relationships between genotype and expression, regulatory QTLs such as eQTLs, sQTLs (splicing QTL), and cis-eQTLs have been comprehensively studied concerning diseases.

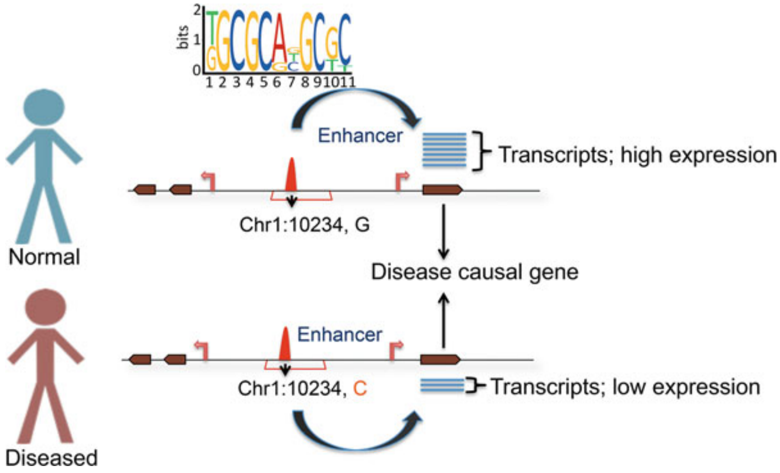


Fig. 6.10 A schematic representation of the potential role of regulatory variants in disease pathogenicity. A point mutation (from G to C) in the enhancer region co-localizing with the transcription factor binding site is causing differential expression of a gene causing pathogenicity

6.6 Available Epigenomic Resources

Post-completion of the human reference genome in 2002 (Human Genome Sequencing Consortium 2004), the analysis of the epigenome was enabled by the introduction of several new high-throughput sequencing protocols. With the aim to identify, catalog, and interpret epigenomic mechanisms, several large initiatives and dedicated academic centers were established, which collectively improve our knowledge of associated epigenetic mechanisms across diverse cell types and species (Table 6.4). These consortia have generated unprecedented, complex, and comprehensive genome-wide epigenomic maps over the past decades. These expanding bodies of chromatin data in the public domain have fostered the development of databases and epigenomic repertoires to fulfill the pressing need for organization of raw and processed epigenomic data in a cordial manner.

Commonly used data resources include 4DGenome, 3CDB, Histone, MethylomeDB, DiseaseMeth, NGSmethDB, MethBase, and MpromDB. Moreover, TF binding motif collections of hundreds of transcription factors such as JASPAR, PreDREM, HOCOMOCO, and TRANSFAC enable systematic analysis of DNA-binding proteins.

Over the years, our model of the epigenetic landscape has become increasingly complex. With the increasing knowledge of molecular mechanisms, epigenetics currently provides a comprehensive global perspective on its role in fundamental cellular processes as well as in causing disorders. Epigenetic mechanisms are tightly interwoven that cooperatively control the wrapping of the DNA for regulation of target genes in a complex network of synergistic and antagonistic interactions.

Table 6.4 Large-scale projects for epigenomic studies

Project name	Start date	Affiliations	Completed and expected data contributions	Selected publication	Access data
Encyclopedia of DNA Elements (ENCODE)	2003	NIH	DNase-seq, RNA-seq, ChIP-seq, and 5C in hundreds of primary human tissues and cell lines	The ENCODE Project Consortium (2012)	http://encodeproject.org/
The Cancer Genome Atlas (TCGA)	2006		DNA methylomes in thousands of patient samples from more than 20 cancer types	Garraway and Lander (2013)	http://cancergenome.nih.gov/
Roadmap Epigenomics Project	2008		DNase-seq, RNA-seq, ChIP-seq, and MethylC-seq in hundreds of normal primary cells, hESC, and hESC-derived cells	Bernstein et al. (2010)	http://www.epigenomebrowser.org/
International Cancer Genome Consortium (ICGC)	2008	15 countries, includes TCGA	DNA methylation profiles in thousands of patient samples from 50 different cancers	The International Cancer Genome Consortium (2010)	http://dcc.icgc.org/
International Human Epigenome Consortium (IHEC)	2010	7 countries, includes BLUEPRINT, roadmap	Goal: 1000 epigenomes in 250 cell types	American Association for Cancer Research Human Epigenome Task Force, European Union, Network of Excellence, Scientific Advisory Board (2008)	http://ihec-epigenomes.org
FANTOM	2000	RIKEN	CAGE data, promoter atlas, and enhancer atlas of more than 800 primary cells, cancer lines, and tissues	The FANTOM Consortium et al. (2014) and Andersson et al. (2014)	fantom.gsc.riken.jp

Untangling this network, both mechanistically and statistically, and linking the outcomes to disease and development are central goals of the modern epigenetics. Current NGS methods have revolutionized our capacities to decipher different building units of the dynamic chromatin such as nucleosomes, open chromatin regions, and DNA-binding proteins (histones and TF).

Taking advantage of existing unprecedented volume of NGS data, novel hypotheses and computational methods have emerged and are still developing. This provides a unique opportunity to integrate various techniques, further elucidate chromatin dynamics and foster new hypotheses uncovering molecular mechanism of gene regulation. With an ultimate goal to uncover the complicated relationship of in vivo system, a new frontier of whole-genome analysis is the amalgamation of data from several thousands of experiments. Upon integration of genomic and epigenomic data including histone modification, chromatin accessibility, nucleosome dynamics, RNA expression, transcription factor binding, and sequence-based genome annotation, a comprehensive view of chromatin structure and function can be obtained that enables the discovery of the complex underlying mechanism of gene regulation. Recently, new experimental and computational methods have emerged integrating single-cell DNA methylation with single-cell RNA-seq data (Clark et al. 2016) overcoming the issues with cell-type composition in complex tissues and batch effects between epigenomic and transcriptomic analyses. Overall, recent studies and methods in the field of regulatory genomics have remarkably improved our ability to identify variants altering gene expression. However, the identification of variants creating de novo active transcriptional sites remains an unsolved issue. Moreover, a complete understanding of regulatory mechanisms is only feasible if tissue-specific, high-resolution information on 3D interactions of REs is taken into account, a type of data that is available for a small number of cell lines and might be more readily accessible in the near future.

In summary, given our still limited knowledge, the past decades have witnessed remarkable progresses in our understanding of mechanisms of gene regulation. Combined efforts of both experimental and computational research have made such great advancements possible. The number of bioinformatics tools developed within the last few years reflects the growing interest in this field. The diversity of the proposed strategies has highlighted their advantages, challenges, and weaknesses. These next-generation sequencing methods have been instrumental to decipher important biological questions in the field of gene regulation. Moreover, a continuous expansion of experimental and computational methods and the availability of benchmark datasets are likely to further improve in silico enhancer prediction as well as enhancer-target gene linkage.

Acknowledgments I thank Ajinkya Deogade and Sachin Pundhir for their useful comments while writing this book chapter. I appreciate *La Caixa international Ph.D. scholarship program* at the Centre for Genomic Regulation (CRG), Barcelona, Spain, for providing financial support.

References

Akalin A, Kormaksson M, Li S et al (2012) MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13:R87

American Association for Cancer Research Human Epigenome Task Force, European Union, Network of Excellence, Scientific Advisory Board (2008) Moving AHEAD with an international human epigenome project. Nature 454:711–715. https://doi.org/10.1038/454711a

Andersen MC, Engström PG, Lithwick S et al (2008) In Silico detection of sequence variations modifying transcriptional regulation. PLoS Comput Biol 4:e5. https://doi.org/10.1371/journal.pcbi.0040005

Andersson R, Gebhard C, Miguel-Escalada I et al (2014) An atlas of active enhancers across human cell types and tissues. Nature 507:455–461. https://doi.org/10.1038/nature12787

Andrey G, Montavon T, Mascrez B et al (2013) A switch between topological domains underlies HoxD genes collinearity in mouse limbs

Aryee MJ, Jaffe AE, Corrada-Bravo H et al (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30:1363–1369. https://doi.org/10.1093/bioinformatics/btu049

Bailey TL, Boden M, Buske FA et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208. https://doi.org/10.1093/nar/gkp335

Bailey T, Krajewski P, Ladunga I et al (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol 9:e1003326. https://doi.org/10.1371/journal.pcbi.1003326

Bernstein BE, Stamatoyannopoulos JA, Costello JF et al (2010) The NIH roadmap Epigenomics mapping consortium. Nat Biotechnol 28:1045–1048. https://doi.org/10.1038/nbt1010-1045

Bock C (2012) Analysing and interpreting DNA methylation data. Nat Rev Genet 13:705–719

Boyle AP, Guinney J, Crawford GE, Furey TS (2008) F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics 24:2537–2538. https://doi.org/10.1093/bioinformatics/btn480

Boyle AP, Hong EL, Hariharan M et al (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22:1790–1797. https://doi.org/10.1101/gr.137323.112

Buenrostro JD, Giresi PG, Zaba LC et al (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10:1213–1218. https://doi.org/10.1038/nmeth.2688

Choi JK (2013) “Open” chromatin: histone acetylation, linker histones & histone variants. https://doi.org/10.14288/1.0165590

Clark SJ, Lee HJ, Smallwood SA et al (2016) Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. Genome Biol 17:72. https://doi.org/10.1186/s13059-016-0944-x

de Wit E, de Laat W (2012) A decade of 3C technologies: insights into nuclear organization. Genes Dev 26:11–24. https://doi.org/10.1101/gad.179804.111

Degner JF, Pai AA, Pique-Regi R et al (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482:390–394. https://doi.org/10.1038/nature10808

Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14:390–403. https://doi.org/10.1038/nrg3454

Dolzhenko E, Smith AD (2014) Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinf 15:215. https://doi.org/10.1186/1471-2105-15-215

Down TA, Rakyan VK, Turner DJ et al (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 26:779–785. https://doi.org/10.1038/nbt1414

Dunham I, Kundaje A, Aldred SF et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. https://doi.org/10.1038/nature11247

Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9:215–216. https://doi.org/10.1038/nmeth.1906

Erwin GD, Oksenberg N, Truty RM et al (2014) Integrating diverse datasets improves developmental enhancer prediction. PLoS Comput Biol 10:e1003677. https://doi.org/10.1371/journal.pcbi.1003677

FANTOM Consortium and the RIKEN PMI and CLST (DGT), ARR F, Kawaji H et al (2014) A promoter-level mammalian expression atlas. Nature 507:462–470. https://doi.org/10.1038/nature13182

Felsenfeld G (2014) A brief history of epigenetics. Cold Spring Harb Perspect Biol. https://doi.org/10.1101/cshperspect.a018200

Feng H, Conneely KN, Wu H (2014) A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res 42: e69–e69. https://doi.org/10.1093/nar/gku154

Fernández M, Miranda-Saavedra D (2012) Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic Acids Res 40: e77. https://doi.org/10.1093/nar/gks149

Firpi HA, Ucar D, Tan K (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics 26:1579–1586. https://doi.org/10.1093/bioinformatics/btq248

Fishilevich S, Nudel R, Rappaport N et al (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford). https://doi.org/10.1093/database/bax028

Forcato M, Nicoletti C, Pal K et al (2017) Comparison of computational methods for hi-C data analysis. https://doi.org/10.1038/nmeth.4325

Frommer M, McDonald LE, Millar DS et al (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 89:1827–1831. https://doi.org/10.1073/PNAS.89.5.1827

Fu Y, Liu Z, Lou S et al (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol 15:480. https://doi.org/10.1186/s13059-014-0480-5

Fujita PA, Rhead B, Zweig AS et al (2011) The UCSC genome browser database: update 2011. Nucleic Acids Res 39:D876–D882. https://doi.org/10.1093/nar/gkq963

Garraway LA, Lander ES (2013) Lessons from the Cancer genome. Cell 153:17–37. https://doi.org/10.1016/J.CELL.2013.03.002

Giresi PG, Kim J, McDaniell RM et al (2007) FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. Genome Res 17:877–885. https://doi.org/10.1101/gr.5533506

Hansen KD, Langmead B, Irizarry RA (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol 13:R83

Hebestreit K, Dugas M, Klein H-U (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics 29:1647–1653. https://doi.org/10.1093/bioinformatics/btt263

Heintzman ND, Hon GC, Hawkins RD et al (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459:108–112. https://doi.org/10.1038/nature07829

Heinz S, Benner C, Spann N et al (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38:576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Henry VJ, Bandrowski AE, Pepin A-S et al (2014) OMICtools: an informative directory for multi-omic data analysis. Database (Oxford). https://doi.org/10.1093/database/bau069

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 U S A 106:9362–9367. https://doi.org/10.1073/pnas.0903103106

Hoffman MM, Buske OJ, Wang J et al (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods 9:473–476. https://doi.org/10.1038/nmeth.1937

Hotchkiss RD (1948) The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. J Biol Chem 175:315–332

Hudson TJ, Anderson W, Aretz A et al (2010) International network of cancer genome projects. Nature 464:993–998. https://doi.org/10.1038/nature08987

Human Genome Sequencing Consortium I (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945. https://doi.org/10.1038/nature03001

Jaffe AE, Murakami P, Lee H et al (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol 41:200–209. https://doi.org/10.1093/ije/dyr238

Ji H, Jiang H, Ma W et al (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26:1293–1300. https://doi.org/10.1038/nbt.1505

John S, Sabo PJ, Thurman RE et al (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet 43:264–268. https://doi.org/10.1038/ng.759

Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26:1351–1359. https://doi.org/10.1038/nbt.1508

Kircher M, Witten DM, Jain P et al (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315. https://doi.org/10.1038/ng.2892

Kleftogiannis D, Kalnis P, Bajic VB (2015) DEEP: a general computational framework for predicting enhancers. Nucleic Acids Res 43:e6. https://doi.org/10.1093/nar/gku1058

Krueger F, Kreck B, Franke A, Andrews SR (2012) DNA methylome analysis using short bisulfite sequencing data. Nat Methods 9:145–151. https://doi.org/10.1038/nmeth.1828

Ku CS, Naidoo N, Wu M, Soong R (2011) Studying the epigenome using next generation sequencing. J Med Genet 48:721–730. https://doi.org/10.1136/jmedgenet-2011-100242

Landt SG, Marinov GK, Kundaje A et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813–1831. https://doi.org/10.1101/gr.136184.111

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–483. https://doi.org/10.1093/bib/bbq015

Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. Ann Appl Stat 5:1752–1779. https://doi.org/10.1214/11-AOAS466

Li MJ, Wang LY, Xia Z et al (2013) GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. Nucleic Acids Res 41:W150–W158. https://doi.org/10.1093/nar/gkt456

Martens JHA, Stunnenberg HG (2013) BLUEPRINT: mapping human blood cell epigenomes. Haematologica 98:1487–1489. https://doi.org/10.3324/haematol.2013.094243

Marx V (2012) READING THE SECOND GENOMIC CODE

Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7:29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623

Maunakea AK, Nagarajan RP, Bilenky M et al (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466:253–257. https://doi.org/10.1038/nature09165

McLean CY, Bristor D, Hiller M et al (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28:495–501. https://doi.org/10.1038/nbt.1630

Montefiori L, Hernandez L, Zhang Z et al (2017) Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. Sci Rep 7:2451. https://doi.org/10.1038/s41598-017-02547-w

Park Y, Figueroa ME, Rozek LS, Sartor MA (2014) MethylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics 30:2414–2422. https://doi.org/10.1093/bioinformatics/btu339

Pique-Regi R, Degner JF, Pai AA et al (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res 21:447–455. https://doi.org/10.1101/gr.112623.110

Plank JL, Dean A (2014) Enhancer function: mechanistic and genome-wide insights come together. Mol Cell 55:5–14. https://doi.org/10.1016/j.molcel.2014.06.015

Rajagopal N, Xie W, Li Y et al (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput Biol 9:e1002968. https://doi.org/10.1371/journal.pcbi.1002968

Rashid NU, Giresi PG, Ibrahim JG et al (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol 12:R67. https://doi.org/10.1186/gb-2011-12-7-r67

Risca VI, Greenleaf WJ (2015) Unraveling the 3D genome: genomics tools for multiscale exploration. Trends Genet 31:357–372. https://doi.org/10.1016/j.tig.2015.03.010

Ritchie GRS, Dunham I, Zeggini E, Flicek P (2014) Functional annotation of noncoding sequence variants. Nat Methods 11:294–296. https://doi.org/10.1038/nmeth.2832

Rivera CM, Ren B (2013) Mapping human epigenomes. Cell 155:39–55. https://doi.org/10.1016/j.cell.2013.09.011

Robertson KD (2005) DNA methylation and human disease. Nat Rev Genet 6:597–610. https://doi.org/10.1038/nrg1655

Robinson MD, Strbenac D, Stirzaker C et al (2012) Copy-number-aware differential analysis of quantitative DNA sequencing data. Genome Res 22:2489–2496. https://doi.org/10.1101/gr.139055.112

Robinson MD, Kahraman A, Law CW et al (2014) Statistical methods for detecting differentially methylated loci and regions. Front Genet 5:324. https://doi.org/10.3389/fgene.2014.00324

Ross-Innes CS, Stark R, Teschendorff AE et al (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 481:389–393. https://doi.org/10.1038/nature10730

Schep A, Buenrostro JD, Denny SK et al (2015) Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. bioR xiv 16642. doi: https://doi.org/10.1101/016642

Schones DE, Cui K, Cuddapah S et al (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132:887–898. https://doi.org/10.1016/j.cell.2008.02.022

Sheffield NC, Furey TS (2012) Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. Genes (Basel) 3:651–670. https://doi.org/10.3390/genes3040651

Sherwood RI, Hashimoto T, O'Donnell CW et al (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol 32:171–178. https://doi.org/10.1038/nbt.2798

Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15:272–286. https://doi.org/10.1038/nrg3682

Song L, Crawford GE (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 2010: pdb.prot5384. https://doi.org/10.1101/pdb.prot5384

Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 403:41–45. https://doi.org/10.1038/47412

Stockwell PA, Chatterjee A, Rodger EJ, Morison IM (2014) DMAP: differential methylation analysis package for RRBS and WGBS data. Bioinformatics 30:1814–1822. https://doi.org/10.1093/bioinformatics/btu126

Sun D, Xi Y, Rodriguez B et al (2014) MOABS: model based analysis of bisulfite sequencing data. Genome Biol 15:R38. https://doi.org/10.1186/gb-2014-15-2-r38

Tan M, Luo H, Lee S et al (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell 146:1016–1028. https://doi.org/10.1016/j.cell.2011.08.008

Teng M, Ichikawa S, Padgett LR et al (2012) regSNPs: a strategy for prioritizing regulatory single nucleotide substitutions. Bioinformatics 28:1879–1886. https://doi.org/10.1093/bioinformatics/bts275

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192. https://doi.org/10.1093/bib/bbs017

Tsompana M, Buck MJ (2014) Chromatin accessibility: a window into the genome. Epigenetics Chromatin 7:33. https://doi.org/10.1186/1756-8935-7-33

Waddington CH (2012) The epigenotype. 1942. Int J Epidemiol 41:10–13. https://doi.org/10.1093/ije/dyr184

Wang Q, Carroll JS, Brown M (2005) Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. Mol Cell 19:631–642. https://doi.org/10.1016/j.molcel.2005.07.018

Wang D, Yan L, Hu Q et al (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics 28:729–730. https://doi.org/10.1093/bioinformatics/bts013

Warden CD, Lee H, Tompkins JD et al (2013) COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res 41:e117–e117. https://doi.org/10.1093/nar/gkt242

Weingarten-Gabbay S, Segal E (2014) A shared architecture for promoters and enhancers. Nat Genet 46:1253–1254. https://doi.org/10.1038/ng.3152

Whitaker JW, Nguyen TT, Zhu Y et al (2015) Computational schemes for the prediction and annotation of enhancers from epigenomic assays. Methods 72:86–94. https://doi.org/10.1016/j.ymeth.2014.10.008

Wu C, Morris JR (2001) Genes, genetics, and epigenetics: a correspondence. Science 293:1103–1105. https://doi.org/10.1126/science.293.5532.1103

Won K-J, Zhang X, Wang T et al (2013) Comparative annotation of functional regions in the human genome using epigenomic data. Nucleic Acids Res 41:4423–4432. https://doi.org/10.1093/nar/gkt143

Xu F, Zhang K, Grunstein M et al (2005) Acetylation in histone H3 globular domain regulates gene expression in yeast. Cell 121:375–385. https://doi.org/10.1016/j.cell.2005.03.011

Zhang HZH, Fiume E, Ms OC (2002) shape matching of 3D contours using normalized Fourier descriptors. Proc SMI Shape Model Int 2002:3–6. https://doi.org/10.1109/SMI.2002.1003554

Zhang Y, Liu T, Meyer CA et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhang B, Zhou Y, Lin N et al (2013) Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome Res 23:1522–1540. https://doi.org/10.1101/gr.156539.113

Data Mining to Detect Common, Unique, and Polymorphic Simple Sequence Repeats 7

Aditi Kapil, C. K. Jha, and Asheesh Shanker

7.1 Introduction

Data mining is at its simplest a process (manual or automated) to extract information from large amount of data and transform it into an understandable pattern for further use. Mined information can be stored and made available through databases so that it can be easily managed and accessed. Data mining plays crucial role in bioinformatics which is an interdisciplinary field with an aim to develop different methods and software tools to store, analyze, and further utilize the stored biological information for various purposes. Incorporating computational data mining in bioinformatics helps in better organization of huge biological data and making it available as a data source for other scientific research. Many data mining tools (online and offline) are available for examining both sequential and structural forms of biological data, and a large number of specialized online databases are also present. They offer very interactive features to easily retrieve and reuse the required information of interest. In the present chapter, computational data mining of simple sequence repeats (SSRs) has been discussed.

Simple sequence repeats, also known as microsatellites, are repetitive DNA sequences of 1–6 nucleotides (mono–hexa) that are repeated tandemly many times at a locus (Vogt 1990), for example:

A. Kapil
Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan, India

C. K. Jha
Department of Computer Science, Banasthali Vidyapith, Rajasthan, India

A. Shanker (✉)
Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan, India

Department of Bioinformatics, Central University of South Bihar, Gaya, Bihar, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_7

Mononucleotide	$(G)_{12}$	GGGGGGGGGGGG
Dinucleotide	$(AT)_6$	ATATATATATAT
Trinucleotide	$(ATG)_4$	ATGATGATGATG
Tetranucleotide	$(CTGA)_3$	CTGACTGACTGA
Pentanucleotide	$(GAAT)_3$	GAATGAATGAAT
Hexanucleotide	$(CGAATG)_3$	CGAATGCGAATGCGAATG

Depending upon the arrangement of repeat motifs, SSRs can be perfect, compound, or imperfect (Bachmann and Bare 2004), for example:

Perfect SSR (uninterrupted repeat)	GTGTGTGTGTGTGTGT
Compound SSR (contains >1 repeat motifs)	$(ATG)_n(GC)_n$
Imperfect SSR (repeat with atleast a single base interruption)	ACACACACA**T**ACACACAC

Slippage mechanism during the DNA replication or errors in repair or recombination processes have been considered to create these repeats (Tautz and Schlotterer 1994; Levinson and Gutman 1987). SSRs are found in the genome sequences of eukaryotic as well as prokaryotic organisms (Field and Wills 1996) and are ubiquitously present in both noncoding and coding regions (Katti et al. 2001); however, they are more frequent in noncoding regions (Hancock 1995). A number of evidences suggest that noncoding SSRs may also be functionally significant (Kashi et al. 1997).

SSRs are also found in the chloroplast (Kapil et al. 2014) and mitochondrial (Kumar et al. 2014) genomes. Plant genomes contain large number of SSRs (mostly di- and tri- repeats) scattered over many loci of chromosome (Shanker et al. 2007a, b; Jones et al. 2009; Gutiérrez-Ozuna and Hamilton 2017). The hypervariability, reproducibility, specificity, and codominant nature of SSRs together make them potential molecular markers (Squirrell et al. 2003). Moreover, they are widely used for population-level evolutionary studies and genotyping (Sung et al. 2010). The conserved flanking DNA sequences of SSRs are used to design primers for PCR which in turn amplify the SSR in target species (Botstein et al. 1980; Kabra et al. 2016).

Due to insertion or deletion of repeat motif, SSRs show high level of polymorphism (Tautz and Renz 1984) which may be essential in the evolution of gene regulation (Moxon and Wills 1999). Gerber et al. (1994) demonstrated that simple homopolymeric stretches of proline or glutamine amino acids, encoded by rapidly evolving trinucleotide SSRs, may cause modulation of transcription factor activity. Sometimes transcriptional activity may also be affected by the length of SSRs lying in promoter regions (Kashi et al. 1997). SSR mining through available conventional biotechnological methods is quite tiresome, lengthy, and costly (Kumpatla and Mukhopadhyaya 2005), so now they are being replaced by *in silico* methods, since use of computational approaches for data mining of available massive sequential data from open data sources allows more economical and efficient way of extraction of SSRs (Shanker et al. 2007a, b).

A large number of complete genome sequences and expressed sequence tags (ESTs) are available at National Center for Biotechnology Information (NCBI) which is a major resource for molecular biology information with access to many public databases. Many of these genomes have been mined for SSRs (Coenye and Vandamme 2005; Batwal et al. 2011; Kabra et al. 2016). Earlier, a large amount of sequence data downloaded from NCBI was mined and online databases of obtained information were developed (Kumar et al. 2014; Kapil et al. 2014; Kabra et al. 2016). These databases contain detailed information of perfect, imperfect, and compound SSRs. A number of interactive features for easy and efficient searching are provided. To mine perfect and compound SSRs, microsatellite identification tool (MISA; http://pgrc.ipk-gatersleben.de/misa/download/.pl) was used on FASTA formatted nucleotide sequences. A minimum length criterion (≥12 mono-, ≥6 di-, ≥4 tri-, and ≥ 3 for tetra-, penta-, and hexanucleotide repeats) of repeating motif with a difference of 0 between two SSRs was taken. To mine imperfect SSRs, Imperfect Microsatellite Extractor (IMEx 2.0; Mudunuri and Nagarajaram, 2007) with 10% imperfection percentage was used. The results of MISA and IMEx were parsed with Perl scripts. To generate PCR primers, Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/; Untergrasser et al. 2012) with default parameters was used considering 200 bases of SSR upstream and downstream regions each. A brief description of these databases is as follows:

MitoSatPlant (Kumar et al. 2014; www.compubio.in/mitosatplant/) is an online database of mitochondrial microsatellites (mtSSRs) of Viridiplantae. mtSSRs hold importance in various fields like plant phylogenetics, genome mapping, and population genetics which made MitoSatPlant a very useful resource for plant scientists. ChloroSSRdb (Kapil et al. 2014; www.compubio.in/chlorossrdb/) is a repository of perfect and imperfect chloroplast simple sequence repeats (cpSSRs) of green plants. To calculate relative frequency, chi-square statistics, and correlation coefficient of perfect and imperfect SSRs, statistical analyses were performed for each species. Along with this, a graphical representation of the comparison of chi-square values of imperfect and perfect SSRs is also provided which will help in understanding evolutionary pattern of cpSSRs. CyanoSat (Kabra et al. 2016; www.compubio.in/CyanoSat/) is a database of cyanobacterial perfect and imperfect microsatellites.

In order to explain data mining to detect common, unique, and putative polymorphic SSRs using a methodology developed by our group (Kabra et al. 2016), a case study on complete chloroplast genome sequences of three species of genus *Triticum* (*Triticum aestivum*, NC_002762; *Triticum monococcum*, NC_021760; and *Triticum urartu*, NC_021762) is discussed here. The genus *Triticum* belongs to the tribe Triticeae of the Pooideae subfamily of grasses. It has been used in many cytogenetic and taxonomic studies (Barkworth 1992; Heslop-Harrison 1992). Bread wheat (*Triticum aestivum*), a hexaploid, is an important member of the tribe and is produced at a very large scale in the world. Earlier, molecular markers have been used in study of genetic diversity and phylogeny among the species of Triticeae (Ogihara and Tsunewaki 1988; Dvorak and Zhang 1992) out of which SSRs become the markers of choice (Gupta and Varshney 2000). Moreover, studies were also conducted in wheat genome for genome mapping (Gupta et al. 2002), physical mapping (Roder et al. 1998), and gene tagging (Roy et al. 1999) using SSRs. In present study, only perfect SSRs were mined which were further utilized to detect putative polymorphic, unique, and common SSRs.

7.2 Mining of Perfect SSRs

FASTA (*.fna) and GenBank (*.gbk) formatted files of chloroplast genome sequences of the selected species were downloaded from NCBI (www.ncbi.nlm.nih.gov/). Only perfect SSRs were mined as discussed above. The analysis of three *Triticum* species yielded 60 perfect SSRs. The highest density of SSRs was observed in *T. aestivum* (1 SSR/5.38 kb) followed by *T. monococcum* (1 SSR/6.47 kb) and *T. urartu* (1 SSR/6.81 kb) which is in accordance with the decreasing order of their chloroplast genome size. Earlier, studies were conducted to analyze the relationship between the genome size and SSR content. From these studies, mixed conclusions were drawn wherein some suggest that the abundance of SSR is directly proportional to the genome size (Primmer et al. 1997), whereas other studies proved that they are inversely proportional (Morgante et al. 2002). The observed density of SSR for *T. aestivum* was higher than the previously observed SSR density for the wheat chromosome arm 3AS-specific library (1/10.4 kb; Sehgal et al. 2012) but similar to wheat chloroplast genome (1 SSR/5.38 kb; Tomar et al. 2014). Higher density of SSRs was observed in *T. aestivum* and *T. monococcum* when compared to cpSSR density of rice (1SSR/6.5 kb; Rajendrakumar et al. 2007) while lower in *T. urartu* (1 SSR/6.81 kb). Similar decreasing pattern was observed in the average length of SSR, viz., *T. aestivum* (13.04 bp) > *T. monococcum* (12.44 bp) > *T. urartu* (12.41 bp). The information of SSRs in three *Triticum* species is represented in Table 7.1.

Among the identified perfect cpSSRs, only 13 (21.67%) were present in the coding regions which are considered to be highly conserved. Therefore, the designed

Table 7.1 Information of SSRs identified in chloroplast genomes of genus *Triticum*

	Organisms		
Parameter	*T. aestivum*	*T. monococcum*	*T. urartu*
Chloroplast genome size	134545 bp	116399 bp	115773 bp
Total SSRs identified	25	18	17
Density of SSR	1 SSR/5.38 kb	1 SSR/6.47 kb	1 SSR/6.81 kb
Coding region	5 (20%)	4 (22.22%)	4 (23.52%)
Average length of SSR	13.04 bp	12.44 bp	12.41 bp
Repeat type			
Mononucleotides	10 (40%)	6 (33.33%)	4 (23.53%)
Dinucleotides	1 (4%)	–	–
Trinucleotides	3 (12%)	3 (16.66%)	3 (17.65%)
Tetranucleotides	8 (32%)	8 (44.44%)	9 (52.95%)
Pentanucleotides	3 (12%)	1 (5.55%)	1 (5.88%)

primers corresponding to coding SSRs can be used to develop molecular markers and genetic diversity studies. Tetranucleotides were the most abundant repeat type with a frequency of 9 (52.95%) in *T. urartu* and 8 in both *T. monococcum* (44.44%) and *T. aestivum* (32%). Trinucleotide repeats with a frequency of 3, bearing similar motifs, length, and region, were present in all. Moreover, 20 mononucleotide repeats, 5 pentanucleotide repeats, and 1 dinucleotide repeat was found. Hexanucleotide repeats were not detected in these chloroplast genomes.

7.3 Identification of Putative pSSRs

The mined cpSSRs were used to retrieve putative polymorphic SSRs (pSSRs) with the help of Perl scripts. Length variations in SSRs with identical repeating units in all other species were identified, for example, motif AC repeating six times, i.e., $(AC)_6$ with a total length of 12 nucleotides, and motif AC repeating seven times, i.e., $(AC)_7$ with a total length of 14 nucleotides. The upstream and downstream regions of both the SSRs containing 200 nucleotides for each were retrieved from FASTA sequence. A reciprocal similarity search of the flanking regions of SSR was performed using BlastN (Altschul et al. 1997). Reciprocal similarity search is a two-step process where in first step one sequence is taken as database and the other as query and in the second step query sequence will become database and the database sequence in previous step will be treated as query. Significant match thus obtained was reported as putative pSSR. MapChart (Voorrips 2002) was used to represent the location of identified SSRs and to highlight the identified putative pSSRs on the respective chloroplast genomes.

Of the identified 60 perfect SSRs, length variation was detected in 20 SSRs, and only 3 (Fig. 7.1) were identified as putative pSSRs (Table 7.2). All the identified putative pSSRs were present in intergenic regions. Mononucleotide repeat (A) was the only putative pSSR detected in all species, and these can be utilized for species identification. The length of identified pSSRs ranged from 12 to 15 nucleotides, and these were found only in noncoding regions of chloroplast genomes. *T. monococcum* and *T. urartu* take up same PCR primer pair; however, *T. aestivum* differs in right primer. These primers can be used for the validation of identified pSSRs and to test genetic variability among *Triticum* species.

7.4 Identification of Common SSRs

To detect common SSRs, identical repeating units with equal length of SSRs in all other species were considered, for example, motif GT repeating 7 times, i.e., $(GT)_7$ with a total length of 14 nucleotides, and motif GT repeating 7 times, i.e., $(GT)_7$ with a total length of 14 nucleotides in another species. A significant match of SSRs with flanking regions in reciprocal similarity search was recorded as common SSRs.

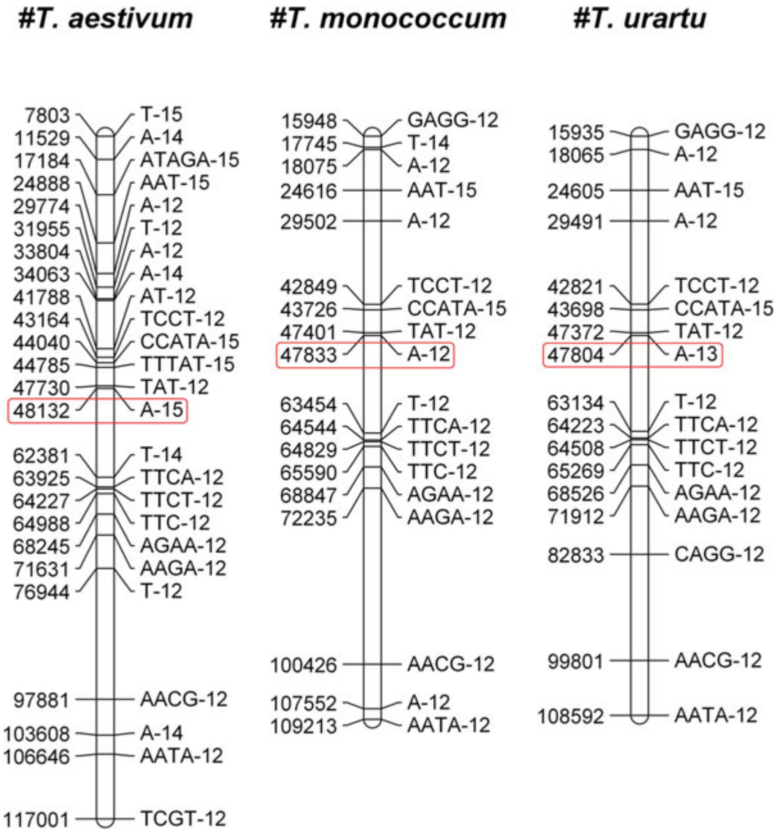


Fig. 7.1 Position of SSRs and putative pSSR identified in the chloroplast genomes of genus *Triticum*

A total of 12 such cpSSRs in all the 3 *Triticum* species (Table 7.3) were found that are common, and 2 SSRs were detected to be common between *T. monococcum* and *T. urartu* (Fig. 7.2). Out of these 12 SSRs, 4 repeats lie in coding regions.

7.5 Identification of Unique SSRs

To identify unique SSRs, identical repeating units among all the species (including species to which the query sequence belongs to) were used. Repeats with no significant match of flanking regions in any of the species considered were treated as unique. A total of 13 unique SSRs were identified (Table 7.4). Locations of these

Table 7.2 Putative pSSRs identified in genus *Triticum* with designed PCR primers

Organism name	Motif	Length	Start	End	Left primer	Right primer
T. aestivum	A	15	48132	48146	TCCTCGTGTCACCAGTTCAA	CCGCGCACATTACTTAGCAC
T. monococcum	A	12	47833	47844	TCCTCGTGTCACCAGTTCAA	GGGAATCTCACCCCCTTCTT
T. urartu	A	13	47804	47816	TCCTCGTGTCACCAGTTCAA	GGGAATCTCACCCCCTTCTT

Table 7.3 Common SSRs identified in genus *Triticum*

	T. aestivum	*T. monococcum*	*T. urartu*
	Start-end	Start-end	Start-end
	Left primer→	Left primer→	Left primer→
Motif-length	→Right primer	→Right primer	→Right primer
A-12	29774-29785	29502-29513	29491-29502
	AAAAGCTGCCCTACGAGGTC	AAAAGCTGCCCTACGAGGTC	AAAAGCTGCCCTACGAGGTC
	TCTCTCATTTCCGACGCGAA	TCTCTCATTTCCGACGCGAA	TCTCTCATTTCCGACGCGAA
AAT-15	24888-24902	24616-24630	24605-24619
	AGCGTCGAGGTATTTGTGCA	AGCGTCGAGGTATTTGTGCA	AGCGTCGAGGTATTTGTGCA
	ATGACGCGTTGATCCAGGTT	ATGACGCGTTGATCCAGGTT	ATGACGCGTTGATCCAGGTT
TAT-12	47730-47741	47401-47412	47372-47383
	AGTTGTGAGGGTTCAAGTCCC	AGTTGTGAGGGTTCAAGTCCC	AGTTGTGAGGGTTCAAGTCCC
	CCATTGAGTTCTCTTCGCATTCC	CCATTGAGTTCTCTTCGCATTCC	CCATTGAGTTCTCTTCGCATTCC
TTC-12	64988-64999	65590-65601	65269-65280
	AGAGCATAGAAAAGGCGGGG	AGAGCATAGAAAAGGCGGGG	AGAGCATAGAAAAGGCGGGG
	TTCCCTTGGCCATGAACCTC	TTCCCTTGGCCATGAACCTC	TTCCCTTGGCCATGAACCTC
AACG-12	97881-97892	100426-100437	99801-99812
	AGTGCTCTCTCCTCCGACTT	AGTGCTCTCTCCTCCGACTT	AGTGCTCTCTCCTCCGACTT
	TCGATTTGATCAGGCCGTGT	TCGATTTGATCAGGCCGTGT	TCGATTTGATCAGGCCGTGT
AAGA-12	71631-71642	72235-72246	71912-71923
	GGGAAGAGAAAAGTCAAGAAGCC	GGGAAGAGAAAAGTCAAGAAGCC	GGGAAGAGAAAAGTCAAGAAGCC
	TCATCTCGTACGGCTCAAGC	TCATCTCGTACGGCTCAAGC	TCATCTCGTACGGCTCAAGC

AATA-12	106646-106657	109213-109224	108592-108603
	GTTCTCGTGGTCCAGAATCCA	GTTCTCGTGGTCCAGAATCCA	GTTCTCGTGGTCCAGAATCCA
	GCTTCTCTTGCCTTACCAGGA	GCTTCTCTTGCCTTACCAGGA	GCTTCTCTTGCCTTACCAGGA
AGAA-12	68245-68256	68847-68858	68526-68537
	CTTTTGGAACACCAATGGGCA	CTTTTGGAACACCAATGGGCA	CTTTTGGAACACCAATGGGCA
	AGCATTCGAATCACCCATTCCT	AGCATTCAAATCACCCATTCCT	AGCATTCAAATCACCCATTCCT
TCCT-12	43164-43175	42849-42860	42821-42832
	ACCCAGTCGCTCACTAATTGA	ACCCAGTCGCTCACTAATTGA	ACCCAGTCGCTCACTAATTGA
	GGGCTTTCTACATAGGGATCGT	GGGCTTTCTACATAGGGATCGT	GGGCTTTCTACATAGGGATCGT
TTCA-12	63925-63936	64544-64555	64223-64234
	ACATCGGGTTTTGGAGACCC	ACATCGGGTTTTGGAGACCC	ACATCGGGTTTTGGAGACCC
	TGGTAGCGCGTTTGTTTTGG	TGGTAGCGCGTTTGTTTTGG	TGGTAGCGCGTTTGTTTTGG
TTCT-12	64227-64238	64829-64840	64508-64519
	CCAAAACAAACGCGCTACCA	CCAAAACAAACGCGCTACCA	CCAAAACAAACGCGCTACCA
	AGAAGAAATGACACGAGGGTTCT	AGAAGAAATGACACGAGGGTTCT	AGAAGAAATGACACGAGGGTTCT
CCATA-15	44040-44054	43726-43740	43698-43712
	TCCCAACCATTCTTCCCAGC	TCCCAACCATTCTTCCCAGC	TCCCAACCATTCTTCCCAGC
	CCTTCCGTCGTGTATCCTCG	CCTTCCGTCGTGTATCCTCG	CCTTCCGTCGTGTATCCTCG
GAGG-12	–	15948-15959	15935-15946
		TCAATTGGTCAGAGCACCGC	TCAATTGGTCAGAGCACCGC
		GTGTCTTTCTTTATCGATTGGGAGT	TGTGTCTTTCTTTATCGATTGGGA
A-12	–	18075-18086	18065-18076
		TCGTGGAGTCCCTTCTTGAA	TCGTGGAGTCCCTTCTTGAA
		ACGGTAGTGGCCAAAATGGT	ACGGTAGTGGCCAAAATGGT

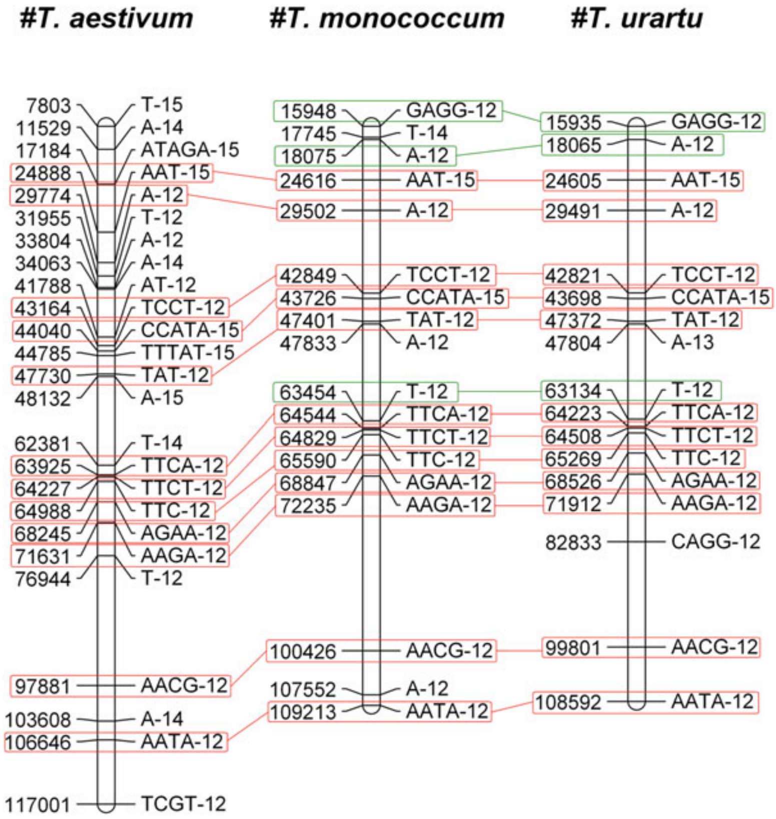


Fig. 7.2 Position of common SSRs identified in the chloroplast genomes of genus *Triticum*

SSRs in respective chloroplast genomes are represented in Fig. 7.3. Maximum frequency of unique SSRs was found in *T. aestivum* (10) and minimum in *T. urartu* (1). Two unique SSRs were detected in *T. monococcum*. Mononucleotides were most abundant repeat type found as unique SSRs. Except for one SSR in *T. aestivum*, i.e., $(TCGT)_3$, which codes for 4.5S ribosomal RNA, the rest of all unique SSRs lie in intergenic regions and will also help in the identification of species.

The methodology used in this analysis can be applied to detect unique, common, and putative polymorphic SSRs in nucleotide sequences of other organisms.

Table 7.4 Unique SSRs identified in genus *Triticum*

Organism name	Motif	Length	Start	End	Left primer	Right primer
T. aestivum	A	14	11529	11542	AGCCGAGCCATTCATTCCTT	TAGGCTGTGGAGAGATGGCT
	A	14	103608	103621	AGAATGGGTTTAGTTGGTTAAATTCA	TGGAACAAGAGAGCTGTTTCA
	T	15	7803	7817	TCCTGGACGCGAGGAGTAAT	TTGGAACGTGGAGAGATGGC
	T	12	31955	31966	GCCGCAGCCTATATAGGTGA	TCTAGCCTGACTCCACCCTC
	T	14	62381	62394	TGGTTTTGCCTGGTTCCGAA	AGATTTTGGCTGCTATTCCGA
	T	12	76944	76955	AGCAATAGTGTCCTTGCCCA	TCGAGTTTTTGGCGCGATTG
	AT	12	41788	41799	AGTCCTCCTCTTTCCGGACA	CGAACACTTGCCTCGGATTG
	TCGT	12	117001	117012	TCGATTTGATCAGGCCGTGT	AGTGCTCTCTCCTCCGACTT
	ATAGA	15	17184	17198	TGGTTGGAATCTTAAAGTGTGGT	TCCTTCCATGGATTCTTTGGTCA
	TTTAT	15	44785	44799	AGCGTATCTTATGCAAACGGA	AGGTACAACCGCAACCACTC
T. monococcum	A	12	107552	107563	TGTGTTTCTATTGGGCAAAGCA	ACCATGGCGGCTAACTTCAA
	T	14	17745	17758	TGTCCTGAATAGAAGAAGCGGG	AAGGCTGCAGGGTTTAGGTC
T. urartu	CAGG	12	82833	82844	GGGGGAAGAGTGAGAGAGGT	GAATCGTACCTGGCCTCACC

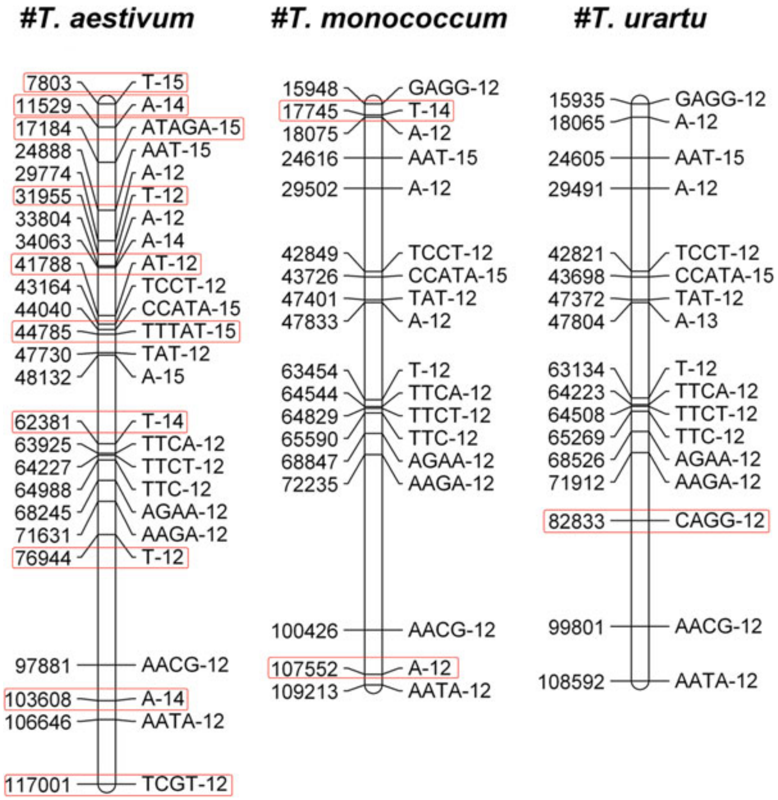


Fig. 7.3 Position of unique SSRs identified in the chloroplast genomes of genus *Triticum*

References

Altschul SF, Madden TL, Schaffer 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 25:3389–3402

Bachmann L, Bare PTJ (2004) Allelic variation, fragment length analysis and population genetic model: a case study on *Drosophilla* microsatellites. Zool Syst Evol Res 42:215–222

Barkworth ME (1992) Taxonomy of the Triticeae: a historical perspective. Hereditas 116:1–14

Batwal S, Sitaraman S, Ranade S, Khandekar P, Bajaj S (2011) Analysis of distribution and significance of simple sequence repeats in enteric bacteria *Shigella dysenteriae* SD197. Bioinformation 6:348–351

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map inman using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Coenye T, Vandamme P (2005) Characterization of mononucleotide repeats in sequenced prokaryotic genomes. DNA Res 12:221–233

Dvorak J, Zhang H-B (1992) Application of molecular tools for study of the phylogeny of diploid and polyploid taxa in Triticeae. Hereditas 166:37–42

Field D, Wills C (1996) Long, polymorphic microsatellites in simple organisms. Proc Biol Sci 263:209–215

Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, Schaffner W (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263:808–811

Gupta PK, Varshney RK (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica 113:163–185

Gupta PK, Balyan HS, Edwards KJ, Isaac P, Korzun V, Roder M, Jourdrier P, Schlatter AR, Dubcovsky J, de la Pena RC, Khairallah M, Hayden M, Keller B, Wang R, Hardouin JP, Jack P, Leroy P (2002) Genetic mapping of 66 new SSR loci in bread wheat. Theor Appl Genet 105:413–422

Gutiérrez-Ozuna R, Hamilton MB (2017) Identification and characterization of microsatellite loci in the tuliptree, *Liriodendron tulipifera* (Magnoliaceae). Appl Plant Sci 5(8):pii: apps.1700032. https://doi.org/10.3732/apps.1700032

Hancock JM (1995) The contribution of slippage-like processes to genome evolution. J Mol Evol 41:1038–1047

Heslop-Harrison JS (1992) Molecular cytogenetics, cytology and genomic comparisons in the Triticeae. Hereditas 116:93–99

Jones N, Ougham H, Thomas H, Pasakinskiense I (2009) Markers and mapping revisited: finding your gene. New Phytol 183:935–966

Kabra R, Kapil A, Attarwala K, Rai PK, Shanker A (2016) Identification of common, unique and polymorphic microsatellites among 73 cyanobacterial genomes. World J Microbiol Biotechnol 32:71

Kaila T, Chaduvla PK, Rawal HC, Saxena S, Tyagi A, Mithra SVA, Solanke AU, Kalia P, Sharma TR, Singh NK, Gaikwad K (2017) Chloroplast genome sequence of Clusterbean (*Cyamopsis tetragonoloba* L.): genome structure and comparative analysis. Genes (Basel) 8(9):E212. https://doi.org/10.3390/genes8090212

Kapil A, Rai PK, Shanker A (2014) ChloroSSRdb: a repository of perfect and imperfect chloroplastic simple sequence repeats (cpSSRs) of green plants. Database 2014:1–5

Kashi Y, King D, Soller M (1997) Simple sequence repeats as a source of quantitative genetic variation. Trends Genet 13:74–78

Katti MV, Rajenkar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167

Kumar M, Kapil A, Shanker A (2014) MitoSatPlant: mitochondrial microsatellites database of Viridiplantae. Mitochondrion 19:334–337

Kumpatla SV, Mukhopadhyaya S (2005) Mining and survey of simple sequence repeats in expressed sequence tags in dicotyledonous species. Genome 48:985–998

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Moxon ER, Wills C (1999) DNA microsatellites: agents of evolution. Sci Am 280:94–99

Mudunuri SB, Nagarajaram HA (2007) IMEx: imperfect microsatellite extractor. Bioinformatics 23:1181–1187

Ogihara Y, Tsunewaki K (1988) Diversity and evolution of chloroplast DNA in *Triticum and Aegilops* as revealed by restriction fragment analysis. Theor Appl Genet 76:321–332

Primmer CR, Raudsepp T, Chowdary BP, Moller AP, Ellegren H (1997) Low frequency of microsatellites in the avian genome. Genome Res 7:471–482

Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM (2007) Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. Bioinformatics 23:1–4

Roder MS, Korzun V, Gill BS, Ganal MW (1998) The physical mapping of microsatellite markers in wheat. Genome 41:278–283

Roy JK, Prasad M, Varshney RK, Balyan HS, Blake TK, Dhaliwal HS, Singh H, Edwards KJ, Gupta PK (1999) Identification of a microsatellite on chromosomes 6B and a STS on 7D of bread wheat showing an association with preharvest sprouting tolerance. Theor Appl Genet 99:336–340

Sehgal SK, Li W, Rabinowicz PD, Chan A, Simkova H, Dolezel J, Gill BS (2012) Chromosome arm-specific BAC end sequences permit comparative analysis of homoeologous chromosomes and genomes of polyploid wheat. BMC Plant Biol 12:64

Shanker A, Bhargava A, Bajpai R, Singh S, Srivastava S, Sharma V (2007a) Bioinformatically mined simple sequence repeats in UniGene of *Citrus sinensis*. Sci Hort 113:353–361

Shanker A, Singh A, Sharma V (2007b) *In silico* mining in expressed sequences of *Neurospora crassa* for identification and abundance of microsatellites. Microbiol Res 162:250–256

Squirrell J, Hollingsworth PM, Woodhead M, Russell J, Low AJ, Gibby M, Powell W (2003) How much effort is required to isolate nuclear microsatellites from plants? Mol Ecol 12:1339–1348

Sung W, Tucker A, Bergeron RD, Lynch M, Thomas WK (2010) Simple sequence repeat variation in the *Daphnia pulex* genome. BMC Genomics 11:691

Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res 12:4127–4138

Tautz D, Schlotterer C (1994) Simple Sequences. Curr Opin Genet Dev 4:832–837

Tomar RSS, Deshmukh RK, Naik K, Tomar SMS (2014) Development of chloroplast-specific microsatellite markers for molecular characterization of alloplasmic lines and phylogenetic analysis in wheat. Plant Breed 133:12–18

Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3: new capabilities and interfaces. Nucleic Acids Res 40:e115

Vogt P (1990) Potentially genetic functions of tandemly repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code". Hum Genet 84:301–336

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

R-Programming for Genome-Wide Data Analysis

8

Arunima Shilpi and Shraddha Dubey

8.1 Introduction

Organisms on this planet are specified by myriad genomes which differ from individual to individual. Genome contains the biological information of an organism and is needed to sustain life. Human genome and all other cellular life forms have genome made up of deoxyribonucleic acid (DNA); only a few viruses have ribonucleic acid (RNA) genomes. With the advancement of sequencing technologies, several genome projects have been successfully accomplished. Consequently, a large amount of genomic data is available and easily accessible from various genomic databases. This genomic dataset is further analysed for better understanding of organisms and the interaction with the environment (Zhao and Tan 2006). In the intervening period of time, this genomic data and its analysis provide the scientific community with noble challenges which stimulates perplexing array of new algorithms and relevant software. Due to this several new approaches and outputs are generated for a wide variety of genomic data. The easy accessibility of biological data makes it convenient for researchers to isolate the gene responsible for a particular symptom in a diseased person (Lander et al. 2001). Drugs for several non-curable dreadful diseases are now available in market based on the genomic data analysis. The effectiveness of the genome data analysis has reached to new heights with the availability of various programming languages incorporating the statistical analysis and validation of results. Several statistical

Author contributed equally with all other contributors. Arunima Shilpi and Shraddha Dubey

A. Shilpi (✉)
Department of Life Science, National Institute of Technology, Rourkela, Odisha, India

S. Dubey
Department of Bioinformatics, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_8

tools are available for the users to generate most appropriate output with high accuracy and precision values (Zhao and Tan 2006).

With the recent advancement in high-throughput technology comprising of next-generation sequencing (NGS), which produce millions of sequences in a single instrument run, it is now possible to catalogue genetic variants in population to understand disease and its evolution. The extensive genome-wide analysis generates ample of data, and it requires computational call to translate the output into high-quality variant calls. Unified analytic framework has been developed to analyse and interpret the variation in genetic dataset characterized by single nucleotide polymorphism to achieve sensitive and specific findings in characterizing the disease diagnosis and prognosis. The genome-wide data analysis incorporates several software tools which are further divided into many categories such as tools for segregation analysis, tools for linkage analysis, tools for mapping polygenic traits and experimental crosses including mapping of quantitative trait loci. Further, it includes genetic association studies, phylogenetic analysis, family history of organisms, and microarray analysis using genomic datasets (Lander et al. 2001).

Earlier, Fortran programming language was used for the development of Likelihood in Pedigrees (LIPED) program which estimates recombination fractions for linkage analysis (O'Neill et al. 1998). Further, several other software tools were designed for linkage and segregation analysis. Apart from this, vigorous efforts have been done for the development of various software tools for facilitating the analysis task. MORGAN is one of the programs based on Bayesian methods (Brooks et al. 2000). These programs were written in different computer languages such as C, C++, Fortran, Pascal, Java, Perl, Stata, SAS, and S-PLUS. These were tested for specific computer systems and take input data in specific format, and sometimes it becomes difficult for the users to use the output generated by these programs (Guo and Lange 2000). Old parsing algorithms have been used for some programs, while some lacks the graphical user interface which sometimes becomes a tough task for the end users operating them, with no customized utilities for these programs if operated for a large dataset. For writing the customized utility programs and for operating the tools, different programming languages and skills are required which often results into redundant work, low maintenance, and lack of validation.

As a result of which for a data analyst, it becomes sometimes tedious to operate various software and keep check on every input and output programs which results into lack of accuracy and precision. Ideally a computer program must be designed in a language which is easily operated on any computer systems, having excellent data management and established algorithms and clear documentation, which provides a graphical user interface (GUI), and which manages large data set in batches. The programs must be designed in a language which is eloquent and flexible for the users to track the errors and modify accordingly the source codes. The input system must be designed in a way such that it has connectivity with online and offline data sources such as the large genomic databases available publicly.

However, all these features are difficult to achieve by programmers, but with the recent development of computational techniques and operating systems, such a platform have become possible. Several programs based on Linux operating system have overcome these pre-existing problems, and in recent years R-statistical programming has undergone revolution in high-throughput analysis and interpretation of genomic data. R could potentially serve as an integrated platform for genetic data analysis.

8.2 R: A Brief Overview

R is a de facto programming language for statisticians as it holds wide application in high-throughput data analysis (Clayton and Leung 2007). R is freely available with an open source interface and is platform independent. Moreover, R is a scripting language, so it provides an ease to assemble several packages, adds on personalized routines, and provides link to pipeline for processing raw data to desired results. Consequently, it minimizes the time and effort for the analysis of complex data. R is compatible with most of the operating systems like Linux, Unix, Windows, and Mac. R is an object-oriented programming language and provides useful application for representation of classes and methods. In addition, R provides graphic interface (R-studio) which facilitates in reading and editing the data having variable formats (SPSS, Stata, SAS, dBase). R can be run both as GUI and in batch mode which allow the user to customize their needs. It also connects common gateway interface (CGI) and HTML/XML outputs. The packages constituting RODBC (Sanders et al. 1997) and RMySQL are beneficial for connecting the MySQL databases with open database connectivity.

The greater strength in implementing R lies on the fact that it has enormous modules and packages available freely in the repositories primarily Bioconductor (http://www.bioconductor.org/) and Comprehensive R Archive Network (CRAN; http://cran.r-project.org) (Gentleman et al. 2004). The packages have been autonomously developed from the core program and can be directly obtained from central repositories. The availability of these packages saves time and can be a shortcut to the desired results. Several packages have been made available for the analysis of data in genome-wide association studies (Amos 2007). These packages have wide application in importing data having varying formats, preprocessing, quality control, and computing statistical significance. Most recent methodologies demand the updated R packages which are made available as per the requirement. A large number of algorithms and methods are being published in order to minimize the turnover time for commercial application.

8.3 R Bioconductor in the Study of Next-Generation Sequencing Data

The advent of next-generation sequencing technologies has revolutionized in genome-wide analysis of sequences. High-throughput sequencing brought by Illumina, Roche, ABI, and Helicos generates enormous data. The data generated by NGS primarily 1000 Genome Project exclusively holds data in terabases. R Bioconductor provides good support for Illumina- and Affymetrix-based data analysis and also has some support to Roche 454. It can be successfully implemented in study of several research questions including:

(a) Analysis of variation in DNA-seq data: Bioconductor helps in analysis of whole-genome sequence or exomic DNA. Genome-wide analysis dictates in identification of indels, single nucleotide polymorphism (SNPs), CNV (copy number variation), and methylation pattern.
(b) Analysis of gene expression or RNA-seq data: Bioconductor guides in genome-wide study of reverse complement mRNA from the transcriptome.
(c) Analysis of ChIP-seq data. ChIP-seq or chromatin immune-precipitation is analysed to identify the presence of regulatory elements mainly the transcription factor associated with genomic DNA.
(d) Analysis of metagenomics data: Study of metagenomics data directs towards sequence analysis of multiple species, analysis of microbial niches, and phylogenetic study of species.

8.4 File Formats As an Input in Bioconductor Packages

Enormous data generated through high-throughput sequencing has to be converted into variable format for input. Different packages have their own input format as described below:

(a) FASTQ: FASTQ format is like FASTA format differing in few contexts like it also has quality score which can be read from Illumina machine. It is a text-based file format such that each record constitutes an identifier, followed by sequences and (+) sign, and at the end is the quality score in ASCII format.
(b) BAM (binary format of sequence alignment/mapping): Results obtained from alignment of the query sequence (reads) to the reference genome are stored in BAM file format. Mapping tools such as Bowtie, BWA, and STAR are implemented to align the sequences.
(c) bigWig (genome browser signal wiggle files in indexed binary format): User can transform BAM file into bigWig format in order to visualize the number of reads mapped to the reference genome in the form of continuous signal.
(d) bigBed (genome browser bed files in indexed binary format): ENCODE data analysis produces annotation files for RNA-seq, CHIP-seq, and bisulphite sequencing data. These annotated files can be visualized in the bigBed format in UCSC genome browser.

(e) VCF (variant call format): Gene sequence variation in the form of indel, SNP, and structural variants is represented in standard text file format which is also labelled as VCF format. The file format was developed due to upcoming of DNA sequencing and large-scale genotype project.

8.5 Method for Input and Output in R

Analysis in R begins with input of data stacked in databases and files. Once the data is imported, necessary functions are applied to analyse the data producing an output. Input of the data can be brought about by basic functions such as read.table (tabular data) and read.CSV (CSV files).

```
# Input
setwd(“home/user_name/directory_name”) # set up the path to which
the file in directory
x <- local(get(load(“filename.R”))) # file obtained is stored in
local variable x
# output
save(output_file, file= “output_file.RData”) # save output file with
extension .RData
#Input
Data <- “home/user_name/directory_name/filename”
x <- read.table(Data, sep = “”, header= T ) # Read input file and
mention any delamination or header if required
#output
write.table(output_file, file = “output_file.txt”, sep = “\t”) # save
output file as (txt/csv) with or without delamination
```

8.6 Data Types in R

Usually, while executing programming language, a user is required to use the data containers or variables to store the data. These variables are like reserved memory which stores values. Therefore, whenever variables are created, space is reserved in memory. Variables can store various data types mainly integer, float, double floating point, character, Boolean, etc. Unlike C and Java, the variables are not assigned to any data type in R. These variables are allocated to R objects. There are several types of R objects, some of which constitute vectors, list, arrays, matrices, factors, data frames, etc. Some of the other variables are logical, numeric, integer, and complex data type. Basic data types in R objects are called vectors which contain elements of variable classes. Here are the examples to see how variable data type can be implemented.

8.6.1 Vector

```
 colour <- c (“blue” , “black”, “white”)
 print (class(colour))
 Result
 [1] character
```

8.6.2 List

```
 y <- list (c(3,4, 5), 22.4, cos)
 print(y)
 Result
 [[1]]
 [1] 3 4 5
 [[2]]
 [1] 22.4
 [[3]]
 [1] function (x) .Primitive("cos")
```

8.6.3 Matrices

```
 y <- matrix( c('1', '2', '3', '4', '5', '6'), nrow = 3, ncol =2,
byrow = T)
 print(y)
 Result
 [,1] [,2]
 [1,] 1 2
 [2,] 3 4
 [3,] 5 6
```

8.6.4 Arrays

```
 y <- array(c ('black' , 'white'), dim = c(2,2,3))
 print(y)
 Results
 , , 1
 [,1] [,2]
 [1,] "black" "black"
 [2,] "white" "white"
 , , 2
 [,1] [,2]
 [1,] "black" "black"
 [2,] "white" "white"
 , , 3
 [,1] [,2]
 [1,] "black" "black"
 [2,] "white" "white"
```

8.6.5 Data Frames

```
Metabolic_rate <- data.frame (
 gender = c("Male", "Female"),
 weight = c(75, 60),
 height = c(170, 153),
 Age = c(35, 28))
print(Metabolic_rate)
Result
 gender weight height Age
1 Male 75 170 35
2 Female 60 153 28
```

8.7 Functions in R

To carry out a specified task, a piece of code written in a programming language acts as function. There are large numbers of functions in R. Some of the functions are enlisted in Table 8.1. Moreover, additional advantage of studying R is that user can also define their own function. Classification of functions can be numeric, character, or statistical.

8.8 Genome-Wide Analysis of Expression Data

Recent inventions in DNA microarray and Illumina platform-based NGS technology have provided high quality of gene expression and transcriptome activity to create genome-wide profile of cellular function in an organism. These methods quantify both gene and isoform level expression estimates, including identification and annotation of novel transcripts. The Human Genome Project have revealed the presence of probably 20,000 to 25,000 protein haploid coding genes such that only 15% encodes for coding RNA (exon), while the rest constitute non-coding genes mainly the introns and regulatory sequences. These genes undergo alternative splicing forming multiple transcripts or isoforms. The information is available for more than 100,000 spliced variants ((https://www.genome.gov/). These gene and isoform expression estimates determine the physiological changes and the

Table 8.1 List of primary functions in R

Functions	Description
Numerical	
sqrt(y)	Calculating square root of y
trunc(y)	Truncate y to manage the output
round(y, digit = n)	Round of y to n decimal place
cos(y), sin (y), tan(y)	Determining cos, sin, and tan value of y
log(y)	Calculating natural logarithm of y
exp(y)	Calculating exponential value for y
Character	
install.packages	Install R packages from online repository
%in% (match function)	Matching the elements of one vector with another
substr(y, start = n1, stop = n2)	Used to extract or replace substring y that starts with n1 and stops at n2, for example,
	Substr(mathematics, 1,4) is "math"
sub(old, new, y)	Old pattern is replaced by new in string y, for example, y < – "he lives in Delhi"
	Sub("Delhi", "Mumbai now", y) returns "he lives in Mumbai now"
grep(pattern, y, ignore. cases = T, fixed = F)	Pattern y is searched, and if fixed is false, it returns regular expression otherwise text string
	Grep("X", c("w", "X", "y"), fixed = T)
toupper(y)	Coverts lowercase to uppercase
tolower(y)	Converts uppercase to lowercase
Statistical	
mean(y)	Calculating mean for object y
median (y)	Calculating median for object y
sd(y)	Calculating standard deviation of y
quantile(vec, probs)	Quantile value for numeric vector is assigned in terms of probability which lies between 0 and 1; for example, quantile (vec, c(0.04, 0.96)) corresponds to 4% and 96% probable findings
t.test(y)	To find confidence interval for population mean (y) and level of significance is determined in terms of p-value
wilcox.test(y, conf.int = T)	Implemented to compute median's confidence interval
shapiro.test(y)	To determine the normal distribution of data
cor.test(x, y)	To determine the correlation between two variables x and y. For normal distribution default is Pearson
cor.test(x, y, method = spearman)	Correlation between two variables having non-normal distribution is determined by Spearman
aov(y ~ x, data = data frame)	One-way ANOVA is extended form of two-sample t-test which covers more than two independent groups

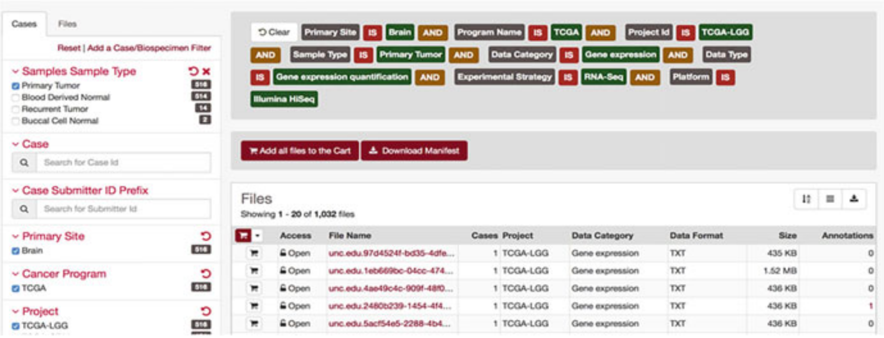
aberrant expression associated with several diseases including cancer. The databases like The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov), International Cancer Genome Consortium (ICGC; https://icgc.org), and

Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo) are the repositories for expression profile of tumour and normal samples in different cancers. Bioconductor offers several packages for differential expression analysis of high-throughput sequence data. The sequence information contained in FASTQ files are aligned to reference genome. The count matrix is generated based on the number of reads/fragments for each gene/isoform aligned to the reference genome. The software used for alignment and quantification to generate count matrix are *RSEM* (Li and Dewey 2011), *Kallisto* (Bray et al. 2016), *Sailfish* (Patro et al. 2014), and *Salmon* (Patro et al. 2017). These raw count matrices generated for different samples are normalized, and statistically significant differential gene expression (DGE) is determined using R packages such as *limma* (Law et al. 2014), *DESeq1/2* (Love et al. 2014), *EBseq* (Leng et al. 2013), *edgeR* (Robinson et al. 2010), and *baySeq* (Hardcastle and Kelly 2010). Stepwise DGE analysis using *limma and edgeR* package is as follows:

```
Installation of package on R-interface
source("http://www.bioconductor.org/biocLite.R")
biocLite("limma")
biocLite("edgeR")
library(limma)
library(edgeR)
help.start()
```

(a) Download raw count gene expression dataset for tumour and normal samples from TCGA GDC data portal (https://portal.gdc.cancer.gov) (Fig. 8.1).

(b) Assemble the raw counts for tumour and normal samples to build data matrix of dimensions (T_data; 20531 × 516) and (N_data; 20531 × 5), respectively. They are combined to build a new matrix of dimension (20531 × 521) (Fig. 8.2);



Samples Sample Type
☑ Primary Tumor 516
□ Blood Derived Normal 514
□ Recurrent Tumor 14
□ Buccal Cell Normal 2

Case

Case Submitter ID Prefix

Primary Site
☑ Brain 516

Cancer Program
☑ TCGA 516

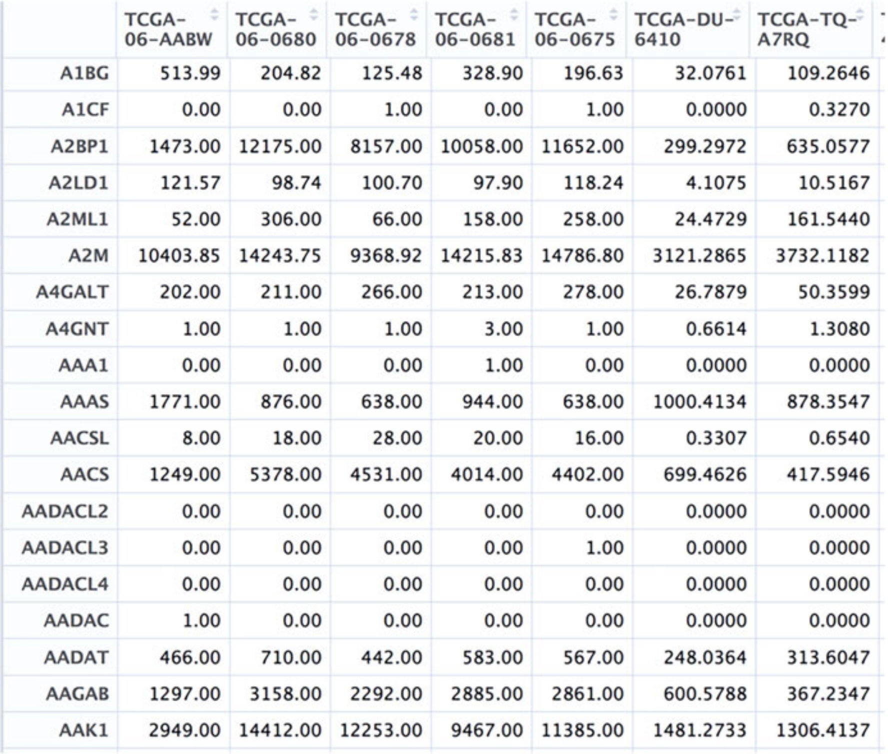
Project
☑ TCGA-LGG 516

Primary Site IS Brain AND Program Name IS TCGA AND Project Id IS TCGA-LGG AND Sample Type IS Primary Tumor AND Data Category IS Gene expression AND Data Type IS Gene expression quantification AND Experimental Strategy IS RNA-Seq AND Platform IS Illumina HiSeq

Files
Showing 1 - 20 of 1,032 files

Access	File Name	Cases	Project	Data Category	Data Format	Size	Annotations
Open	unc.edu.97d4524f-bd35-4dfe...	1	TCGA-LGG	Gene expression	TXT	435 KB	0
Open	unc.edu.1eb669bc-04cc-474...	1	TCGA-LGG	Gene expression	TXT	1.52 MB	0
Open	unc.edu.4ae49c4c-909f-48f0...	1	TCGA-LGG	Gene expression	TXT	436 KB	0
Open	unc.edu.2480b239-1454-4f4...	1	TCGA-LGG	Gene expression	TXT	436 KB	1
Open	unc.edu.5acf54e5-2288-4b4...	1	TCGA-LGG	Gene expression	TXT	436 KB	0

Fig. 8.1 Selection of gene expression dataset from TCGA GDC portal

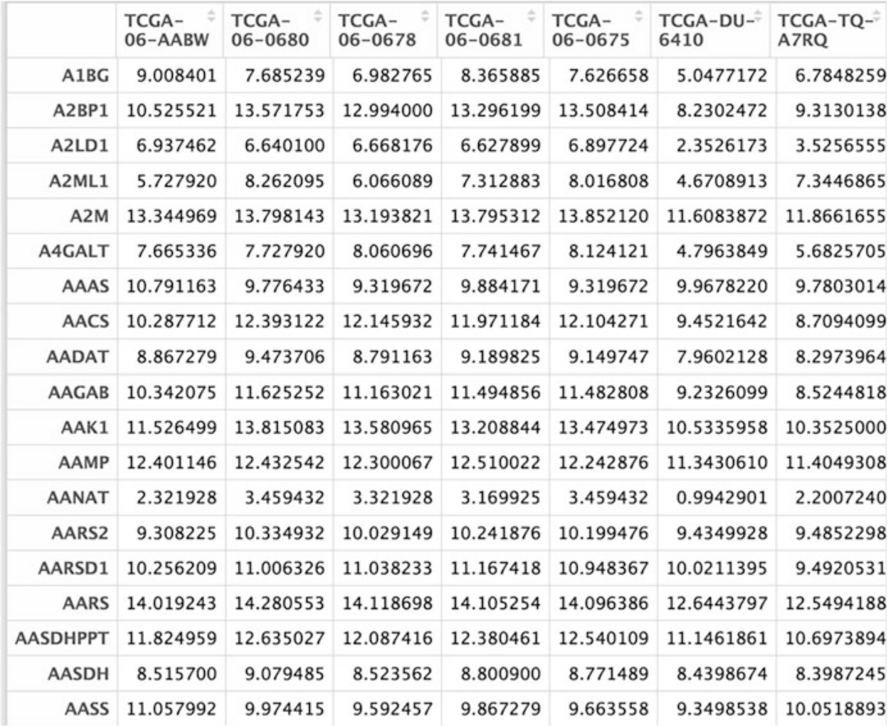


	TCGA-06-AABW	TCGA-06-0680	TCGA-06-0678	TCGA-06-0681	TCGA-06-0675	TCGA-DU-6410	TCGA-TQ-A7RQ
A1BG	513.99	204.82	125.48	328.90	196.63	32.0761	109.2646
A1CF	0.00	0.00	1.00	0.00	1.00	0.0000	0.3270
A2BP1	1473.00	12175.00	8157.00	10058.00	11652.00	299.2972	635.0577
A2LD1	121.57	98.74	100.70	97.90	118.24	4.1075	10.5167
A2ML1	52.00	306.00	66.00	158.00	258.00	24.4729	161.5440
A2M	10403.85	14243.75	9368.92	14215.83	14786.80	3121.2865	3732.1182
A4GALT	202.00	211.00	266.00	213.00	278.00	26.7879	50.3599
A4GNT	1.00	1.00	1.00	3.00	1.00	0.6614	1.3080
AAA1	0.00	0.00	0.00	1.00	0.00	0.0000	0.0000
AAAS	1771.00	876.00	638.00	944.00	638.00	1000.4134	878.3547
AACSL	8.00	18.00	28.00	20.00	16.00	0.3307	0.6540
AACS	1249.00	5378.00	4531.00	4014.00	4402.00	699.4626	417.5946
AADACL2	0.00	0.00	0.00	0.00	0.00	0.0000	0.0000
AADACL3	0.00	0.00	0.00	0.00	1.00	0.0000	0.0000
AADACL4	0.00	0.00	0.00	0.00	0.00	0.0000	0.0000
AADAC	1.00	0.00	0.00	0.00	0.00	0.0000	0.0000
AADAT	466.00	710.00	442.00	583.00	567.00	248.0364	313.6047
AAGAB	1297.00	3158.00	2292.00	2885.00	2861.00	600.5788	367.2347
AAK1	2949.00	14412.00	12253.00	9467.00	11385.00	1481.2733	1306.4137

Fig. 8.2 Overview of data matrix such that the row name is the gene id and column name is the sample

(c) The expression matrix is normalized to remove the low raw counts, and the background correction is done to subtract the low intensity expression. Generally the upper-quartile normalization process is used where the read count for each gene is divided by the 75th percentile of the total read counts. The following steps are used for normalization:

```
   group <- factor(c(rep(0, times=5), rep(1, times=516)))
 # grouping of normal and tumor samples into two class
 y <- DGEList(counts=combined_data,group=group)
 # Determination of the differential gene list for the dataset into
two groups
 z <- calcNormFactors(y, method = "upperquartile")
 # Upper quartile normalizes the data to its 75% percentile
 keep <- rowSums(cpm(z)>1) >= 2
  # Filter out genes such that total raw-count is lesser than
1. This filtration is user dependent.
 z1 <- z[keep, , keep.lib.sizes=FALSE].
```



	TCGA-06-AABW	TCGA-06-0680	TCGA-06-0678	TCGA-06-0681	TCGA-06-0675	TCGA-DU-6410	TCGA-TQ-A7RQ
A1BG	9.008401	7.685239	6.982765	8.365885	7.626658	5.0477172	6.7848259
A2BP1	10.525521	13.571753	12.994000	13.296199	13.508414	8.2302472	9.3130138
A2LD1	6.937462	6.640100	6.668176	6.627899	6.897724	2.3526173	3.5256555
A2ML1	5.727920	8.262095	6.066089	7.312883	8.016808	4.6708913	7.3446865
A2M	13.344969	13.798143	13.193821	13.795312	13.852120	11.6083872	11.8661655
A4GALT	7.665336	7.727920	8.060696	7.741467	8.124121	4.7963849	5.6825705
AAAS	10.791163	9.776433	9.319672	9.884171	9.319672	9.9678220	9.7803014
AACS	10.287712	12.393122	12.145932	11.971184	12.104271	9.4521642	8.7094099
AADAT	8.867279	9.473706	8.791163	9.189825	9.149747	7.9602128	8.2973964
AAGAB	10.342075	11.625252	11.163021	11.494856	11.482808	9.2326099	8.5244818
AAK1	11.526499	13.815083	13.580965	13.208844	13.474973	10.5335958	10.3525000
AAMP	12.401146	12.432542	12.300067	12.510022	12.242876	11.3430610	11.4049308
AANAT	2.321928	3.459432	3.321928	3.169925	3.459432	0.9942901	2.2007240
AARS2	9.308225	10.334932	10.029149	10.241876	10.199476	9.4349928	9.4852298
AARSD1	10.256209	11.006326	11.038233	11.167418	10.948367	10.0211395	9.4920531
AARS	14.019243	14.280553	14.118698	14.105254	14.096386	12.6443797	12.5494188
AASDHPPT	11.824959	12.635027	12.087416	12.380461	12.540109	11.1461861	10.6973894
AASDH	8.515700	9.079485	8.523562	8.800900	8.771489	8.4398674	8.3987245
AASS	11.057992	9.974415	9.592457	9.867279	9.663558	9.3498538	10.0518893

Fig. 8.3 $\log_2$ transformation of the matrix with the raw count values

```
# keep remaining in the list
combined_norm_counts <- data.frame(as.matrix(z1$counts))

 # Generate matrix for genes selected after filtration. The dimension of the matrix after filtration is 17347 x 521
```

(d) The raw count can be transformed to $\log_2$ scale for calculating fold change values and to determine up- and downregulated genes (Fig. 8.3).

```
combined_norm_log <- log2(combined_norm_counts+1)
```

(e) Differential gene analysis

```
   design <- data.frame(TCGA_id = colnames(combined_norm_log))
row.names(design) <- design$TCGA_id
design$Normal_ovary <- 1
design$OSCvsNormal_ovary <- 1
design$OSCvsNormal_ovary[1:ncol(N_Data)] <- 0
design$TCGA_id <- NULL
```

```
 design <- as.matrix(design)
 # Design matrix is created to determine the two group of dataset
as Tumor and the control (Normal) to compute the difference in
expression.
 fit_LT <- lmFit(combined_norm_log, design)
 #Lmfit: function is used to convert the expression estimates into
weights for quality assessment and discarding poor quality arrays.
 fit_LT <- eBayes(fit_LT)
 # eBayes function is moderated F-statistic test that combines with
t-statistic to compare the expression between the groups and output
as log2 Fold change, p-value, adjusted p-value.
 LT <- toptable(fit_LT, coef=ncol(design),number = dim(fit_LT)[1],
p.value = 1) # different cut-off
```

The toptable function determines the content of the table based on p-value cut-off (Fig. 8.4).

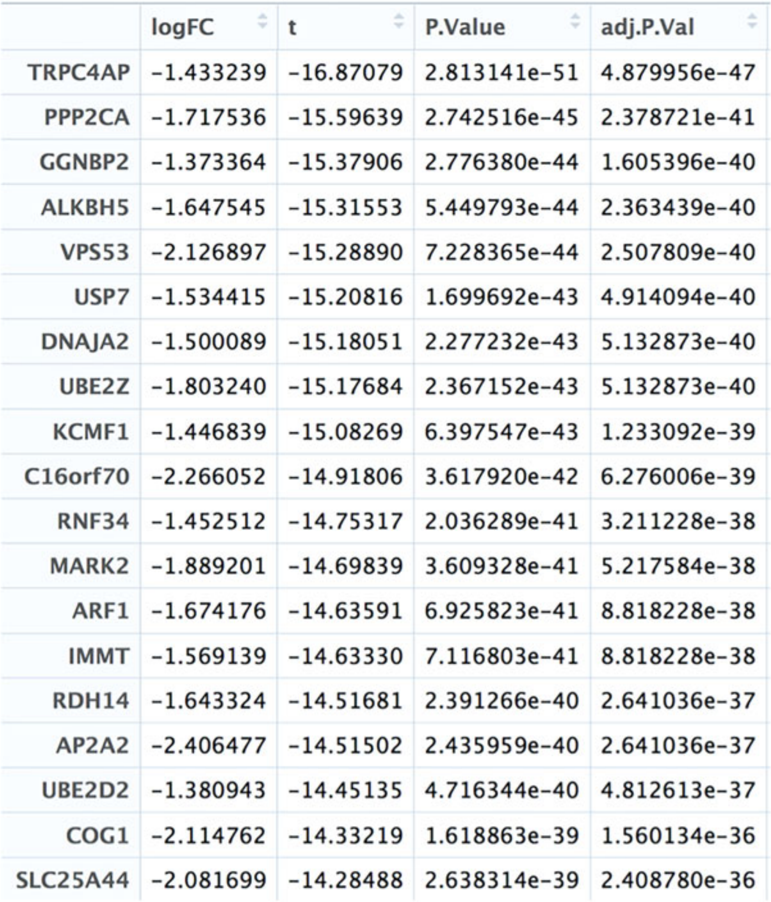
8.9 R Packages for Analysis of Large Data

The enormous amount of data generated through sequencing requires large memory for storage. Moreover, it confronts challenge to identify algorithm or design a tool for handling of data with an ease. R provides an interface where an efficient code restricts large data into subset as an input for subsequent statistical evaluation. Some of the packages that can be implemented in the analysis of large genetic data are described in Table 8.2.

```
# Command to install package in R interface
install.packages("pacakagename", repos = http://cran.us.r-project.
org", type = "source")
# repos is the repository url address which contain the desired
package
# type corresponds to installing source package
 OR
install.packages("packagename") # Can be directly downloaded from
the following command
```

Besides, there are packages that can be implemented in phylogenetic analysis, data manipulation, normalization, and population genetics. These packages are:

- APE (analysis of phylogenetic and evolutionary data): Reading and plotting of phylogenetic tree is brought about by the functions provided in the package. It provides functions for reading and plotting phylogenetic trees in Newick format. It helps in comparing the nucleotide or protein sequences in phylogenetic



	logFC	t	P.Value	adj.P.Val
TRPC4AP	-1.433239	-16.87079	2.813141e-51	4.879956e-47
PPP2CA	-1.717536	-15.59639	2.742516e-45	2.378721e-41
GGNBP2	-1.373364	-15.37906	2.776380e-44	1.605396e-40
ALKBH5	-1.647545	-15.31553	5.449793e-44	2.363439e-40
VPS53	-2.126897	-15.28890	7.228365e-44	2.507809e-40
USP7	-1.534415	-15.20816	1.699692e-43	4.914094e-40
DNAJA2	-1.500089	-15.18051	2.277232e-43	5.132873e-40
UBE2Z	-1.803240	-15.17684	2.367152e-43	5.132873e-40
KCMF1	-1.446839	-15.08269	6.397547e-43	1.233092e-39
C16orf70	-2.266052	-14.91806	3.617920e-42	6.276006e-39
RNF34	-1.452512	-14.75317	2.036289e-41	3.211228e-38
MARK2	-1.889201	-14.69839	3.609328e-41	5.217584e-38
ARF1	-1.674176	-14.63591	6.925823e-41	8.818228e-38
IMMT	-1.569139	-14.63330	7.116803e-41	8.818228e-38
RDH14	-1.643324	-14.51681	2.391266e-40	2.641036e-37
AP2A2	-2.406477	-14.51502	2.435959e-40	2.641036e-37
UBE2D2	-1.380943	-14.45135	4.716344e-40	4.812613e-37
COG1	-2.114762	-14.33219	1.618863e-39	1.560134e-36
SLC25A44	-2.081699	-14.28488	2.638314e-39	2.408780e-36

Fig. 8.4 Displaying top 18 differentially expressed genes which details about log fold change, *t*-statistics, significance in terms of *p*-value, and adjusted *p*-value

framework. The comparison of data leads to the identification of macro and micro evolution of species. The phylogenetic distance determines the era of evolution.

- BIM (Bayesian mapping of intervals): Functions are used to interpret quantitative trait loci Bayesian mapping (Satagopan et al. 1996).
- Bqtl: It is a QTL (quantitative trait loci) mapping package implemented in analysis of recombinants and inbred crossed lines. Functions are based on Bayesian and likelihood tools.
- Genetics: Classes and modules are being employed for handling genetic data. Variable classes are being implemented to identify genotype and haplotype markers on chromosomes. Different functions are being implemented in calculating allelic frequency-based Hardy-Weinberg equation and estimating linkage disequilibrium.

Table 8.2 Packages in R for genome-wide study

Data concept	Packages
Input/output	ShortRead (fastQ), rtracklayer (wig, bed, gff), GenomicRanges, Rsamtools (bam), BSgenome, Biostrings, VariantAnnotation (vcf), readxl, Google Sheets, MonetDBLite
Alignment	DECIPHER, MSA, bios2mds, seqinR, Rsubread, GraphAlignment, R-Coffee, GitHub, gmapR
Annotation	AnnotationDbi, VariantAnnotation, TxDb.*, annotate, NLP, biomaRt, Annotables, GOsummaries, GenomicFeatures GitHub, dcGOR, ChIPpeakAnno, AnnotationHub
Visualization	iPlots, ggvis, ggplot, visualize, quantmod, dygraphs, googleVis, metricsgraphics, RColorBrewer, shiny, flexdashboard, rcdimple, plotly
Quality assessment	DatABEL, GenABEL, MetABEL,MixABEL ParallABEL, PredictABEL, ProbABEL RepeatABEL, VariABEL, OmicABEL
DNA methylation	methylPipe, BiSeq, bsseq, ChAMP, COHCAP, comet, DMRcate, DMRforPairs, lumi, MassArray, methyAnalysis, methylumi
ChIP-seq	ChIPpeakAnno, motifStack, rGADEM, ChIPXpress, les, iChip, Starr, MotIV
RNA-seq	ArrayExpressHTS, dcGSA, DEGseq, DEXSeq, DER Finder, easyRNASeq, globalSeq, metaSeq, NOISeq, rnaSeqMap, SeqGSEA, sSeq, subSeq, transcript
SNPs	BBCAnalyzer, crlmm, beadarraySNP, deepSNV, GMRP, logicFS, GWASTools, oligo, SNPchip, trio, snpStats, SeqVarTools
Exon array	Limma, puma, oligo, TIN, siggenes, PECA, snm SELEX, motifRG, BCRANK, TFBSTools
Motifs	CNAnorm, CrispRVariants, myvariant, CINdex, myvariant, SomaticSignatures
Genomic variation	BasicSTARRseq, geneplast, GSReg, ISoLDE, tRanslatome, IVAS, InPAS
Gene regulation	digitmetagenomeSeq, phyloseq, mmnet, rRDP
Microbiome	ABarray, affyPara, AnnotationForge, beadarraySNP, bioCancer, BioQC, coMET,
Workflow	DNABarcodes, flowcatchR, flowClust, flowMap, GSVA, ideogram, Heatplus, LBE

- hapassoc: Functions are being used in inferring trait associated with haplotypes and simultaneously other co-variables in linear model. Functions can also identify uncertain haplotype and missing genotype for few SNPs (Burkett et al. 2004).
- haplo.score: It comprises of routines and functions to assess the haplotypes in wide variety of traits including ordinal, binary, and quantitative, based upon score (Schaid et al. 2002). The analysis is based upon the assumption that all the variables are unrelated and haplotype is equivocal. It envisages on global haplotype pattern, and the significance is computed based upon the p-values.
- haplo.stats: It contains multiple S-plus/R functions to evaluate the presence of haplotypes indirectly. The statistical analysis is based on assumption that all the subjects are discrete and the haplotypes are indistinct (Lake 2003). The genetic marker identified is anticipated to exhibit codominance. The important functions recognized in haplo.stats are halpo.score, halplo.gem, and haplo.em.

- hierfstat: The function is implemented in assessment of hierarchical F-value in the genetic haploid and diploid set of data, and level of significance is computed.
- hwde: It constitutes the model that fits for genotypic disequilibrium, and the analysis accounts for the interaction between the loci of first order (Weir and Wilson 1986).
- Kinship: The package constitutes variable functions, mainly Coxme, which is implanted to analyse the data based upon cox proportional hazard model. Routines are used to create n by n matrices which define the genetic relationship amid of two individuals, pedigree, which provides functions for creating the pedigree plots, and bdsmatrix, which constitutes number of classes for block diagonal matrices.
- IdDesign: This package is built to identify the linkage disequilibrium. The presence of biallelic quantitative trait loci (QTL) and the markers are detected based on deterministic power calculation together with Bayes factor (Sen et al. 2005).
- LDheatmap: Package has function to generate heat map associated with linkage disequilibrium for particular SNP.
- PHLOGR: Various functions are implemented for analysis and manipulation of phylogenetic dataset.
- qtlDesign: Contains functions to determine QTL experiments.
- R/gap: Constitute integrated package for analysis of genetic data for both family and human population. It holds several functions for calculation of sample size, probability of having particular disease within a family or in population, calculation of kinship, statistical linkage analysis, and association between the genetic markers. Some of the functions identified are hwe.hardy for Hardy-Weinberg analysis involving SNPs and polymorphic satellites; s2k for analysis of single locus associated with genomic control and polymorphic markers; gene counting; gcp for interpretation of haplotype for all chromosomes (Zhao 2004); kbyl, tbyt for computing linkage disequilibrium associated with SNPs and multi-allelic markers; htr for haplotype extraction based upon regression analysis; and kin. morgan for simple calculation of kinship.
- Rmetasim: It provides an interface between metasim and R. It helps in building individual-based population genetics using metasim simulation.

R provides an ease to data handling as it incorporates multiple packages having an application in research, neuroimaging, disease mapping, and social network analysis. Further, multiple functions in packages can serve in variety of data analysis including determination of haplotype frequencies, assigning of probable haplotypes, and heat map for linkage disequilibrium. Most standard feature of R in data management, graphics, and statistical analysis renders it to be valuable for microarray and next-generation sequencing data analysis. Besides, packages are also enriched in phylogenetic analysis, population studies, quantitative trait analysis, and QTL mapping in human pedigrees. However, ease of drafting packages from the available code will not obstruct the researcher in subsequent analysis of complex data. In summary R constitutes an integrated platform for genome-wide data analysis. It will be rewarding for theoretical and applied scientist in software development for long term.

Acknowledgements Authors Arunima Shilpi and Shraddha Dubey are thankful to the Department of Bioinformatics at NIT Rourkela and Maulana Azad National Institute of Technology, Bhopal, respectively, for providing financial support and dry lab facilities.

References

Amos CI (2007) Successful design and conduct of genome-wide association studies. Hum Mol Genet 16(2):R220–R225

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34(5):525–527

Brooks SP, Catchpole EA, Morgan BJ, Barry SC (2000) On the Bayesian analysis of ring-recovery data. Biometrics 56(3):951–956

Burkett K, McNeney B, Graham J (2004) A note on inference of trait associations with SNP haplotypes and other attributes in generalized linear models. Hum Hered 57(4):200–206

Clayton D, Leung HT (2007) An R package for analysis of whole-genome association studies. Hum Hered 64(1):45–51

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10): R80

Guo SW, Lange K (2000) Genetic mapping of complex traits: promises, problems, and prospects. Theor Popul Biol 57(1):1–11

Hardcastle TJ, Kelly KA (2010) BaySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11:422

Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ (2003) Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 55(1):56–65

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, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. Nature 409(6822):860–921

Law CW, Chen Y, Shi W, Smyth GK (2014) Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15(2):R29

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29(8):1035–1043

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12(1):323

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550

O'Neill RJ, O'Neill MJ, Graves JA (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature 393 (6680):68–72

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14(4):417–419

Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol 32(5):462–464

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140

Sanders NW, Mann NH 3rd, Spengler DM (1997) Web client and ODBC access to legacy database information: a low cost approach. Proc AMIA Annu Fall Symp:799–803

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) A bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144(2):805–816

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70 (2):425–434

Sen S, Satagopan JM, Churchill GA (2005) Quantitative trait locus study design from an information perspective. Genetics 170(1):447–464

Weir BS, Wilson SR (1986) Log-linear models for linked loci. Biometrics 42(3):665–670

Zhao JH (2004) 2LD, GENECOUNTING and HAP: computer programs for linkage disequilibrium analysis. Bioinformatics 20(8):1325–1326

Zhao JH, Tan Q (2006) Integrated analysis of genetic data with R. Hum Genomics 2(4):258–265

Computational Approaches to Studying Molecular Phylogenetics

9

Benu Atri and Olivier Lichtarge

Even if we didn't have a single fossil, the evidence for evolution would be absolutely secure because of comparative anatomy, comparative biochemistry, and geographical distribution

Dr. Richard Dawkins (The Blind Watchmaker)

9.1 Introduction

Molecular phylogenetics is the study of evolutionary history, development, and relationships among organisms using molecular sequence or structure data (DNA, RNA, or proteins). The premise of any phylogenetic analysis is the hypothesis that the two organisms or sequences are evolutionarily related. Phylogenetics involves representation of said relationships in the form of branches of a tree. Features including the location of a species and the length of the branches from a given node to the end depend on the information obtained from methodically comparing sequences and hence making inferences about the relatedness of the sequences to each other and a common ancestor.

In this chapter, we show how one can use sequence alignment information to visualize evolutionary relationships between organisms in the form of a *phylogenetic tree*. We will briefly discuss the common terminology used in molecular

B. Atri (✉)
Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX, USA

O. Lichtarge
Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX, USA

Center for Computational and Integrative Biomedical Research (CIBR), Baylor College of Medicine, Houston, TX, USA

Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_9

phylogenetics, the most common tree-building methods describing in detail character-based, distance-based, and Bayesian inference methods and how to decide which one to use, computing the accuracy of phylogenetic inferences statistically, i.e., to assess the level of confidence, one has in an inferred tree through methods like bootstrapping and jackknifing. Next, we consider the underlying assumptions of various models of evolution that different tree-building algorithms use, and finally, we list a number of popular tools and briefly go over a couple of applications of phylogenetic analyses.

9.1.1 Terminology

A *tree* is a 2D graphical structure used to model the evolutionary relationships between groups of organisms and sequences represented by nodes connected to each other by branches. *Rooted trees* have a common ancestor represented by a single lineage at the base that connects all the branches of the tree. The common node is called the *root*. On the other hand, *unrooted trees* depict relationships among organisms but do not have a common ancestor and do not require the knowledge of an ancestral root.

A *rooted bifurcating* tree has exactly two descendants arising from each ancestral node (interior node), i.e., it is a *binary tree*. An *unrooted bifurcating* tree is also binary with three neighbors at each node. *Multifurcating* refers to more than two daughter branches. A *labeled tree* has values assigned to the nodes, whereas an *unlabeled tree* is simply the overall topology of the tree.

A *clade* is a group of organisms that include all descendants of a common ancestor, for example, birds or insects. A *cladogram* is a tree generated using cladistics. Cladistics refers to the biological classification of categorizing organisms based on shared characteristics that can be traced to the group's most recent ancestor, which are not present in further distant ancestors. A *dendrogram* is a tree diagram used to illustrate clustering.

A reference group used to determine evolutionary relationships among three or more monophyletic groups of organisms is known as an outgroup. It is chosen based on whether it is closely related to the other groups but not more than any single one of the groups to each other. *Operational taxonomic units* (OTUs) are terminal branches represented by sequences for which molecular sequence data or structure is known.

Branch length is the estimated amount of evolutionary time taken from a node to the end. A tree with branch lengths indicated is known as a *weighted tree*.

Phylogenetics aims to classify all organisms by making groups and subgroups of those that share biological information with a common ancestor thereby allowing us to reconstruct ancient evolutionary events as well as relationships and perhaps missing links with a level of certainty. The most common phylogenetic tree-building methods are either (1) character or tree-searching based or (2) algorithmic or distance based (Fig. 9.1). Moreover, the criteria to choose the most suitable phylogenetic tree-building method are shown in Fig. 9.2.

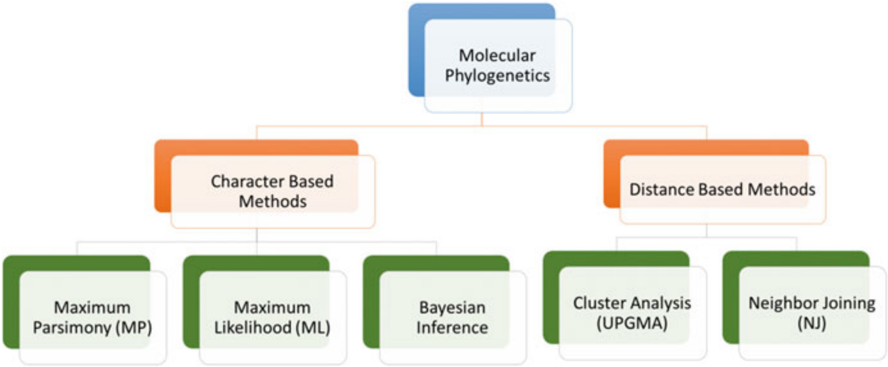


Fig. 9.1 Tree-building algorithms

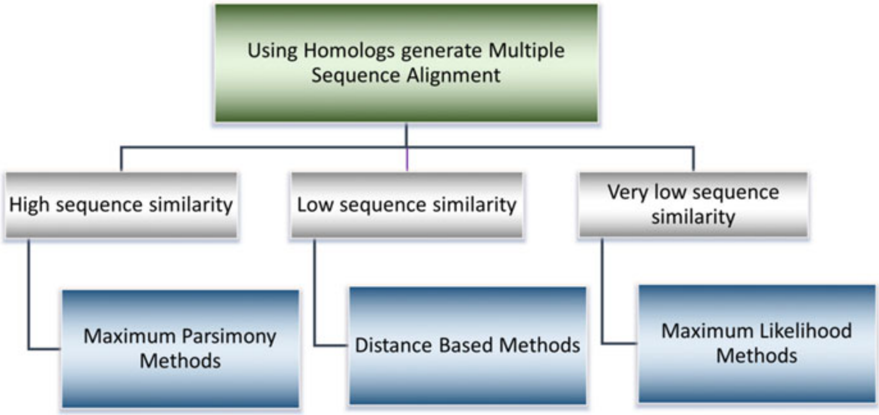

Using Homologs generate Multiple Sequence Alignment
High sequence similarity
Low sequence similarity
Very low sequence similarity
Maximum Parsimony Methods
Distance Based Methods
Maximum Likelihood Methods

Fig. 9.2 Choosing the most suitable phylogenetic tree-building method will ultimately depend upon the quality of sequences and the degree of similarity between the homologs. (Adapted from David W. Mount 2004)

9.1.2 Character-Based Methods

Character-based methods rely on character states, which refer to an alignment of molecular sequences – nucleotides or proteins. Three ways of character-based phylogenetic analysis are maximum parsimony, maximum likelihood, and Bayesian inference.

9.1.2.1 Maximum Parsimony (MP) or Minimum Evolution Method

Edwards and Cavalli-Sforza (1964) proposed to analyze gene frequency data and subsequently applied to sequence data and phylogenetics by Fitch (1971) and Felsenstein (1983), among others; this method attempts to reduce the number of steps needed to explain the observed data. It states that the most preferred tree is one with the least number of steps (most parsimonious) or character changes. A step can be defined as the change from one-character state to another. Character state, in the context of molecular biology, refers to the four nucleotide bases, A, T, C, and G (for DNA) or the 20 amino acids (for proteins).

MP defines a function to score an input tree topology. All trees are scores one by one, by determining how many steps were taken to obtain the distribution of each character in a tree. The characters are evaluated column by column. The trees are then ranked and compared to each other based on the set optimality criterion. The shortest or the most parsimonious tree for the given problem represents the preferred relationships among the sequences that explain the divergence pattern among sequences.

This method involves two steps:

1. Calculating the amount of character change required by a tree.
2. Searching all the returned tree topologies for one with the minimum length. If the number of trees returned is large, the search can become computationally expensive.

Trees resulting from MP are usually unrooted. While one can infer relationships, there is no way to infer the time of divergence. For that, choosing an outgroup (a branch taken outside of the tree) will provide a relative understanding of time. The choice of an outgroup is very crucial, and an incorrect choice could result in a tree with incorrect relationships between the sequences.

Drawbacks

1. Does not always estimate the most statistically significant tree.
2. Could lead to a large number of tree topologies, thereby, slowing the method.
3. Often computes out many equally most parsimonious trees (MPTs), and a large number of such trees are considered as a failure of the algorithm. Inherently, the method lacks any information on how sensitive these conclusions are. Using statistical resampling procedures such as jackknifing or bootstrapping, uncertainties can be quantified.

9.1.2.2 Maximum Likelihood (ML)

This method uses probabilities to find a phylogenetic tree that can best explain the observed divergence patterns (Felsenstein 1973, 1983). It is similar to MP in a way that ML algorithm also moves along the alignment, one column at a time. Since all possible trees are evaluated, this method works well with smaller datasets. Like MP, ML also considers the fewest steps, i.e., for each of the possible trees, the number of mutations needed to achieve the observed patterns of divergence in the sequences is considered, and the more mutations needed to get a given tree, the less likely it is the true tree. That is because the rate of mutations is generally small.

However, the uniqueness of the ML method lies in the fact that it allows one to incorporate evolutionary models like Jukes-Cantor69 or Kimura models (see Sect.

9.1.5). These models explain the variations in nucleotides; therefore, provide ML the additional power to analyze sequence datasets with more divergences.

Steps

1. Choosing an evolutionary model to provide the probability/rates of nucleotide substitutions (transitions/transversions).
2. Sequence alignment.
3. Within each column, the substitutions are evaluated to how well they fit a set of trees describing the phylogenetic relationships among the sequences. Based on the succession of mutations needed to reach the given sequence data, the trees are given a *likelihood score*, the product of the rate of substitutions in each branch of the tree (which in turn is calculated as the product of substitution rate in a branch to its branch length). Because each tree can have multiple substitution possibilities, usually a combined probability (sum of the probability of each set of changes) is considered.
4. Sequentially, all columns are analyzed. Column likelihoods are multiplied to obtain the *tree likelihood score*.
5. The tree with the highest likelihood score (hence the name maximum likelihood) is identified. Usually, these scores are reported in terms of *log likelihood*.

Such a methodical manner of scoring each column within several trees can quickly become computationally expensive, so first a small subset of trees is evaluated and eventually more trees are added to the set. Several random iterations must be completed, and a flexible rate of evolution must be considered to avoid any biases. While ML is a computationally expensive method, lately, it is less of an issue due to the current advancements in computational power.

9.1.2.3 Bayesian Inference Method

This method employs a likelihood function to generate a posterior probability of trees based on a prior probability evolutionary model producing the best-fit tree for a given data using Bayes theorem (Huelsenbeck et al. 2001). It can be represented as follows:

$$\Pr\left[\text{Tree} \mid \text{Data}\right] = \frac{\Pr\left[\text{Data} \mid \text{Tree}\right] \times \Pr\left[\text{Tree}\right]}{\Pr\left[\text{Data}\right]} \tag{9.1}$$

Here, combining the prior probability of a phylogeny Pr [Tree] with the likelihood Pr [Data l Tree] or the posterior probability of the tree could be used as a proxy for the correctness of the tree. That is to say that the higher the posterior probability, the better the estimated tree is.

In order to computationally approximate the posterior probability, the most useful and popular method is the Markov Chain Monte Carlo (MCMC) method. The incorporation of MCMC algorithm and increase in computing speeds have made Bayesian methods very popular. The idea behind MCMC is to construct a Markov chain, which is a sequence or chain of random samples taken throughout the parameter space at specified intervals. MCMC allows sampling based on posterior probabilities.

Steps

1. Create a new tree by randomly perturbing the current tree.
2. Either accepts the new tree or rejects it based on the probability using MCMC algorithm. The Metropolis-Hastings method is a widely used MCMC algorithm that allows random sampling from a dataset with complicated and multidimensional (multi-parametric) distribution probabilities (Metropolis 1953; Hastings 1970).
3. If accepted, repeat steps 1 and 2.

Since the results of a Bayesian approach are directly related to the evolutionary model, in order to avoid incorrect phylogeny, it is important to choose a model that fits the observed data well.

Advantages

1. High computational efficiency.
2. Complicated datasets and models of evolution can be handled with ease.
3. It is considered a better method because it gives the probability of the estimated tree in the context of the given problem/dataset.

Mr. Bayes, a computer program, is the most commonly used implementation of the Bayesian method (Huelsenbeck and Ronquist 2001, 2003). The software uses both MCMC and MCMC coupled with Metropolis-Hastings algorithms for performing the Bayesian method for phylogenetic inference.

Another implementation of the Bayesian phylogenetic inference method is the open-source software Phycas (Lewis et al. 2015), which uses marginal likelihoods and the generalized time-reversible evolutionary model. Phycas takes input in Nexus format.

9.1.3 Distance-Based Methods

Distance-based methods infer phylogenetic relationships using a matrix of genetic or evolutionary distances between the sequences, and then a tree is fit to those relationships. Evolutionary distance measures sequence divergences, which occur due to the acquisition of mutations over the course of evolution.

Easiest ways to calculate divergence is to count the number of mismatches between two aligned sequences also known as the observed distance (p). However, this is not necessarily the best measure, especially if there is a high degree of divergence, as not all the substitutions that occurred will be revealed. That also holds true for any reversion mutations, in which case there will not be an observed difference. This issue can be fixed by using a model of evolution with rules related to evolutionary divergence. Failing to use accurate assumptions while choosing the evolutionary model could result in artifacts, false conclusions, and a tree representing wrong phylogenetic relationships. It is important to remember that the ultimate goal is to infer a tree that correctly and best fits the observed patterns of divergence.

There are two types of tree inference methods based on evolutionary distance: cluster analysis methods and neighbor-joining methods.

9.1.3.1 Cluster Analysis/Unweighted Pair Group Method with Arithmetic Mean

Sneath and Sokal (1973) developed this method to generate tree topologies for ultrametric data. Ultrametric data assume the molecular clock hypothesis, i.e., all sequences have evolved at a constant rate throughout the tree (Swofford 1996) and meet the following for taxa A, B, and C:

$$d_{\mathrm{AC}} \leq \max(\, d_{\mathrm{AB}}, d_{\mathrm{BC}}),$$

where, $d =$ evolutionary distance

As the molecular clock is assumed, all trees are rooted trees and all end or terminal nodes are equidistant from the tree root. Ultrametric data are rarely encountered with real sequences; therefore, if the molecular clock is wrongly assumed, it could be misleading. The most common cluster analysis is the unweighted pair group method with arithmetic mean (UPGMA) method that uses sequential or agglomerative clustering. A tree is built sequentially by grouping the sequences in a pairwise manner, starting from the most similar pair, with the lowest value of the genetic distance. The input to this method is a matrix of genetic/evolutionary distances, which captures the divergences between the sequences.

Steps

1. Find the shortest pairwise distance.
2. Join the two sequences with the shortest distance.
3. Calculate the depth of the new branch using $= \frac{1}{2}$ shortest distance.
4. Tip to tip path length $=$ the shortest distance.
5. Calculate mean pairwise distance with the remaining sequences and create a new matrix.
6. All tip-to-tip distances via the root will have the same total distance equal to the total sum.

This method has some inherent limitations. For example, UPGMA is only responsive to equal evolutionary rates. If one of the sequences has acquired more mutations over time, we might obtain a tree with erroneous topology using this method. Additionally, for non-ultrametric distances, one needs to correct for unequal rates of mutation by comparing with a reference sequence, also sometimes known as an outgroup.

The UPGMA weighs each sequence equally; the averaging of distances depends on the total number of items in the cluster/group. For example, A, B (grouped), and C are grouped into a new node ABC. Then the distance of ABC to any other node (D or E) is calculated as

$$d_{\mathrm{ABC\_D}} = \frac{N_{\mathrm{AB}}\, d_{\mathrm{ABD}} + N_{\mathrm{c}}\, d_{\mathrm{CD}}}{N_{\mathrm{AB}} + N_{\mathrm{c}}} \tag{9.2}$$

where, N_{AB} = number of items/units in the cluster AB (two in this case) and

$$N_{\mathrm{c}} = 1$$

A version of UPGMA called the weighted pair group method with arithmetic mean or WPGMA weighs the member most recently added to the group the same as all the previous members of the group. The averaging of distances is not based on the total number of items in the groups, and for the example above, the distance of ABC to any other node (D or E) is

$$d_{\mathrm{ABC\_D}} = \frac{d_{\mathrm{ABD}} + d_{\mathrm{CD}}}{2} \tag{9.3}$$

When the data are ultrametric, UPGMA and WPGMA give the same result.

We must remember that any phylogenetic analyses and all inferences derived from such analyses are only postulations, and enough evidence must be collected to support such hypotheses. One must be careful of pitfalls like the fact that while most of the biological information are passed from parents to the offspring in a vertical fashion, there could also be the possibility of horizontal gene transfer, DNA transformation, transposon-mediated shuffling, etc.

9.1.3.2 Neighbor Joining

Neighbor joining (NJ) involves bottom-up clustering and is one of the most commonly used tree inference methods (Saitou and Nei 1987). The NJ method employs the star decomposition method. The input is a distance matrix and an initial star network tree. The input data does not have to be ultrametric.

Steps

1. Calculate matrix M, based on the distance matrix d, given by

$$\mathrm{M}_{\mathrm{s1,s2}} = (n-2)d_{\mathrm{s1,s2}} - \sum_{k=1}^{n} d_{\mathrm{s1},\ k} - \sum_{k=1}^{n} d_{\mathrm{s2},\ k} \tag{9.4}$$

where,

$d_{\mathrm{s1,\ s2}}$ is the distance between sequences $\mathrm{s}_1$and s_2
and n is the total number of sequences.

2. Find the pair of sequences (s1, s2) for which $M_{s1,\ s2}$ is the smallest.
3. Connect this pair to a node that is connected to the central node. To calculate the distance of each of pair of sequences (s1, s2) to the new node, n_0, use

for s_1:

$$d_{s1,n_0} = \frac{d_{s1,s2}}{2} + \frac{1}{2\,(n-2)}\left[\sum_{k=1}^{n} d_{s1,k} - \sum_{k=1}^{n} d_{s2,k}\right] \tag{9.5}$$

for s_2:

$$d_{s2,n_0} = d_{s1,s2} - d_{s1,n_0} \tag{9.6}$$

where, s_1 and s_2 are the paired sequences in step 2 and n_0 is the new node.

4. Calculate the distance from each of the remaining sequences to the new node n_0, using

$$d_{n_0,k} = \frac{d_{s1,s2}}{2} + d_{s2,k} - d_{s1,s2} \tag{9.7}$$

where, n_0 = new node

k = sequence or node we want to find the distance of from node n_0
s_1, s_2 = pair joined in step 2

5. Replace the pair (s1, s2) with new node n_0. This pruning is done to change the new node n_0 to a terminal node and to get a smaller tree.
6. Repeat the calculations as given in steps 1 through 4; the algorithm is repeated until left with two nodes connected with one branch.

Advantages
Fast and suited for large datasets, permits a variable evolution rate among the sequences, considers the possibility of multiple substitutions, and allows for corrections

Disadvantages
The result is one possible tree and depends on the assumed evolutionary model.

- It is worth restating that the neighbor-joining method does not assume the molecular clock hypothesis and generates an unrooted tree, while the UPGMA method assumes the molecular clock hypothesis and generates a rooted tree.

9.1.4 Statistically Computing Accuracy of Inferences

It is important to question the level of confidence one has in the inferred tree. The jackknife and bootstrap are statistical techniques for empirically approximating the variability of any estimate. They differ but are of the same family of methods. A third resampling technique is permutation testing.

9.1.4.1 Bootstrapping

Bootstrapping (Efron, Halloran and Holmes 1996) is one of the most popular resampling procedures used to assess the reliability of branches of a phylogenetic tree. Felsenstein first applied the bootstrapping methodology to phylogenetic analyses (Felsenstein 1985). It can provide the confidence for each unit or taxon of the tree in a two-step process:

Steps

1. Generate new datasets (several) by sampling columns of characters from the original dataset at random with replacement, to ensure the sampling of each column with equal probability. Each new dataset is of the same size (number of columns) as the original, but the actual columns are sampled in different ways than the original. The process of resampling and tree construction is repeated several times (100–1000), and the percentage of times a branch is given a value of a certain score is noted.
2. Compute the relative number of times a given branch appears in the tree. This number is called the *bootstrapping value*, which is a measure of the accuracy of the inferred tree branch, suggesting how close it is to representing true evolutionary relationships.

Hillis and Bull (1993) suggested that under favorable conditions such as equal rates of change, and more or less symmetric branches, bootstrapping values of 70% can be considered equal to 95% probability that the inferred tree is the true tree. While under unfavorable conditions, bootstrapping of more than 50% could mean a higher chance of false positives, i.e., giving a higher score to a wrong tree. One must, therefore, be careful to evaluate the conditions under which the tree is generated and the assumptions (e.g. molecular clock) that were met by the original data.

Problems Encountered Sites may not evolve independently, sites may not come from a common distribution (but can consider them sampled from a mixture of possible distributions), and bootstrapping does not correct biases in phylogeny methods

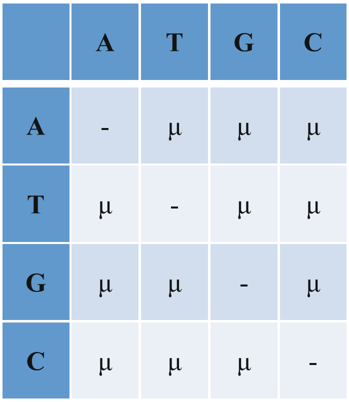
9.1.4.2 Jackknifing

Another method for estimating support was first used by Mueller and Ayala in 1982 for phylogenetic analyses. The jackknife, which is the older of the two, involves dropping one observation at a time from one's sample and calculating the estimate each time. It is very similar to bootstrapping, but jackknifing does not resample the data. Instead, it uses subsets of the data, which is referred to as resampling without replacement. The goal is to create a smaller dataset to see if removing a subset changes the tree in large part to estimate its influence. Jackknifing is less commonly used than bootstrapping.

9.1.5 Models of Evolution

Most of the phylogenetic analyses methods require some inference or probability-based statistical models of how DNA or protein sequences evolve. Such *evolutionary models* compute the probability of nucleotide or amino acid change as well as correct for changes during evolution. Evolution of sequences happens on a variety of timescales; therefore, it is easier to define these models regarding the rate of change between the different states (substitutions). An evolutionary model is of great usefulness when it can fit the data well and provides accurate predictions. When conducting phylogenetic analyses on different data (e.g., coding and noncoding sequences), one must make sure that the model fits all of the data well, assuming different amounts of selection pressure on different types of sequences. The choice of which model to use depends on the underlying assumptions of the model, and a wrongly chosen model could lead to underestimated variations and incorrect branch lengths.

JC69 Model (Jukes and Cantor 1969) This model is the simplest and most restrictive model that assumes equal rates of substitution between all nucleotides. This model has only one parameter μ, the rate of change. For example, in the case of nucleotides, a 4X4 matrix can be created with the rates of nucleotide substitutions (Fig. 9.3).



	A	T	G	C
A	-	μ	μ	μ
T	μ	-	μ	μ
G	μ	μ	-	μ
C	μ	μ	μ	-

Fig. 9.3 JC69 model assumes equal rate of substitution in nucleotides

K-80 Model (Kimura 1980) This consists of a slightly more complex combination of two rates (or parameters). In the case of nucleotides, these two can be classified as rates of transitions and transversions. It makes sense to separate these two rates since transitions occur more frequently than transversions.

F-81 Model (Felsenstein 1981) It is an extension of the JC69 model. It assumes all substitutions to be equal, but base frequencies can vary from 0.25 ($A \neq C \neq T \neq G$). Base frequencies are the values for the rate (n) by which any nucleotide i changes to nucleotide *j*.

HKY-85 Model (Hasegawa et al. 1985) This combines the K80 and F81 models and assumes different rates of substitutions between each nucleotide (base frequency) and also different rates for transitions and transversions.

TN-93 Model (Tamura and Nei 1993) This includes new parameters that consider different rates of substitutions for pyrimidines and purines (two types of transitions).

Generalized Time-Reversible Model (GTR Model) (Tavare 1986) It is the most general, independent, and time-reversible model. It assumes six classes of substitutions and base frequencies are not the same.

9.2 Phylogenetic Tools

There are a large number of online tools available to construct phylogenetic trees. Phylogeny.fr (www.phylogeny.fr/) provides a robust tool for nonspecialists for reconstructing and analyzing phylogenetic relationships between molecular sequences (Dereeper et al. 2008, 2010).

MEGA (http://www.megasoftware.net/) is an integrated tool for conducting sequence alignment, inferring phylogenetic trees, estimating divergence times, mining online databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses (Kumar et al. 1994; Tamura et al. 2013).

PHYLogeny Inference Package (PHYLIP) is a free package of programs for inferring phylogenies (Felsenstein 1989, 2013; Revell 2013). It is distributed as source code, documentation, and different types of executable files. PHYLIP is available at http://evolution.genetics.washington.edu/phylip.html. PHYLIP includes methods for parsimony, distance matrix, and likelihood methods and also performs bootstrapping and consensus tree generation.

Phylogenetic Analysis Using Parsimony (PAUP) is a commercially available software for inferring phylogenies (Swofford 1991) from discrete character data using maximum parsimony, i.e., searches for minimum-length trees resulting in trees that minimize the amount of evolutionary steps needed to explain the divergence patterns in the available data. PAUP (http://paup.csit.fsu.edu/) includes maximum likelihood and distance methods as well.

MacClade is another popular downloadable phylogenetic tool which provides an interactive view to visualize phylogenies (Maddison and Maddison 1999). MacClade data editor has several editing and visualization options. It is available at http://macclade.org/macclade.html

A more comprehensive list of available tools can be found at

http://molbiol-tools.ca/Phylogeny.htm and http://evolution.genetics.washington.edu/phylip/software.html.

9.3 Evolutionary Trace (ET) and Evolutionary Action

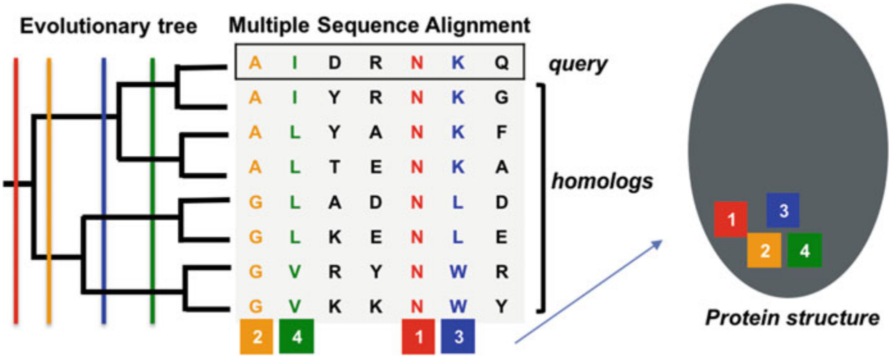
To guide mutagenesis experiments (and other studies), phylogenetics can provide a reliable and cohesive computational analysis of the protein evolution by analyzing the proteins' sequences, structures, and functions. One such phylogenetic strategy is called the evolutionary trace (Lichtarge et al. 1996), which is based on two features of protein's functional domains:

A high level of identity is seen among amino acids that share a common ancestor, and

Under evolutionary pressure, domains that are functionally important for any protein will resist new mutations and try to maintain function and integrity.

Analysis For a given protein, ET uses an alignment of homologs from sequence databases and generates a multiple sequence alignment. The output is the whole alignment and the phylogenetic tree, and most importantly, each residue is assigned a rank or a score based on its functional importance. This rank comes from the columns (residue positions) in the multiple sequence alignment by tracing whether variations in a particular residue during evolution show a relationship with large or small divergences among orthologs and paralogs (given by the tree). These ranks reflect the functional importance of that residue in the overall protein. From these ranks, we can formulate hypotheses on the molecular determinants of activity and specificity and rationally target experiments to the most relevant sites of the protein. The concept is illustrated in Fig. 9.4. The first trace is computed with the entire family in one group (red). The second trace is done with the family divided into two classes defined by the first two branches of the tree (orange). The third trace is done with the family split into the three groups defined by the first three branches of the tree (blue). This is repeated up until the family is divided into N classes, where N is the total number of sequences (green). Thus, each residue eventually becomes class specific; some do so when the division is into fewer branches than others that need finer division. A residue's evolutionary rank, therefore, can be defined as the minimum number of branches into which it is necessary to divide the family for this residue to become class specific.

Applications Evolutionary trace is a well-validated method to identify functional sites and their residue determinants in proteins. ET ranks residues by relative



A	I	D	R	N	K	Q
A	I	Y	R	N	K	G
A	L	Y	A	N	K	F
A	L	T	E	N	K	A
G	L	A	D	N	L	D
G	L	K	E	N	L	E
G	V	R	Y	N	W	R
G	V	K	K	N	W	Y
2	4			1	3	

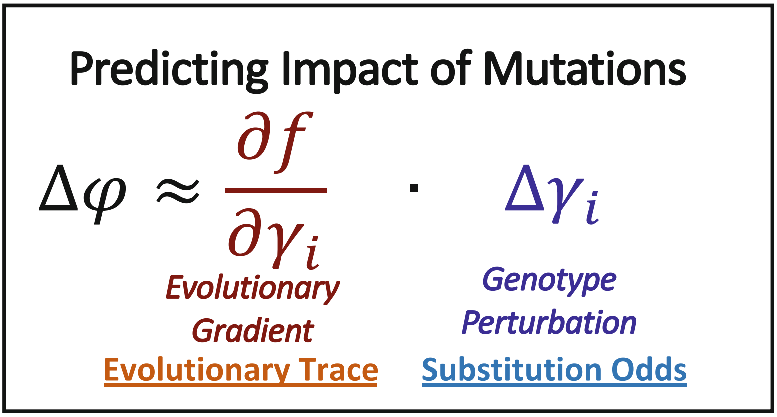
Fig. 9.4 Evolutionary Trace (ET) utilizes a tree-based approach to narrow the analysis to subfamilies within which the sequence similarity is higher. ET is freely available for use at http://mammoth.bcm.tmc.edu/

functional importance based on the variation observed in the phylogenetic tree, and top-ranked residues are seen clustered on the protein's structure revealing known (Madabushi et al. 2002) or putative functional sites (Wilkins et al. 2010).

ET is also a reliable predictor for ligand specificity determinants in allosteric pathways (e.g., pharmacologically important class of G-protein-coupled receptors (GPCRs)). A study by Rodriguez et al. (2010) used a version of ET called Difference-ET which compares ET ranks between functionally distinct branches of the phylogenetic tree. It guided experimental residue swapping to successfully pinpoint ligand specificity determinant residues of two GPCRs, which despite highly similar functions and nearly identical binding site structures operate differently on their respective ligands. These applications make ET a powerful tool for an efficient redesign of protein function and drug targeting. Another use of ET is the Evolutionary Trace Annotation (ETA) Server that predicts enzymatic activity by using functional residue positioning to generate a 3D template or motif. All available annotated structures are searched for matches with the template. Positive matches have been experimentally shown to be predictive of molecular and functional similarity (Ward et al. 2009).

ET has also been shown to be a valuable tool for peptide design revealing evolutionarily relevant domains (Sanae Shoji-Kawata et al. 2009). Moreover, ET helps in narrowing down our choices before mutational analysis, assists in determining the separation of function residues, and provides drug targets to combat antibiotic resistance (Adikesavan et al. 2011).

Finally, a recent application of ET is the quantification of genotypic variations (amino acid mutations in coding regions) to predict their phenotypic impact. This mutational impact prediction tool called as evolutionary action (action) computes the effects of mutations as Action scores reflecting predicted phenotype. It is hypothesized that the phenotypic impact can be computed as the product of evolutionary gradient to the genotypic change. In order to compute the score, the



Predicting Impact of Mutations

$$\Delta\varphi \approx \frac{\partial f}{\partial \gamma_i} \cdot \Delta\gamma_i$$

Fig. 9.5 Action equation computes the phenotypic impact of mutations as the product of evolutionary gradient to the genotype perturbation. The gradient is measured using evolutionary trace ranks of functional importance of each residue and the genotypic variation is measured using substitution odds. Together, the terms of this equation are used to calculate the substitution impact scores or the action scores for all residues of a protein

evolutionary gradient is approximated using ET ranks for residue functional importance and genotypic perturbation is approximated using substitution odds, defined as the rate of amino acid change (Fig. 9.5) (Katsonis and Lichtarge 2014). Action can optionally incorporate structural features as well.

Action estimates phenotypic impact robustly and on a large scale by predicting scores for all possible substitutions at each residue position in a protein.

This predictive power of Action can help in guiding mutational studies of protein function, interpreting the numerous polymorphisms data revealed by exome sequencing, and importantly, distinguishing disease causing from harmless mutations since these variations can be powerful indicators of diseases, especially in a clinical setting.

ET-Related Software and Links

- PyETV (Lua and Lichtarge 2010) is a freely available PYMOL (Delano 2002) plugin to identify the protein's functional determinants and functional sites (http://mammoth.bcm.tmc.edu/traceview/HelpDocs/PyETVHelp/pyInstructions.html).
- Universal Evolutionary Trace (UET at http://mammoth.bcm.tmc.edu/uet/) is a database with pre-computed ET analyses for protein structures and sequences (Lua et al. 2015).
- ETA is available at http://mammoth.bcm.tmc.edu/eta/.
- The EA server is accessible at http://mammoth.bcm.tmc.edu/EvolutionaryAction.

As a concluding remark, it is imperative to say that the primary objective of molecular phylogenetics is to approximate amount of divergence between sequences (DNA or protein) and reconstruct ancient evolutionary events as well as the phylogenetic tree, which represents the probable evolutionary relationships between these sequences. The features of a phylogenetic tree will depend upon the underlying assumptions of our experimental methods and the models used to make our inferences of evolutionary relatedness. In the end, however, every researcher must remember that any phylogenetic analyses and all inferences derived from such analyses are only postulations, and attempts must be made to support such hypotheses with available data or evidences.

Acknowledgments This work is supported by a grant from the NIH Research Project Grant Program (2R01GM079656). The authors are grateful to Dr. David C. Marciano, Dr. Angela Wilkins, and Dr. Rhonald C. Lua for their helpful comments.

References

Adikesavan AK, Katsonis P, Marciano DC et al (2011) Separation of recombination and SOS response in Escherichia coli RecA suggests LexA interaction sites. PLoS Genet 7:e1002244

David WM (2004) Bioinformatics: sequence and genome analysis, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor

DeLano WL (2002) The PyMOL molecular graphics system. http://www.pymol.org

Dereeper A, Audic S, Claverie J-M, Blanc G (2010) BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evol Biol 10:8. https://doi.org/10.1186/1471-2148-10-8

Dereeper A, Guignon V, Blanc G et al (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 36:W465–W469. https://doi.org/10.1093/nar/gkn180

Edwards AWF, Cavalli-Sforza LL (1964) Reconstruction of evolutionary trees. In: Phenetic and phylogenetic classification, vol 6. Systematics Association, London, pp 67–76

Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci USA 93:13429–13434. https://doi.org/10.1073/pnas.93.23.13429

Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22:240–249

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376. https://doi.org/10.1007/BF01734359

Felsenstein J (1983) Statistical inference of phylogenies. J R Stat Soc 126:246–272

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution (NY) 39:783–791. https://doi.org/10.2307/2408678

Felsenstein J (1989) PHYLIP–phylogeny inference package (version 3.2). Cladistics 5:164–166

Felsenstein J (2013) PHYLIP-phylogeny inference package (version 3.695). Department of Genome Sciences, University of Washington, Seattle

Fitch WM (1971) Towards defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406–416

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

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42:182–192. https://doi.org/10.1017/CBO9781107415324.004

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001a) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314

Huelsenbeck JP, Ronquist F (2001b) MrBayes: Bayesian inference of phylogeny. Bioinformatics 17:754–755

Huelsenbeck JP, Ronquist F (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Jukes TH, Cantor CR (1969) Evolution of protein molecules. Academic, New York, pp 21–132

Katsonis P, Lichtarge O (2014) A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein coding variations on fitness. Genome Res 24:2050. https://doi.org/10.1101/gr.176214.114

Kumar S, Tamura K, Nei M (1994) MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput Appl Biosci 10:189–191

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

Lewis PO, Holder MT, Swofford DL (2015) Phycas: software for Bayesian phylogenetic analysis. Syst Biol 64:525–523. https://doi.org/10.1093/sysbio/syu132

Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342–358. https://doi.org/10.1006/jmbi.1996.0167

Lua RC, Lichtarge O (2010) PyETV: a PyMOL evolutionary trace viewer to analyze functional site predictions in protein complexes. Bioinformatics 26:2981–2982. https://doi.org/10.1093/bioinformatics/btq566

Lua RC, Wilson SJ, Konecki DM et al (2015) UET: a database of evolutionarily-predicted functional determinants of protein sequences that cluster as functional sites in protein structures. Nucleic Acids Res 44:D308–D312. https://doi.org/10.1093/nar/gkv1279

Madabushi S, Yao H, Marsh M et al (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 316:139–154. https://doi.org/10.1006/jmbi.2001.5327

Maddison WP, Maddison DR (1999) MacClade: analysis of phylogeny and character evolution (version 3.08). Sinauer Associates, Sunderland

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) J Chem Phys 21:1087

Mueller LD, Ayala FJ (1982) Estimation and interpretation of genetic distance in empirical studies. Genet Res 40:127–137

Revell LJ (2013) Rphylip: an R interface for PHYLIP. R package (Version 0-1.09)

Rodriguez GJ, Yao R, Lichtarge O, Wensel TG (2010) Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. Proc Natl Acad Sci USA 107:9476–9476. https://doi.org/10.1073/pnas.1005260107

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Shoji-Kawata S, RJr S, Leveno M, Campbell GR et al (2009) Identification of a candidate therapeutic autophagy–inducing peptide. Nature 33:1223–1229. https://doi.org/10.3892/ijo

Sneath PHA, Sokal RR (1973) Numerical taxonomy. W.H. Freeman, San Francisco

Swofford DL (1991) PAUP: Phylogenetic analysis using parsimony, (version 3.1) computer program distributed by the Illinois. Natural History Survey, Champaign

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 10:512–526

Tamura K, Stecher G, Peterson D et al (2013) MEGA6: molecular evolutionary genetics analysis (version 6.0). Mol Biol Evol 30:2725–2729. https://doi.org/10.1093/molbev/mst197

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on mathematics in the life sciences. Am Math Soc 17:57–86

Ward RM, Venner E, Daines B et al (2009) Evolutionary trace annotation server: automated enzyme function prediction in protein structures using 3D templates. Bioinformatics 25:1426–1427. https://doi.org/10.1093/bioinformatics/btp160

Wilkins AD, Lua R, Erdin S et al (2010) Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation. Protein Sci 19:1296–1311. https://doi.org/10.1002/pro.406

Structural Bioinformatics: Life Through The 3D Glasses

10

Ankita Punetha, Payel Sarkar, Siddharth Nimkar, Himanshu Sharma, Yoganand KNR, and Siranjeevi Nagaraj

10.1 Introduction

Structural bioinformatics can be considered as synergy of computational and structural biology. The premise of informatics approaches is to uncover the complexity underlying structures and propose hypothesis for understanding the cellular processes. Broadly, it encompasses two aspects – the development of methods for studying structures of biomolecules and the application of these methods in solving biological problems and elucidation of new biological knowledge. The latter mainly involves analyzing three-dimensional (3D) structures and establishing their link to function.

It has been almost 64 years from inception of structural biology, which started with X-ray diffraction studies of DNA double helix by Rosalind Franklin and Maurice Wilkins (Watson and Crick 1953; Wilkins et al. 1953) and followed by structural determination of myoglobin by John Kendrew and Max Perutz (Kendrew et al. 1958). Since then growth in structural biology has been phenomenal, and this particular field has been pranced from understanding simple protein structure to underpinning complex molecular machines such as proteasome and ribosomes (Liu et al. 2017; Li et al. 2016; Groll et al. 2000; Amunts et al. 2015; McClary et al. 2017; Desai et al. 2017; Myasnikov et al. 2016). To decipher the modes of interaction and the consequences, it is essential to know the individual structures, because structure defines the function of macromolecule. Insight into the structure deep down to atomic level assists in manipulating the biological system for powerful therapeutic potential, as in drug designing. Thus, structural biology is given paramount importance in recent days, and it is incomplete without bioinformatics/computational biology. For example, algorithms are required to visualize the molecules, modeling

A. Punetha (✉) · P. Sarkar · S. Nimkar · H. Sharma · Y. KNR · S. Nagaraj
Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati,
Guwahati, Assam, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_10

tools for analyzing molecular interactions, knowing energetically favorable conformations that allow molecular stability, and decipher putative interactions with the environment.

The structural information can be obtained either by experimental methods using structure determination techniques like X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) or can be predicted using bioinformatics tools (Venko et al. 2017; Dorn et al. 2014; Floudas 2007). All this requires knowledge about computational geometry, computer graphics, and algorithms to analyze and deconvolute the crystallographic data, to fit the resulting electron densities to more manageable ball and stick models, and to use distance information from NMR data to solve the structure. The priority in usage of these methods over one another depends on the question that needs to be addressed.

Obtaining macromolecular structure at one conformation hinders us to know its versatility. One static structure refers to a conformation, and there can be many conformations in a particular state. Knowledge of all states completes the conformational landscape. For example, protein with 10 amino acids will have 9 peptide bonds and therefore 18 different dihedral angles. Assuming each bond angles is in two stable conformations, then possible different conformations are 2^{18} (Zwanzig et al. 1992). Knowing the dynamic behavior of the macromolecule in all conformations is a must to capture its various interactions. In other words, understanding of conformational state of proteins enables us to study its mechanism.

Robust techniques are still in the state of infancy, to capture dynamic functionality of the macromolecule machineries, although it is possible to seize static details that could provide sufficient information to reconstruct the structure of the whole interacting system to yield detailed dynamics as *in vivo*. Computer simulations such as molecular dynamics (MD) and Monte Carlo simulations can help in understanding the putative mechanism of these molecular machineries (Hospital et al. 2015; Kroese et al. 2014; Paquet and Viktor 2015; Pandey et al. 2017). Thus, bioinformatics is indispensable in structural biology, though the extent to which it is applied may vary from simple to complex computational programs. For instance, determining structure for visualization from electron density maps (X-ray diffraction) or frequency distribution graph (NMR) or bio-imaging (cryo-EM) in experimental methods is quite simple relative to comparative modeling, molecular threading, *ab initio* structure prediction of proteins with unknown structure, and molecular simulation of the processes for investigating the dynamics detail. However, what we infer from simulation process may not be the same as *in vivo*, if we do not provide the conditions that mimic the *in vivo* system. Run-time errors, wrong coordinates as source for processing, and force fields that do not fit in to solve the targeted question are few issues that need to be addressed. Stringent validation involving evaluation of bond length, bond angle, torsion angles, and free energy cutoff is always required to accept the model. Further, assessment of root-mean-square deviation is vital to measure deviation of predicted structure with the known closely related structure obtained through experimental methods. A holistic approach of using bioinformatics with structural biology can unleash unprecedented information, which can be explored for resurrection from menacing diseases.

10.2 Fundamentals of Macromolecular Structure

The realization of the need of understanding the structural principles to know the functioning led to remarkable growth in structure of macromolecules like deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and protein. The macromolecules can attain various shapes responsible for particular function. Therefore, in order to understand the function of these macromolecules, their structure needs to be understood.

10.2.1 DNA

Deoxyribonucleic acid is the genetic material that carries the biological information on how an organism will grow, develop, maintain, and reproduce. It is a biopolymer composed of repeating units of nucleotides and usually comprises of two strands, which coil around each other to form a double helix. Each nucleotide is made up of a pentose sugar called as deoxyribose (lacks hydroxyl group at the 2nd carbon of the pentose sugar), a nitrogenous base – either purine like adenine (A) and guanine (G) or pyrimidine like thymine (T) and cytosine (C) – and a phosphate group. The backbone of this polynucleotide chain has alternating sugar-phosphate molecules in which the sugar of a nucleotide is covalently linked to phosphate of the next. The hydrogen bonding between the nitrogenous bases of the two polynucleotide strands (A with T and G with C) results in the formation of double-stranded DNA.

The first extraction of DNA dates back to 1869 by Friedrich Miescher (Dahm 2008), but its double-helical nature was revealed in 1953 by Watson and Crick from the X-ray diffraction data of Rosalind Franklin (Watson and Crick 1953; Wilkins et al. 1953). The nucleic acid research focus was soon shifted to fiber diffraction (Arnott 1970; Arnott et al. 1974a, 1976; Rodley et al. 1976), which provided insight into a variety of structures adopted by nucleic acid like single-stranded helices (Arnott et al. 1976), parallel helices (Rich et al. 1961), and triple and quadruple helices (Arnott et al. 1974b). DNA being flexible can exist in various forms depending on the environmental conditions. Its conformation is governed by the sequence, extent and direction of supercoiling, base modifications, level of hydration, ionic strength, and the presence and concentration of metal ions or polyamines in the solution (Basu et al. 1988; Cheng and Pettitt 1992; Ghosh and Bansal 2003; Choi and Majima 2011; Zhou et al. 2015; Porrini et al. 2017; Dickerhoff et al. 2017; Sathyamoorthy et al. 2017; Kriegel et al. 2017a). The publication of B-DNA structure in 1980 revealed it to be a right-handed double helix (Wing et al. 1980). Also the unusual Z-DNA, the left-handed form of DNA structure, was elucidated (Wang et al. 1979). Tremendous growth in the fine structures of DNA followed, which increased our understanding manifolds (Kocman and Plavec 2017; Kriegel et al. 2017b; Porrini et al. 2017; Yella and Bansal 2017; Gajarsky et al. 2017; Artusi et al. 2016; Arcella et al. 2012; Adrian et al. 2012; Choi and Majima 2011; Chou et al. 2003; Wahl and Sundaralingam 1997).

10.2.1.1 Primary Structure

The primary structure of DNA is made of linear sequence of nucleotides linked by phosphodiester bonds. Each nucleotide itself consists of three components:

1. A pentose sugar – a five-carbon sugar
2. A nitrogenous base – adenine, guanine, cytosine, and thymine
3. A phosphate group

The nitrogenous bases have planar aromatic heterocyclic structure and can be categorized into purines and pyrimidines. Adenine and guanine are purines in structure (a nitrogen containing double ring having a six- and a five-membered ring) and form a glycosidic bond between their 9 nitrogen and 1′-OH group of the deoxyribose. Cytosine and thymine are pyrimidines (a nitrogen containing single six-membered ring) and form glycosidic bond between their 1 nitrogen and the 1′-OH of the deoxyribose. The phosphate group forms a bond with the deoxyribose sugar through an ester bond between one of the negatively charged oxygen groups and 5′-OH of the sugar. The nucleotides in a polynucleotide chain are linked by phosphodiester bond between 5′and 3′carbon atoms. The oxygen and nitrogen atoms in the backbone make the chain polar. The order of nucleotides within a DNA forms its sequence and represented by the series of letters of its base.

10.2.1.2 Secondary Structure

In DNA double helix, the two strands are held by intermolecular hydrogen bonding between the bases of the two strands. The purines and pyrimidines bonded by specific hydrogen bonds form planar base pairs. The adenine base pairs with thymine using two hydrogen bonds, while guanine pairs with cytosine using three hydrogen bonds (Zamenhof et al. 1952; Watson and Crick 1953). The secondary structure is determined by the set of interactions between the bases of the two strands, which is responsible for the shape of the molecule. In DNA, the asymmetric attachment of sugar moiety to the bases on the same side of the base pairs dictates the mutual positions of the two sugar-phosphate strands. In the helix, the successive base pair stacking on each other results in two indentions with different dimensions called major and minor groove, formed by atoms at the backbone surface.

10.2.1.3 Tertiary Structure

The tertiary structure of DNA represents the location of atoms in the 3D space taking account of geometrical and steric constraints. The linear polymer chain of DNA folds to form a specific 3D shape, which might result in various structural forms based on the folding (left- or right-handed), size difference between major and minor grooves, length of helix turn, and the number of bases per turn. The tertiary organization of DNA double helix results in its three forms – B-DNA, A-DNA, and Z-DNA.

The B-DNA is right-handed helix and the most common form of DNA under physiological conditions, neutral pH, and low salt concentration. It attains a narrow, elongated structure with narrow minor groove and wide major groove with helix axis being perpendicular to base pairs. The deoxyribose sugar ring of B-DNA has C2′endoconformation, i.e., the C2′atom is above the plane of C4′-O-C1′. The base separation is the same as the helical rise 3.4 Å. The right-handed double helix has ten base pairs per complete turn, with the two polynucleotide chains antiparallel to each other and linked by Watson-Crick base pairs (A-T, G-C). The Watson-Crick base pairing results in asymmetry of the two deoxyribose sugars linked to the bases of an individual pair on the same side of it. The helix winds along, parallel to the sugar-phosphodiester chains with base pairs almost centered over the helix axis. The wide major groove has similar depth (distance of base pairs from the helix axis) as much narrower minor groove. The major groove is richer in base substituents – O6, N6 of purines and N4, and O4 of pyrimidines compared to minor one. The major groove width renders it accessible to proteins. B-DNA occurs at high water concentration, as the hydration of the minor groove appears to favor B-DNA form. The X-ray diffraction analysis of oligonucleotides crystals reveals that even the same sequence can adopt distinct structures, which may differ in propeller twist between bases within a pair to optimize the base stacking, or the two successive base pairs can move relative to each other showing twist, roll, or slide.

The right-handed DNA duplex attains A-DNA form in dehydrating environments, which is shorter and wider than B-form. It occurs at low water concentration. The A-DNA has C3′ atom above the C4′-O-C1′ plane, i.e., it has C3′endoconformation in contrast to the C2′endoconformation of B-DNA. The C3′endoconformation brings both the consecutive phosphate groups on the nucleotide chain closer together reducing the distance between the adjacent nucleotides by 1 Å in A-form relative to B-form. In A-DNA, the base pairs are twisted, tilted, and displaced nearly 5 Å from the helix axis, which results in different groove characteristics. The major groove is deep and narrow and not easily accessible to proteins, while the minor groove is wide and shallow which can be accessed by proteins but has lower information content than the major groove. Thus, the A-DNA has a hollow cylindrical core. The helical rise is consequently reduced to 2.56 Å, and the helix is wider with 11 base pair per turn.

The Z-DNA is a relatively rare left-handed double helix with pronounced zigzag pattern in the phosphodiester backbone (Wang and Vasquez 2007). Its helix is more narrow and elongated than A- and B-DNA with convex outer surface of the major groove and a deep central minor groove. Z-DNA formation can occur when the DNA has alternating purine-pyrimidine sequence with purines and pyrimidines in different conformation, leading to the zigzag pattern. Usually, there is alteration of cytosine and guanine with cytosine at the first position. It occurs when there is a high salt concentration (Bae et al. 2011). In a base pair, Z-DNA has one nucleotide with sugar in the C3′endoconformation (like A-DNA and in contrast to B-DNA) and the

Table 10.1 Comparison of B-, A-, and Z-DNA

		Type of DNA		
S. no.	Property	B-DNA	A-DNA	Z-DNA
1.	Helix sense	Right-handed	Right-handed	Left-handed
2.	Helical diameter (Å)	20	23	18
3.	Number of base pair per turn	10	11	12
4.	Vertical rise per base pair (Å)	3.4	2.56	3.7
5.	Sugar pucker conformation	C2′-endo	C3′-endo	Pyrimidines – C2′-endo Purines – C3′-endo
6.	Conformation of glycosyl bond	Anti	Anti	Pyrimidines-anti Purines-syn

base in synconformation which places the base over the sugar ring (in contrast to anticonformation in A- and B-DNA). The advantage of having base in anticonformation is that it places the base in a position where it can readily form hydrogen bonds with the complementary base on the opposite strand. The duplex in Z-DNA has to accommodate the distortion of this nucleotide in the synconformation, while the adjacent nucleotide of Z-DNA is in the normal C2′endo, anticonformation.

The comparison between the three forms of DNA is shown in Table 10.1.

10.2.1.4 Quaternary Structure

The interactions between distinct nucleic acids or between nucleic acid and proteins define the quaternary structure. It is a higher level of organization like the nucleosome formation that involves DNA-histone binding and their further organization into chromatin fibers. The DNA quaternary structure governs the accessibility of DNA sequence to the transcription machinery for gene expression. Since a portion of DNA is condensed or exposed for transcription, its quaternary structure tends to vary over time.

10.2.2 Quadruplex Structures

The guanine base has the ability to utilize both its faces at once to form hydrogen-bonded arrays, resulting in multi-stranded structures in guanine-rich DNA sequences. Guanine (G) quartet is one such arrangement with four guanines. The G-quartet is stabilized by forming stacked sets of four bases, where first the four guanine bases form a flat plate, which then stacks over another flat plate to form a quadruplex structure. Each four base unit is stabilized by hydrogen bonding between the base edges and metal ion chelation in the center (Burge et al. 2006; Parkinson et al. 2002). Numerous conformations can be formed from a set of four bases, either from different parallel strands that contribute a base to the central structure or from a single strand that folds around a base. Diverse quadruplexes can be formed depending on the length and number of strand involved and also in the intervening non-guanine loop sequence. The diverse topologies adopted by G-quadruplexes include interlocked G-quadruplexes, double-chain-reversal and V-shaped loops, triads, mixed tetrads, adenine-mediated pentads, hexads, and snap-back G-tetrad alignments (Dolinnaya et al. 2016; Huppert 2010; Campbell and Parkinson 2007; Perrone et al. 2017; Kocman and Plavec 2017).

The presence of DNA tetrameric structure was first shown in 1947 (Arnott et al. 1974b), but the biological relevance was discovered in 1995 (Rhodes and Giraldo 1995). The tetrameric arrangement of DNA exists in the G-rich eukaryotic telomeres (at the ends of the linear chromosomes) and also in non-telomeric genomic DNA, e.g., nuclease-hypersensitive promoter regions (Burge et al. 2006), and viral genome, e.g., the human herpes simplex-1 (HSV-1) genome (Artusi et al. 2016). It has been reported that the DNA G-quadruplex structures are involved in gene expression and telomere maintenance (Takahama et al. 2013; Murat and Balasubramanian 2014; Rhodes and Lipps 2015; Fukuhara et al. 2017).

Cells have specialized regions called telomeres that permit chromosomal end replication utilizing enzyme telomerase (Greider and Blackburn 1985) and also protect the DNA ends from the DNA repair systems of the cell from treating them as damage to be corrected (Nugent and Lundblad 1998). The telomeres in human cells usually contain single-stranded DNA with several thousand repeats of TTAGGGG, which loop back to form DNA quadruplex having conformation very different from the usual DNA helix (Wright et al. 1997). The large loop structures in telomeres called T-loops are extensive circle of the single-stranded DNA stabilized by telomere-binding proteins. Slight variations of human telomeric sequences can form different types of G-quadruplex structures (Griffith et al. 1999; Li et al. 2014). Toward the T-loop end, the single-stranded telomere DNA strand disrupts the double-stranded DNA to base pair with one of the strand to form a triple-stranded arrangement termed displacement loop or D-loop (Parkinson et al. 2002). The G-quadruplex formation at telomeric ends seems to negatively regulate the activity of the enzyme telomerase, which maintains telomere length (Patel et al. 2007; Kuryavyi et al. 2010).

Another addition to tetrahelical families are AGCGA-quadruplexes, which comprises of four 5′-AGCGA-3′ tracts stabilized by G-A and G-C base pairs, forming GAGA- and GCGC-quartets, respectively. Residues in the core of the structure are connected with edge-type loops. Sequences of alternating 5′-AGCGA-3′ and 5′-GGG-3′ repeats form AGCGA-quadruplexes instead of G-quadruplexes. These structurally unique AGCGA-quadruplexes have lower sensitivity to cation and pH variation. This indicates their biological significance in regulatory regions of genes responsible for basic cellular processes that are related to neurological disorders, cancer, and abnormalities in bone and cartilage development (Kocman and Plavec 2017).

10.2.3 RNA

Ribonucleic acid (RNA) is also a biopolymer made up of repeating unit of nucleotides and is involved in various biological processes including coding, decoding of genetic information, regulating gene expression, sensing and communicating the responses to cellular signals, and catalyzing the biological reactions. According to the central dogma, the genetic information stored in DNA is transcribed into RNA called messenger RNA (mRNA). The genetic information in the mRNA is then decoded, and specific protein is synthesized on ribosomes. This process also uses other forms of RNAs called the transfer RNA (tRNA) molecules which deliver amino acids to ribosomes and the ribosomal RNA (rRNA) molecules which link the amino acids together to form proteins. In many viruses, RNA is the genetic material.

RNA is usually a single-stranded molecule folded onto itself instead of paired double strand as in DNA. Intramolecular hydrogen bonding and complementary base paring stabilize the folded structure. Although RNA is a single-stranded molecule, it can also form double-stranded structures which are important to its function (Rich 1956). In 1960, the first experimental demonstration of how information can be transferred from DNA to RNA was revealed by the RNA/DNA hybrid structure (Rich 1960). In 1965, the structure of tRNA was worked out, a structure that carried amino acid and arranged them in order that corresponded to sequence in DNA (Holley 1965; Holley et al. 1965), followed by the elucidation of phenylalanine tRNA structure from yeast (Kim et al. 1974; Robertus et al. 1974). Soon, the structural studies of RNA gained interest, and many structures were subsequently deposited (Ferre-D’Amare and Doudna 1999; Doherty and Doudna 2000; Piccirilli and Koldobskaya 2011; Arieti 2014; Ahmed and Ficner 2014; Patel et al. 2017; Nguyen et al. 2017; Gebetsberger and Micura 2017; Schlick and Pyle 2017; Sun et al. 2017; Zhao and Pyle 2017).

10.2.3.1 Primary Structure

The primary structure of RNA is made of linear sequence of nucleotides linked by phosphodiester bonds. Each nucleotide itself consists of three components:

1. A pentose sugar – a five-carbon sugar
2. A nitrogenous base – adenine, guanine, cytosine, and uracil
3. A phosphate group

RNA is similar to DNA in chemical composition except for a few differences. The sugar composition of RNA is ribose (that has additional hydroxyl group at 2′ position in the pentose ring) as compared to the deoxyribose sugar present in DNA (having no hydroxyl group at 2′ position in the pentose ring). The presence of the hydroxyl groups makes RNA more susceptible to hydrolysis. RNA also differs from DNA in having uracil base instead of thymine, which base pairs with adenine. Uracil is an unmethylated form of thymine and lacks methyl group at the 5 position. Other than these differences, RNA and DNA are the same, having the same bonding pattern of sugars, bases, and phosphates to form nucleotide which then binds to form nucleic acid in similar fashion.

As in DNA, RNA nitrogen bases are divided into types – purines and pyrimidines. Adenine and guanine are purines in structure (a nitrogen containing double ring having a six- and a five-membered ring) which form a glycosidic bond between their 9 nitrogen and 1′-OH group of the ribose. Cytosine and uracil are pyrimidines (a nitrogen containing single six-membered ring) and form glycosidic bond between their 1 nitrogen and the 1′-OH of the ribose. The phosphate group forms an ester bond between one of the negatively charged oxygen groups and 5′-OH of the ribose sugar. The nucleotides are linked by phosphodiester bond between 5′ and 3′carbon atoms in a polynucleotide chain. The oxygen and nitrogen atoms in the backbone make the chain polar. The RNA sequence is the order of nucleotides in the polynucleotide chain and represented by the series of letters of its nitrogenous base A, U, G, and C, denoting adenine, uracil, guanine, and cytosine, respectively. Unlike DNA, RNA has much shorter nucleotide chain.

10.2.3.2 Secondary Structure

The secondary structures in RNA result due to two-dimensional (2D) base pair folding in which local sequences have regions of self-complementarity, giving rise to base pairs and turns. The pairing between the complementary bases within single-stranded polynucleotide chain of RNA results in the existence of both single- and double-stranded areas in the same RNA molecule. The secondary structure elements of RNA can be categorized into four basic types – helices, loops, bulges, and junctions (Tinoco and Bustamante 1999).

Double Helix

The antiparallel strands form the helical shape. RNA double helices have structures similar to the A-form of DNA.

Stem-Loop Structures

Stem-loop or hairpin loop is the most common RNA secondary structure, which is formed when the nucleotide chain folds back onto itself to form double-helical portion called stem. Loop is the single-stranded region formed by the unpaired nucleotides. It serves as the building block for larger structural motifs like cloverleaf structures, which are four-helix junctions like in tRNA.

Bulges and Loops

The unpaired nucleotide region in between the long double-helical region resulting from the parting of the double helices on any one side of the strand forms the bulge and on both the strands forms the internal loops. The four-base hairpin arrangement is called tetraloop. Three common families of tertraloops are present in ribosomal RNA – CUUG, UNCG, and GNRA (where N is a nucleotide and R is a purine). Among tetraloops UNCG is the most stable (Hollyfield et al. 1976).

Pseudoknots

Another form of RNA secondary structure is pseudoknot, which is a helical segment resulting from the pairing of nucleotides from the hairpin loop with a single-stranded region outside of the hairpin. Pseudoknots fold into knot-shaped 3D conformations but are not true topological knots. The base pairing occurs that overlaps one another in sequence position. Pseudoknots are found in most classes of RNA and have diverse functions. It was first identified in turnip yellow mosaic virus (Rietveld et al. 1982). Among the pseudoknots H-type fold pseudoknots are best characterized. It has two stems and two loops. The second stem loop is formed as a result of pairing of nucleotides in hairpin loop with bases outside the hairpin stem (Staple and Butcher 2005). Pseudoknots are involved in several important biological processes like the pseudoknot of RNA component of human telomerase that is critical for activity (Chen and Greider 2005).

10.2.3.3 Tertiary Structure

The three-dimensional structure of single-stranded RNA is formed by base pairing in all the self-complementary regions and can be very complex. It consists of the conformations adopted by the double-helical form, which is stabilized by intramolecular hydrogen bonding. It also forms RNA-DNA duplexes, which are mostly A-form because of the additional 2′ hydroxyl of the ribose sugar that interferes with the arrangement of the sugar in the phosphate backbone. Due to this, it becomes difficult for RNA to adopt the highly ordered B-form, but some RNA-DNA duplexes and localized single-strand dinucleotide of RNA do exist in B-form also (Chen et al. 1995; Sedova and Banavali 2015). The A-RNA helix has 11 base pairs per turn, which are tilted and displaced from the helix axis, having C3′ endoconformation of sugar, a narrow and deep major groove, and a wide, shallow minor groove.

The miscellaneous biological functions of RNA are determined by its complex structure being stabilized by both secondary and tertiary interactions. An important tertiary structure motif is RNA triplex, commonly found in many pseudoknots and other structured RNAs. It usually forms through tertiary interactions in the major or minor groove of a Watson-Crick base-paired stem. In isolation a major-groove RNA triplex structure remains stable by forming consecutive major-groove base triples such as U·A-U and C(+)·G-C. Almost all large structured RNAs possess minor-groove RNA triplexes. Since double-stranded RNA stem regions are often involved in biologically important triplex structure formation and protein binding, they hold great potential for sequence-specific targeting of any desired RNA duplexes by triplex formation (Devi et al. 2015).

10.2.3.4 Quaternary Structure

The quaternary structures represent the interactions between separate RNA units or between RNA and proteins like in ribosome or spliceosome.

10.2.3.5 Quadruplex Structures

In G-rich RNA sequences, noncanonical secondary structures held together by Hoogsteen-bonded planar guanine quartets form G-quadruplexes. They occur in transcripts associated with telomeres, in noncoding sequences of primary transcripts, and within mature transcripts. At these specific locations, they play important roles in key cellular functions, including telomere homeostasis, regulation of pre-mRNA processing (splicing and polyadenylation), RNA turnover, and mRNA targeting and translation (Fay et al. 2017). RNA G-quadruplexes govern regulatory mechanisms like the binding of protein factors that modulate G-quadruplex conformation and/or serve as a bridge to recruit additional protein regulators (Dolinnaya et al. 2016; Agarwal et al. 2012; Millevoi et al. 2012). Current methods for identifying RNA G-quadruplex involve the use of short, purified RNA sequences *in vitro* in the absence of competition with secondary structures or protein binding. In case of long functional RNAs and in cellular context, a comparison of RNA and 7-deaza-RNA is used (Weldon et al. 2016; Weldon et al. 2017).

10.2.3.6 Transfer RNA

The yeast phenylalanine tRNA exists in L-shape, with two arms at right angle to each other as revealed by its crystal structure (Kim et al. 1974; Robertus et al. 1974). The arms consisting of short A-helices are held by extensive base-base interactions. The helix-helix stacking is observed. The D stem's short helix is stacked onto the longer double-helical anticodon arm, while the other arm has acceptor stem helix stacked with four base pair helix of T arm. Overall, it displays a cloverleaf structure with interactions between distant parts of structure. It shows nine additional non-Watson-Crick base-base interactions and several triplet interactions at the two-arm junction, which helps to maintain the structural fold.

10.2.4 Protein

Proteins perform innumerable functions that mediate structural and mechanistic basis of various life processes, for which they interact with various other biomolecules and tolerate different physical factors like pH, temperature, and ionic strengths. Functional versatility of proteins can be attributed to variability in their structure that is optimized by the evolution process. The protein is a biopolymer made up of amino acids. Protein structural folding brings particular amino acid residues to vicinity that further helps in enzyme catalysis, transport, metabolic regulation, and structural functions. Thus, various functions of proteins are driven by variability in structure, which in turn is the function of its amino acid sequence.

Diversity and complexity of protein structures possess a great challenge for the researchers in the area of structural biology. Initially, proteins were considered to lack structural regularity as in DNA double helix but later found to contain various types of regular subunits (Pauling and Corey 1951; Ramachandran 1963; Ramachandran et al. 1963; Eisenberg 2003; Amzel and Poljak 1979; Tilton et al. 1992; Mixon et al. 1995; Sammito et al. 2013; Weisser et al. 2017). Proteins are primarily a linear chain of amino acids (in various combinations); these chains fold to form regular structures termed as secondary structure. In protein, secondary structural elements group together to satisfy various intramolecular interactions to form a tertiary structure. Not all but in few cases, tertiary structures associate with each other (intermolecular) to form quaternary structure.

10.2.4.1 Primary Structure

Protein's primary structure is composed of covalently linked amino acids forming a linear polymer chain. Each protein can be identified by unique composition of its amino acids. Amino acids are small organic molecules comprising of a central carbon atom (α-carbon) attached to carboxyl group (–COOH), amino group (–NH_2), a hydrogen atom, and a side-chain group (–R). The proteome comprises of 20 amino acids. The basic structure of amino acids remains the same except the side-chain group. Based on the properties of side-chain group, amino acids are categorized into polar, nonpolar, and charged. Generally amino acids show chirality and hence exhibit two forms (i.e., D and L forms) which are mirror images of each other. Exception to this is glycine, which is an achiral molecule due to the presence of single hydrogen atom as side chain. Cellular machinery prefers and incorporates only L form amino acids.

The protein is formed by linking two amino acids by a covalent bond called peptide bond, which is resultant of condensation reaction between the carboxyl group of the first amino acid and the amino group of the next. Two or more amino acids linked in this way are called peptides. Thus, a protein can be termed as polypeptide.

The peptide bond characteristics have important implications on the polypeptide 3D structure. The peptide bond being planar and rigid imparts rotational freedom to the polypeptide chain only about the bonds formed by the α-carbons (i.e., Cα-N and Cα-C′). These are termed as Phi (ϕ) and Psi (ψ) angles, respectively. Steric hindrance between the residues side chain and the peptide backbone further limits the rotational freedom about the ϕ (Cα-N) and ψ (Cα-C′) angles. Due to this constraint, only few conformations are possible. Based on sterically allowed ϕ and ψ angles of a polypeptide chain, the entire conformational space can be plotted (ϕ vs ψ angles) into allowed and disallowed conformations (Ramachandran et al. 1963). It is called the Ramachandran plot, with exceptions of glycine and proline amino acids. The side chain of glycine has a simple hydrogen molecule, which reduces the steric hindrance to a greater extent, thus increasing its flexibility and expanding the conformational space, whereas proline has markedly reduced conformational flexibility due to the covalent linkage of side chain to the main chain carbon (Cα), which reduces the conformational space.

10.2.4.2 Secondary Structure

The local conformation of the backbone of the polypeptide chain can be termed as the protein secondary structure. Based on the known physical limitations of polypeptide chains, Linus Pauling, Robert Corey, and H. R. Branson (Pauling et al. 1951) predicted protein to possess alpha (α) helix and beta (β) sheets, which were experimentally proven in the course of protein research. Ramachandran plot also maps two major areas of allowed conformation denoting α-helices and β-sheets. A high degree of regularity is displayed by these structures. In the polypeptide chain, a particular ϕ and ψ angle combination is approximately repeated in its secondary structure. Helices and sheets satisfy the peptide bond constraints, but this is not the only factor that explicates their ubiquity. Hydrogen bond formation between the backbone atoms of the partaking residues makes them a highly favorable conformation for the polypeptide chain. In proteins, apart from the regular secondary structural elements like helices and sheets, irregular secondary structural elements are also present that are vital to both structure and function.

Alpha (α) Helix

Helix is a regular coiled structure produced as a resultant of polypeptide backbone curving. These coils are mostly right-handed in proteins. Steric clashes restrict the left-handed coiling of polypeptide backbone. Among the right-handed helices, α-helix is the most predominant form. The amino acid side chains point away from the helical axis, which form the surface of the helix. An α-helix consists of 3.6 amino acids per turn. The helix structure is stabilized by hydrogen bond formation between the oxygen atom of carboxyl group of each residue and the hydrogen atom of the amide group belonging to 4th residue ahead in the helix. Except at the ends, all backbone hydrogen bonds are satisfied within the α-helix. In this arrangement, the carbonyl groups of all amino acids are arranged in the same directions, whereas the amide groups are oriented in opposite way. Here each amino acid of the α-helix acts as a small dipole. Thus, alignment of all the amino acids in the same orientation gives a directionality to α-helix, i.e., negative to positive in C-terminus to N-terminus direction.

Based on the complexity of the side chain, different amino acids have different tendencies to form α-helix. Residues with a higher frequency of occurrence in α-helices are alanine, glutamate, and leucine. Alanine is the most prevalent amino acid in helix, as it has a small side chain that fits well into α-helix, whereas bulky side chain containing amino acids like tryptophan occurs less often. The presence of hydrogen bond donors and acceptors in the side chains of aspartate, asparagine, and serine makes the least preferable amino acids by α-helices as they can form hydrogen bonds with the main chain when in close proximity, thereby disrupting the core helical structure. Glycine and proline are also less in helix as they act as helix breaker. Glycine with its single hydrogen as a side chain has a flexible movement around alpha-carbon (Cα), whereas proline has reduced flexibility due to its ring structure, and absence of NH group introduces kinks in the main chain.

3_{10} Helix and pi (π) Helix

In addition to α-helix, proteins may rarely contain tightly packed 3_{10} helix and loosely packed pi (π) helix. The 3_{10} helix has three residues per turn, with hydrogen bonding occurring between each residue and the residue 3 positions ahead. The seldom-occurring π-helix has 4.4 residues per turn and exhibits hydrogen bonding between each residue and the residue 5 positions ahead. Both 3_{10} helix and π-helix are seen only at the ends of α-helix.

Beta (β) Sheets

In β-sheets, the hydrogen bonding between the main chain C=O and NH groups does not form between the residues of the same strand but with other parts of the polypeptide, which means a single β-strand does not exist in isolation but spatially adjacent to other strands. This results in the formation of twisted, pleated structure called β-pleated sheet, formed by consecutive, spatially adjacent hydrogen-bonded strands. The individual polypeptide chains participating in the sheet formation are termed as β-strands. In these types of structures, dihedral angles (ϕ and ψ angles) are nearly 180° with respect to each other, producing pleated sheet with the residue side chains approximately perpendicular to the pleated plane. These side-chain groups are further oriented in altering positions on opposite sides of the sheet. The β-sheets are of two types – parallel and antiparallel. In both the types, Cα atoms of adjacent strands are aligned closely, and their side-chain groups face in the same direction. In parallel arrangement, adjacent strands orient in the same direction such that their amino terminus (N-terminus) or carboxy terminus (C-terminus) lie adjacent to each other. These are less frequent and can be only formed by β-strands that are very distant in sequence (as in β-α-β motifs). This type of sheet results in less stable nonparallel inter-strand hydrogen bonding. Antiparallel arrangement orients strands in reverse direction, thus bringing the N-terminus of the first strand besides the C-terminus of the adjacent strand. In this configuration, more stable parallel inter-strand hydrogen bonds are formed. This is the more prevalent form of β-sheet configuration.

In addition to abovementioned configurations, β-sheets can also seldom form mixed configuration, containing a mixture of both parallel and antiparallel aligned β-strands. All the β-sheets exhibit some degree of right-handed twist. In topology diagrams, flat arrows pointing in N-terminus to C-terminus direction represent β-strands.

Valine and isoleucine are most commonly found amino acids in β-strands. The reason for not contributing to α-helices can be drawn to the bifurcation at their β-carbon atom that results in steric clashes, thereby destabilizing the secondary structure, while β-strands can readily harbor these amino acids since their side chains are directed outward to the plane that contains the main chain.

Loops and Turns

In a protein, apart from stable α-helix and β-sheets, unordered structures like loops (coil) and turns also exist. These structures often interconnect ordered secondary structural elements. These structures majorly occur on the surface of the protein and generally contain hydrophilic amino acids. Glycine, asparagine, and proline are commonly found in turns. In many proteins, loop regions bear the active site for enzymatic function. Hairpin loops or reverse turns are the most common in the proteins. These are usually made up of 4–5 amino acids. Reverse turns usually increase the compactness of protein structure by reversing the polypeptide chain direction, by folding it to 180°. These structures are usually connected by internal hydrogen bonds and generally contain proline and/or glycine. Proteins can also contain longer (5–15 residues) loops called omega (Ω) loops, which in addition to polypeptide backbone are also networked by interaction of side-chain groups. Other than these, proteins may also contain highly flexible irregular regions termed as random coils.

10.2.4.3 Tertiary Structure

The tertiary structure is the actual form of the protein structure that is responsible for biological function. The various ordered secondary structural elements interact with their side chains of amino acids and fold in three dimensions to form tertiary structures. The folding is driven by the hydrophobic effect, i.e., the hydrophobic side chains fold to the core regions away from the hydrophilic surroundings. In addition to this, other interactions like hydrogen bonding, salt bridges, covalent disulfide bridges, and weak van der Waal forces contribute significantly in tertiary structure building. This three-dimensional folding allows elsewhere located active site residues of peptide chain to associate closely, thus allowing the substrate binding and catalysis process. Though tertiary structures in total appear to be irregular and lack symmetry, they are comprised of smaller conserved super-secondary structures termed as motifs.

Motifs

Motifs act as structural subunits of the protein and comprise of various secondary structural elements, which are arranged in regular patterns. Based on these arrangements, these super-secondary structures are classified into various types enumerated below.

1. *Helix-loop-helix*

Helix-loop-helix (HLH) is a simple motif comprising of two helices interconnected by a shorter loop. These motifs are generally located in the DNA-binding regions of transcription factors. Generally, this motif comprises of longer basic amino acid-containing helix (DNA interacting) connected to a smaller helix.

2. *Helix-turn-helix*

In this motif, two helices are joined by a loop that makes a turn, thus folding back the polypeptide chain. These are commonly found in DNA-binding domains of the proteins.

3. *Beta (β) hairpin*

This motif comprises of two β-strands connected by a loop forming a hairpin bend. The bending causes reversal of strand direction in the peptide chain. β-Hairpin structures either exist as individual motifs or form a continuous antiparallel β-sheet.

4. *Beta-alpha-beta (β-α-β) motif*

The β-α-β motif commonly exists in proteins with parallel β-sheets. The C-terminus of first β-strand is connected to the N-terminus of a second by a loop-α helix-loop. Generally, in 3D structure parallel β-sheet exists in a plane, where intermittent helices are placed above the sheet plane. Varying lengths of loops are observed in different motifs. In some proteins, catalytic sites are found in the loop regions of β-α-β motif.

5. *Greek key motif*

The Greek key motif consists of four adjacent antiparallel strands arranged in the form of an ornamental Greek key. Three β-antiparallel strands of this motif are connected by two hairpin loops, while the fourth is placed adjacent to the first and linked to the third by a longer loop.

Protein Folds and Domains

Protein fold is a large and complex structure formed by combination of simple motifs. The types of folds that a protein can attain are limited and are commonly related to the type of function. An independently folding large subunit of protein with conserved protein fold and/or specific function is called domain. It is quasi-independent modular units with simpler functions. Based on the structural features, domains are classified into following types.

1. *Alpha (α) domains*

The α-domains comprise of only parallel or antiparallel α-helices. Examples include helix bundle and globin fold.

(a) *Four-helix bundle*

This is one of the common folds present in various proteins. In most cases, four antiparallel helices are bundled to pack hydrophobic core at the helix interface and expose the hydrophilic residues to the aqueous solvent.

(b) *Globin fold*

This fold is present in globin family of proteins (e.g., hemoglobin and myoglobin). This fold contains eight helices forming an active site pocket for binding of heme group. In this domain, two helices at the C-terminus form a helix-turn-helix motif, thus arranging themselves antiparallel. The other helices in the remaining domain pack against each other with angle around 50°.

2. *Beta (β) domains*

The β-domains are made up of β-sheets alone. Examples include up and down β-barrels, jelly roll barrel, and β-sandwich.

(a) *Up and down β-barrel*

The large antiparallel β-sheet wraps around in circular fashion so that the strands that would be on the edges of the sheet are spatially adjacent and hydrogen bonded forming a barrel structure with large void in the center. The amino acid side chains alternatively point above and below the sheet. This space in the center acts as a transporting channel in various membrane proteins.

(b) *Jelly roll barrel*

Jelly roll barrel also consists of single sheet wrapped around itself, but here longer loops transverse the channel core, thus leaving no void. The core region consists of hydrophobic residues. It usually consists of eight beta strands arranged in two four-stranded antiparallel beta sheets.

(c) *β-sandwich*

In this fold, two antiparallel β-sheets are arranged in parallel planes stacking each other like a bread sandwich. In contrast to β-barrel, they conceive a hydrophobic core with no void spaces. The number of strands found in such domains may differ from one protein to another. This type of fold is found in immunoglobulins.

3. *α+β-domains*

The secondary structure of α+β-domains is composed of α-helices and β-strands that occur separately along the backbone. The β-strands are therefore mostly anti-parallel. Examples include the ferredoxin fold, the DNA clamp fold, and the SH2 domain.

(a) *Ferredoxin fold*

A ferredoxin fold is a common α+β-protein fold with a signature βαββαβ secondary structure along its backbone. The ferredoxin fold has as a long, symmetric hairpin that is wrapped around once, so that its two terminal β-strands hydrogen bond to the two central β-strands, forming a four-stranded, antiparallel β-sheet covered on one side by two α-helices.

(b) *DNA clamp fold*

A DNA clamp or a sliding clamp is a protein fold that serves as a processivity-promoting factor in DNA replication. It is an α+β-protein that assembles into a multimeric structure that completely encircles the DNA double helix as the polymerase adds nucleotides to the growing strand. The clamp because of its toroidal shape of the assembled multimer cannot dissociate from the template strand, thereby preventing the enzyme from dissociating. Thus, it acts as a critical component of the DNA polymerase III holoenzyme.

(c) *(Src homology 2) SH2 domain*

The SH2 domain is a structurally conserved protein domain contained within the Src oncoprotein and in many other intracellular signal-transducing proteins. It contains two α-helices and seven β-strands and is approximately 100 amino acids in length. It shows high affinity to phosphorylated tyrosine residues and is known to identify a three to six amino acid sequence within a peptide motif.

4. *α-/β-domains*

In these domains, the secondary structure is composed of alternating α-helices and β-strands along the backbone. The β-strands are therefore mostly parallel. They contain either spread or curved β-α-β motifs. Examples include TIM barrel, flavodoxin fold, Rossmann fold, and leucine-rich repeat (LRR).

(a) *α-/β-barrel*

This type of structure is found in triosephosphate isomerase (also termed as TIM barrel). A sheet of four β-α-β-α units is wrapped into a circle, forming internal core made up of parallel β-sheet covered by α-helices as outer layer. Core region is not entirely hollow. Substrates interact with the loops above the barrel.

(b) *Flavodoxin fold*

It is a common three-layered α-/β-protein fold that has a five-stranded parallel β-sheet sandwiched between two α-helical layers.

(c) *Rossmann fold*

This type of fold is routinely found in nucleotide-binding proteins. It contains open twisted parallel β-sheet with α-helices on both sides. A specific spot in a cleft between two parallel sheets connected by a helix acts as nucleotide binding motif.

(d) *α-/β-horseshoe fold*

A leucine-rich repeat (LRR) is a structural motif that forms an α-/β-horseshoe fold. It is composed of repeating 20–30 amino acid stretches that are unusually rich in hydrophobic amino acid leucine. Many such repeat units consisting of β-strand-turn-α helix fold together to form a leucine-rich repeat domain that takes up a horseshoe shape. Seventeen stranded parallel β-sheets form the interior, whereas interconnecting 16 α-helices form the outer covering of the horseshoe. The hydrophobic core formed between the helices and sheets has tightly packed leucine residues. This type of fold is found in placental ribonuclease inhibitor.

10.2.4.4 Quaternary Structure

Many proteins do not function as single folded polypeptide chains; instead they form a non-covalent association with two or more folded polypeptide chains. Each subunit of this multimeric protein is termed as protomer. Proteins with identical protomers are termed as homomeric, whereas proteins with different protomers are known as heteromeric. In some cases, these proteins possess active site in the interface of protomers, whereas in others each protomer carries a separate active site. Similar interactions stabilize both tertiary and quaternary structures. Formation of quaternary structures grants certain advantages like cooperativity in function (e.g., one protomer of hemoglobin bound to oxygen promotes other three peptide chains to bind to substrate), structural assembly (e.g., multiple heterodimers of tubulins associated with each other to form microtubules), and co-localization of different functions which results in various protein multimers, i.e., multifunctional complexes.

10.3 Structure-Function Relationship

The protein is able to perform its biological function by forming stable 3D structure in normal environment. For example, enzymes use a cavity in the surface of their 3D structure called active site, which is accessible to reactants to catalyze the reactions. The multifunctional active sites contain key catalytic machinery of the protein, consisting of one or more residues that are actively involved in catalyzing the reaction and transition-state stabilization. Based on the active site shape and physiochemical properties, only a particular class of molecules can bind and catalyzed. All this depends on the active site attaining proper 3D conformations, which in turn depends on the folding of the polypeptide chain. In general, all proteins rely on specific 3D structure to perform their biological function. All proteins are not enzymes and may have other functions such as molecular recognition like transport proteins need to recognize and carry specific molecules, or the antibodies which need to recognize the foreign proteins, or the interaction of components in signaling pathway or the complex formation. The recognition of other macromolecules is very important in gene expression regulation by DNA-binding protein and formation of nucleoprotein complex like ribosome. The recognition of molecular signal by receptor proteins is important in sensing (e.g., the receptors present in cell nucleus sense steroids).

The basic requirement for molecular recognition requires binding of the molecules in energetically favorable conformation, which depends on complementarity of shapes and physiochemical properties, i.e., they must fit snuggly together and their surface atoms in contact must have complementary properties. Thus, the hydrophobic area of one interacting partner must be in contact with hydrophobic area of the other, and the negatively charged area of one must contact the positively charged area of the other. All this is dependent on the formation of specific 3D structure of proteins. Therefore, the protein function is dependent on its attaining a stable specific 3D structure.

Various approaches have been developed to predict function from the structural information. The basic approach that uses structural data for predicting function of a protein relies on finding globally similar structural features (Sleator and Walsh 2010). However, if the match is not significant, similarities between the functional sites are assessed. Typically, it involves either protein fold comparison, use of local 3D templates, or the local structural feature comparison. Proteins having similar structural features along their entire sequence are more likely to have similar functions (Whisstock and Lesk 2003; Tosatto and Toppo 2006). Some of the popular web services available for quantifying this relationship are DALI (Holm and Laakso 2016), CATHEDRAL (Redfern et al. 2007), SALSAs (Wang et al. 2013), and FLORA (Redfern et al. 2009). The significance of the similarity is assessed based on the number of amino acid residues considered in the alignment and the quality of superposition. Detecting the presence of common motifs distributed over the range of diverse folds within the structure hints the key functional similarity. The analysis of CATH database (Dawson et al. 2017) reveals that the protein domains having the same folds tend to have a specific function, but a few number of additional superfolds can completely change the key function. Recent advancement in the similarity-based scoring methods involves the comparison of protein's internal residue contact that identifies the residues co-located in the range of 8–10 Å in the structure and finally detects additional similarities using conventional global alignment methods.

Though whole fold comparison is the most common method used to assign protein functions, it has some limitations. It does not consider the conservation of the local environment distinctly, which is very important as small changes in the active site residues can cause a complete alteration in protein functions. For example, the function of enzymes and DNA-binding proteins is solely dependent on the conservation of their active site residues. Thus, methods have been developed that compare smaller structural motifs to assign specific functions to proteins. The Catalytic Site Atlas (Furnham et al. 2014) is a protein structure database that stores all manually annotated catalytic site residues of different proteins. It helps to provide a structural template that can be compared to the protein structures of unknown function using a fast search algorithm to transfer and assign the closest Enzyme Commission (EC) numbers. Hydrophobic residues are often eliminated while constructing a structural template because they tend to be buried in the core of the protein. The EzCatDB database houses manually classified enzymatic reactions based on enzyme active site structures, their catalytic mechanisms based on literature, amino acid sequences of enzymes (UniProtKB), the corresponding tertiary structures from the Protein Data Bank (PDB), and ligand information classified in terms of cofactors, substrates, products, and intermediates. It provides various sequence search methods, including the detection of remote homology (Nagano et al. 2015). The structure-function linkage database (SFLD) is a manually curated classification resource describing structure-function relationships for functionally diverse enzyme superfamilies. SFLD enables rational transfer of functional features to unknowns in cases where the members of superfamilies have diverse functions but share an ancestry and some conserved active site features associated with conserved functional attributes and therefore tend to misannotate (Holliday et al. 2017).

Protein surface analysis as well as analysis of the conformation of the active site cleft also provides information on protein function. It can fetch information on small molecule binding and potential protein-protein interaction. The ability of a protein to maintain a unique chemical environment and specific binding pocket conformation aids them to distinguish between their substrates and catalyze reactions effectively. Based on the local structural features, the binding sites in unannotated proteins can be compared against a database of known sites. For example, the web server pocket and void surfaces of amino acid residues (pvSOAR; Binkowski et al. 2004) performs such comparisons.

Recent approaches of protein function assignment include comparison of the physiochemical properties of the active site residues, the charge conservation, hydrophilicity, and information about the electrostatic potential surfaces that helps to identify similarities in the charge distribution pattern in the interaction sites (Dudek et al. 2017; Quester and Schomburg 2011; Ruiz-Blanco and Aguero-Chapin 2017; Wang et al. 2017; Stahl et al. 2017).

10.4 Macromolecular Structure Determination

There are several methods to determine the protein structure like X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy (EM). The priority in usage of these methods over one another depends on the biological question that needs to be addressed. If one has to study small protein with <50 amino acids, NMR is the obvious choice. Not all proteins behave the same, some are easy to crystalize, and some are not amenable for crystallization. For easily crystalizing proteins, X-ray crystallography is preferred. Cryo-EM helps to unravel overall topology of the protein interactions by direct imaging the macromolecular interactions. Other techniques such as electrospray ionization mass spectrometry (ESI-MS) have also been developed to study macromolecular structures that are not accessible by either NMR or X-ray crystallography. All these techniques have some pitfalls. The major issue with X-ray crystallography is the complexity of amino acids, which decides the protein's fate to get into crystal or not. Additionally, the expression of protein in larger quantity for structural studies is often difficult.

Uncertainties are associated in predicting crystallization feasibility of a protein. Solving phase problem in attaining the information from both phase and amplitude is crucial, which will account for complete information about the electron density maps. Isomorphous replacement and anomalous dispersion with heavy atom cluster improve this phase problem. Radiation damage is the common problem with X-ray crystallography and cryo-EM technique. Theoretically, higher electron dose gives higher resolution. However, in practice, exposing higher electron dose damages the sample, and thus only low-resolved structures are attained. This problem can be rectified by exposing the sample with moderate level of electron dose and capturing the signal from different orientations followed by reconstruction to attain highly resolved structure. The major issues with solution techniques such as NMR are that it allows increased conformational flexibility of proteins, resolution is limited, and complex data analysis makes it cumbersome at times.

10.4.1 NMR Spectroscopy

Nuclear magnetic resonance is used to determine the structure of macromolecule in solution at the atomic-level resolution (Sugiki et al. 2017). The purified protein is placed in strong magnetic field and probed with radio waves. The resonance pattern is analyzed to get information about atomic nuclei close to one another and local conformation of bonded atoms. The model of protein is then built using the list of restraints. NMR can also be used to study the various other properties of protein such as changes in protein domain on ligand or substrate binding, enzyme kinetic, etc. at the atomic level. However, it is limited to small or medium proteins because the large protein results in overlapping peaks in NMR spectra. For larger proteins (>30 kDa), more powerful NMR spectrometers are required which are currently unavailable. Commercially up to 900 MHz NMR are available with the latest being 1,020 MHz NMR (Hashi et al. 2015).

10.4.1.1 The Principle of NMR Spectroscopy

All the atoms contain nuclei, which have certain angular velocity and harbor neutrons and proton. Protons are charged particles and when rotated create a spin angular momentum. The spin angular momentum vector characterizes the spin. The rotating nucleus creates a magnetic field in the direction perpendicular to the rotation as described by the right-hand rule (RHR). RHR states that when the fingers of the right hand are curled in the direction of circular motion, the thumb points in the direction of the angular momentum vector. This is called as magnetic moment vector, μ. This magnetic moment and spin angular momentum are directly proportional. When an external magnetic field is applied to a nucleus, the nucleus aligns itself in the external magnetic field. If electromagnetic pulse in the form of radio frequency (RF pulse) is applied to the aligned nucleus, a perturbation in the alignment is created, which is proportional to the external magnetic field and the nuclei under observation. Thus, when an RF pulse is in *on* state, the alignment of the nuclei gets disturbed, and when the RF pulse is in *off* state, the nuclei try to realign itself with the external magnetic field. It is measured as declining amplitude with time and is called as free induction decay. This gives a measure of frequency and decay as a function of time. In order to get the spectrum with a particular peak for particular nuclei, the data is subjected to Fourier transform. Each peak in an NMR spectrum defines certain magnetically different nuclei. The presence of other nuclei in vicinity in the form of atomic bonds, van der Waals interaction, ionic interaction, etc. will have an effect on the position of the peak. This is termed as chemical shift. The chemical shift for an NMR signal is normally measured in Hertz (Hz) shifted relative to a reference signal of tetramethylsilane.

10.4.1.2 Protein NMR

For protein NMR mainly H^1, C^{13}, and N^{15} NMR spectra are measured. Since the abundance of these nuclei except H^1 is very less in nature, the protein is labeled with C^{13} and N^{15} isotopes. Labeling is done while growing bacterial culture in the media containing these nuclei in the form of nutrients. N^{15} and C^{13} labeling are commonly referred to as double labeling. The protein is produced by expression from bacteria, which are grown on minimal medium supplemented with $^{15}NH_4Cl$ and ^{13}C-glucose.

In order to determine protein structure, two-dimensional (2D) NMR is preferred over one-dimensional (1D) NMR, where only a single type of nucleus is taken into consideration. The following methods are used to obtain the 2D data.

10.4.1.3 Correlation Spectroscopy (COSY) and Total Correlation Spectroscopy (TOCSY)

It is the most popular and widely used method for 2D NMR. COSY transfers the magnetization through chemical bonds between the adjacent atoms. A COSY data shows the frequencies of a single atom on both axes. The diagonal peaks have the same frequency coordinates on both the sides and hence appear on the diagonal. The cross peaks are due to the phenomenon called as magnetization transfer, which indicates that two nuclei are coupled.

TOCSY differs with COSY in the fact that in TOCSY magnetization is transferred to all the protons that are connected to the adjacent atom, i.e., the magnetization is transferred from primary to secondary atom and then to tertiary atom.

10.4.1.4 Heteronuclear Correlation Spectroscopy

A heteronuclear correlation spectroscopy gives the data based on the interaction/coupling between two different nuclei types.

10.4.1.5 Heteronuclear Single-Quantum Correlation Spectroscopy (HSQC) and Heteronuclear Multiple-Quantum Correlation Spectroscopy (HMQC)

HSQC is used to detect the correlation between two nuclei of different types, which are separated by a single bond. This method gives one peak per coupled nuclei, and the coordinate for this peak is the chemical shifts in the same coupled nuclei.

HMQC is similar and gives identical spectra to HSQC but uses a different pulse program. However, HSQC is often considered better than HMQC.

10.4.1.6 Nuclear Overhauser Effect Spectroscopy (NOESY)

This method detects the correlation between the two nuclei, which are not bonded but are closely placed in space. The spectrum which is obtained is similar to COSY with both cross and diagonal peaks. Here cross peaks arise due to correlation through space rather than through bond.

10.4.2 X-Ray Crystallography

X-ray crystallography employs X-rays to determine the atomic structure of a molecule. It is by far the best method to solve protein structure (Ilari and Savino 2008, 2017; Yang et al. 2004). X-rays are diffracted from protein crystal, and the diffraction angle and the intensities of diffracted rays are calculated to create a 3D view of electron density. This electron density is then used to solve the molecular structure.

X-ray crystallography can be said to share resemblance with microscopy. In visible microscopy, the shortest wavelength used is around 300 nm, and it is sufficient to visualize cells and subcellular structures. With electron microscopy, the wavelengths used can go as less as 10 nm. In order to understand the protein structure with the distances between the atoms, around ~1 Å X-rays are used. X rays used ranges between 0.5 and 1.5 Å in wavelengths. The structure determination using X-ray crystallography requires protein crystal as the diffraction pattern from a single protein molecule is too weak to be measured. A protein crystal is a solid material in which each protein molecule is arranged in a highly ordered microscopic structure, forming a lattice that extends in all the directions. If the internal structure of the protein crystal is highly ordered, X-rays will be diffracted to high angles and high resolution. On the other hand, poor crystal packing leads to lower resolution, and the data generated is not useful to solve the molecular structure.

10.4.2.1 Protein Crystallization

The process begins with purification of protein to be crystallized. Protein can either be isolated from its source or it can be overexpressed and isolated from an expression platform. The following points should be followed to obtain a good crystal:

1. Protein should be pure and homogenous.
2. Protein should be active and properly folded.
3. Protein should be soluble.
4. Concentration of protein should be as high as possible. Typically, more than 15 mg/ml is used.
5. Protein should be monodisperse and there should be no aggregation.

If any of the above criteria is not met, it becomes very difficult to obtain crystals, and one has to modify expression and purification conditions such as pH, salt concentration, etc. The solubility of protein depends on the interactions with other compounds present in the solution. At physiological conditions, proteins are soluble, but as the concentration rises, the protein tends to precipitate, a process called as

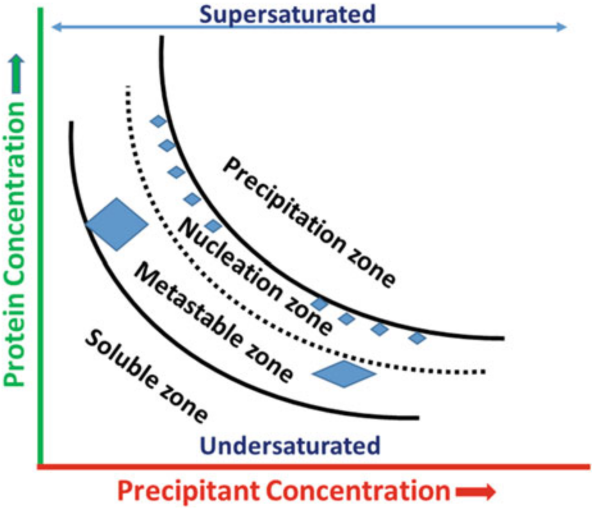


Fig. 10.1 The effect of titration of protein with precipitant

salting out. The basic idea behind crystallization is to slowly salt out protein to form crystals. Precipitant concentration is increased gradually, which allows protein to enter metastable state leading to crystal formation (Fig. 10.1).

Many factors influence the crystal formation, such as the following:

1. Protein purity – If the protein is not pure, the lattice will not be properly formed, which will lead to the disintegration of crystals.
2. pH of the solution – Protein tends to precipitate and form crystal near its pI as the charge on protein becomes null, which leads to easy precipitation.
3. Concentration of protein – If the protein concentration is too low, it tends to remain in soluble form, while the molecular crowding due to high concentration of protein easily forms crystals.
4. Temperature – It affects the rate of precipitation and hence the crystal formation. Typically, 4 and 18 °C are used.
5. Precipitant – Different proteins tend to precipitate with different precipitants. Hence, the choice of precipitant depends on the protein.
6. Additives – The use of additives is to increase intermolecular attraction between the proteins molecules, or it may help in decreasing the interaction between the solvent and protein, thereby increasing the propensity of protein to crystallize.

10.4.2.2 Methods of Crystallization

Vapor Diffusion or Hanging Drop

It is the most common method used for protein crystallization, also called hanging drop method. In this method, a known volume of drop of concentrated protein is mixed with certain volume of crystallization buffer containing precipitant and is allowed to equilibrate with a large reservoir of the same crystallization buffer

Fig. 10.2 Protein drop setting using hanging drop method

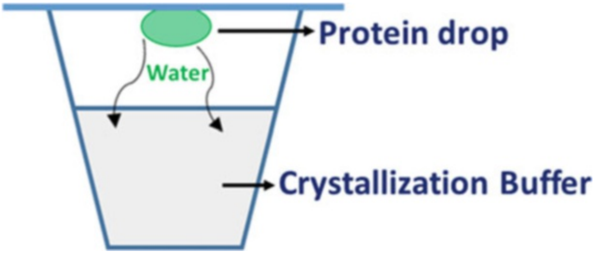
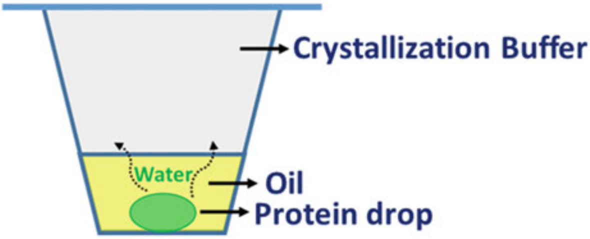


Fig. 10.3 Protein drop setting using sitting drop method



(Fig. 10.2). Initially the precipitant concentration in the drop is less, and the water concentration is more as compared to the reservoir buffer. However, as the system tends to achieve equilibrium, water from droplet evaporates into reservoir, thereby increasing the precipitant concentration in the droplet. This slow increase in precipitant concentration leads to crystal formation.

Microbatch Crystallization

In microbatch method or sitting drop method, the protein drop and reagent are combined and sealed in a plate, tube, and container or sealed under a layer of oil (Fig. 10.3). It can be categorized in two types:

1. *Microbatch under oil*

In microbatch under oil method, the protein drop is placed at the bottom of the tank, and it is then covered with a layer of oil – either paraffin oil or Al's oil (a mixture of 1:1 silicon oil and paraffin oil). Oil acts as a barrier between the reservoir and the protein drop, allowing little to no diffusion of water through the oil. Microbatch under Al's oil permits diffusion of water from the drop through the oil, hence allowing for concentration of the sample and the reagents in the drop.

2. *Microbatch without oil*

Microbatch can be performed without oil. For example, batch crystallization experiments used for small molecules that involve larger volumes in the order of milliliters rather than micro- or nanoliters. Such experiments are performed in a sealed container, with or without the possibility of evaporation, and usually involve temperature control. No oil is used to cover the protein and reagent. Microbatch without oil can also be performed on a micro- or nanoliter scale in a sealed plate, which is termed as drop drop crystallization.

Micro-dialysis

This method employs a semipermeable membrane across which precipitant can pass, whereas larger molecule like protein cannot pass. A salt gradient is established across the membrane, which allows slow diffusion of precipitant into protein drop.

Data Collection

The second step in X-ray crystallography is bombardment of protein crystal with high-intensity X-rays. Four types of X-ray sources are available for protein crystallography:

1. Bombardment of metal (Cr, Cu, or Mo) with high-energy electron beam
2. From a synchrotron radiation source
3. From a radioactive decay that generates the X-rays
4. By exposing substance to primary beam of X-rays to generate secondary X-rays

X-rays once generated are then shot to protein crystal. Most of the X-ray pass through the crystal without any diffraction, but some X-rays are scattered from the electron, and this phenomenon is called as X-ray diffraction. Although these waves cancel one another out in most directions through destructive interference, they add constructively in a few specific directions, determined by Bragg's law:

$$2d \sin \theta = n\lambda$$

Here d is the spacing between diffracting planes, θ is the incident angle, n is any integer, and λ is the wavelength of the beam. These specific directions appear as spots on the diffraction pattern called reflections.

Phase Problem

In order to solve the structure, phase information is required. The destructive interference of waves lead to phase problems as no diffraction pattern is obtained in such case. There are few ways to solve phase problem. If the coordinates are already present from similar protein structure, molecular replacement can be used to solve the phase problem. Molecular replacement takes coordinates from the existing system and tries to fit into the experimental data until a good match is obtained. If it is successful, it can be used to create electron density map. In other methods, heavy atoms are allowed to diffuse into the crystal without affecting the crystal lattice. Since heavy atoms are large in size, it is assumed that one unit cell will have only one heavy atom. Heavy atoms are electron dense and hence give a very clear diffraction pattern. The diffraction data is also collected without the heavy atoms. The difference between the two data can allow easy calculation of phase using vector simulation method. This method is called as isomorphous replacement. Once the phase is known, electron density map can be created using Fourier transform. Sometimes atoms which cause significant scattering are used such as sulfur or metals from metalloproteins. Sulfur can be easily replaced with selenium to solve the phase problem. If multiple wavelengths from sources such as synchrotron are used, then it is termed as multiwavelength anomalous diffraction (MAD), and if single wavelength is used, then it is called as single-wavelength anomalous diffraction (SAD). The next step is the creation of model using the electron density map. Model building starts first with fitting protein backbone to electron density map. The amount of details depends upon the resolution of the data. Once the backbone is fitted, the protein chains are fitted. Building a model is like solving a jigsaw puzzle. The best-fitted model is taken for structure refinement. In refining, the model is further improved, which results in better phase and resolution. The solved structure is then deposited in PDB.

10.4.3 Electron Microscopy

Electron microscopy is used to determine structures of large macromolecular complexes. It uses a beam of electrons to image the molecule directly. The resolution of microscopy depends upon the wavelength of light used, which can be increased by decreasing the wavelength. The wavelength of electron waves is 0.1 million times smaller than visible electromagnetic waves; hence it has a higher resolution. The resolution is given by the Rayleigh formula:

$$r = \frac{1.22\,\lambda}{2n\sin\theta}$$

Here r is the resolving power, λ is the wavelength of the beam, n is the refractive index of the view medium between the objective lens and the object, and θ is the semi-angle of collection of the magnifying lens.

It uses electromagnetic and electrostatic lenses to control the electron beams and direct it toward the specimen. Majorly two types of electron microscopes are used, namely, transmission electron microscope (TEM) in which electrons are transmitted through ultrathin section of specimen and scanning electron microscope (SEM) in which specimen is scanned with beams of electron. TEM has one order higher magnitude of resolution than SEM.

10.4.4 Cryo-electron Microscopy (Cryo-EM)

For determining the structure of protein, TEM at cryogenic temperature is used where the samples are cooled with the help of liquid nitrogen. This technique is called as cryo-electron microscopy. Cryo-EM has gained popularity in structural biology and is being used either singly or with NMR/X-ray crystallography to understand the molecular structure. This technique is specialized in visualizing viruses, organelles, large protein complexes, or nucleic acid molecules (Frank 2017; Bai et al. 2015; Subramaniam et al. 2016; Skiniotis and Southworth 2016; Razi et al. 2017). It requires quick freezing of the biological sample using liquid nitrogen so that the innate structure of the sample remains preserved and the aqueous environment around it is not disturbed. In contrast to X-ray crystallography where protein needs to be crystallized, which can be expensive, cryo-EM can be used to visualize samples without staining and with maintaining the native environment around the protein. However, minimum size of protein or protein complex that can be visualized is around 170 kDa with a resolution of around 2.2 Å. For smaller molecules NMR or X-ray crystallography needs to be used. The development of methods to quickly freeze the samples into thin layer made cryo-EM more popular. The liquid nitrogen reduces the radiation damage caused by high-intensity electron beams. Other than liquid nitrogen, liquid helium has also been used as a cryogen.

Recent advances have revolutionized cryo-EM, which include the use of direct electron detectors that yield images of unprecedented quality, better movie-processing methods that correct the beam-induced sample movements, new classification methods that separate images of different structures, and field emission gun microscope that provides stronger signal with higher resolution. These technological breakthroughs have enabled cryo-EM to achieve near-atomic resolution structural information for a wide variety of biological complexes (Frank 2017; Bai et al. 2015; Orlov et al. 2017).

10.4.5 Cryo-electron Tomography (Cryo-ET)

Cryo-electron tomography is a powerful technique that can image the native cellular environment (Asano et al. 2016; Bharat et al. 2015). It relies on the intrinsic contrast of frozen cellular material for direct identification of macromolecules. In cryo-ET, multiple 2D projections of biological sample are computationally integrated to reconstruct its 3D image. Multiple images are taken with every image tilted at a certain angle as compared to the previous image, and then all images are merged to create a complete 3D image. This allows densities to be resolved in 3D that would otherwise overlap in 2D projection images. To increase signal-to-noise ratio and resolution, the structures present in multiple copies within tomograms are extracted, aligned, and averaged. This reconstruction approach is termed subtomogram averaging and can produce 3D pictures (tomograms) of complex objects such as asymmetric viruses, cellular organelles, or whole cells (Bharat and Scheres 2016; Wan and Briggs 2016). Subtomogram averaging or single particle tomography (SPT) is gaining enormous momentum and becoming a widely used technique, owing to its potential for in situ structural biology at subnanometer resolution. With recent advances in sample preparation, detector technology, and phase plate imaging, it can be applied to unambiguously determine the structures of macromolecular complexes that exhibit compositional and conformational heterogeneity, both in situ and *in vitro* (Galaz-Montoya and Ludtke 2017).

The limitation with cryo-ET is that the samples must be cut into thin sections to allow proper freezing and TEM images to be taken. If the sample is too thick, then it must be sliced into fine sections to obtain better image. Another limitation is that the samples have to be kept at cryo-temperatures to avoid radiation damage, which limits the 3D resolution of the sample.

10.5 Structural Data Representation

With the increase in structural information of macromolecules, there aroused the need to represent the structural data in uniform format, so that it can be easily accessed and compared. For that purpose, different formats were formed in which the structural data can be represented and stored in respective databases like the PDB (Rose et al. 2017) and the nucleic acid database (NDB; Narayanan et al. 2014). The data is usually represented in the PDB format or the dictionary built representations like macromolecular Crystallographic Information File (mmCIF).

10.5.1 The Protein Data Bank Format

PDB at Brookhaven National Laboratory was established in requirement of a common repository for biological macromolecular structural information by Walter Hamilton in 1971 (Bernstein et al. 1977; Berman et al. 2000b; Berman et al. 2000a) with addition of advanced extensions in the format in 1992 and 1996. A detailed description of the format is provided in the PDB Contents Guide, which enumerates the field formats for each PBD record, remarks, and defines the convention for naming atoms, residues, and nucleotides. The PDB format consists of a collection of fixed format records that describe the atomic coordinates, the refinement details, experimental details of structure determination, biochemical features, secondary structural assignments, hydrogen bonding, active site, and biological assemblies. This uniform representation has enabled comparative analysis of the data.

10.5.2 mmCIF: Dictionary-Based Approach

The macromolecular Crystallographic Information File archives the information about crystallographic experiments and results (Hall et al. 1991; Hall 1991), which is also the accepted format of articles in *Acta Crystallographica C*, a scientific journal. The International Union of Crystallography (IUCr) in 1990 formed group to expand the dictionary to satisfactorily describe the macromolecular crystallographic experiment and its results, which included description of all records in a PDB entry. Subsequently, many improvements were incorporated to provide sufficient data names, which would help in writing the experimental section of a structure paper. Tools were also developed so that mmCIF data files could be easily accessed and validated using computer programs. The structure of the dictionary was further improved to deal with complexity of macromolecules data, and the Dictionary Definition Language (DDL; Westbrook and Hall 1995) was used. Soon it was realized that it was not sufficient. Its data typing was not efficient with missing links among data items. This led to the development of enhanced DDL (DDL2). The dictionary was placed on World Wide Web, and mmCIF list server was used to receive comments from the community, which resulted in continuous correction and update of the dictionary. mmCIF dictionary version 1.0 containing 1700 definitions was released in 1997 after the review of the IUCr committee that supervises the dictionary development. The dictionary extensions were managed using a scientific journal as model with proposed extensions being sent to the specialized editors of the mmCIF dictionary for scientific review and then sent to technical editors. New definitions came with succeeding years, which were incorporated in the mmCIF dictionary version 2. To parse and access CIF and mmCIF, software libraries were produced for many languages including C, C++, Java, Fortran, Perl, and Python. The syntax of mmCIF data files and dictionaries is similar to the syntax of core CIF (used for describing small molecule crystallography) and is derived from the Self-defining Text Archive and Retrieval (STAR; Hall 1991) grammar. The mmCIF simplest data file has paired collection of data item names and values.

10.5.3 Dictionaries of Other Data

These dictionaries contain the number of contents that were not covered in mmCIF dictionary but are developed on the same methodology used for mmCIF data and are consistent with its data representation. For example, imgCIF dictionary details the crystallographic data in ASCII and binary formats from image detectors, symmetry extension adds crystallographic symmetry details, cryo-electron microscopy extension adds the structure and volume data for 3D EM experiments, BioSync dictionary describes the features and facilities available at synchrotron beamlines, MDB dictionary provides homology models, and PDB exchange dictionary provides data internally used by PDB and data required to describe high-throughput structure determination. Thus, a single file format cannot be used for all users and application. Application program interfaces (API) are used to access data to avoid file format issues. Data is accessed collection of functions, procedures, and methods depending on the language used which is standardized by Object Management Group (OMG) using Common Object Request Broker Architecture (CORBA). The language- and platform-independent programmable interfaces are defined using interface definition language (IDL), which is supported by CORBA. Thus, CORBA supports the cross-platform access and often called middleware. The mmCIF data representation in CORBA IDL for macromolecular structure provides efficient program access to all the data in PDB entries.

Each of the representation of macromolecular structural data has their own strength and weaknesses. The PDB format is accessible with simple tools, while the mmCIF format based on data dictionary provides comprehensive ontology, precise definitions, and examples with robust metadata model, which can be used to perform thorough checks on individual data and of internal consistency of data items.

10.6 Macromolecular Structure Prediction

Structure of a biomolecule is required to appreciate the functional dynamics of the living system. The protein 3D structure enables us to understand its function and mechanism of action. The information about a protein structure and its interactions with ligands and other proteins, nucleic acids, are also essential for pharmaceutical industry in structure-based drug discovery and drug design. The structural information is less as compared to the sequence information because the experimental structure determination is a slow process and also not possible for many. This result in gap between sequence and structural knowledge is called the sequence-structure gap. The computational methods provide structural information of the proteins whose experimental structure is not available. This unavailability may be due to the difficulties in obtaining the protein (at various steps – cloning, expression, and purification, amount obtained) or failures in experimental determination (may be too large for NMR analysis or cannot be crystallized for X-ray diffraction or other difficulties in using Cryo-EM). In such cases, protein modeling helps to predict the structure of proteins from its sequence.

Structure prediction is the prediction of the relative position of every protein atom in 3D space using the information from the protein sequence. According to the theoretical basis, the prediction methods can be characterized into knowledge-based (like comparative modeling/homology modeling and fold prediction/threading) or *ab initio*. The knowledge-based methods predict structures using information from the databases of known structures. It assumes that a sequence similar to the sequence of known structure will adopt a similar structure. The *ab initio* methods on the other hand predict structure based on fundamental physical principles using quantum mechanics and statistical thermodynamics. This method attempts to calculate and minimize free energy. The difficulties arise due to current computational power that is not sufficient to model proteins with enough solvent molecules, as this forms enormous system with thousand atoms making it difficult to calculate exact free energies. Suitable approximations of free energy are therefore required, which still capture the essentials of protein folding.

The approaches to predict 3D structure are selected based on sequence identity of the target protein with the available homologous sequences with known 3D structure. The accuracy of the structure prediction is measured in terms of root-mean-square deviation (RMSD) between the α-carbon positions in predicted and actual structure of the target and depends on the target-template sequence identity. RMSDs less than 1.0 Å represent good prediction but it is difficult to achieve. If the percentage sequence identity is 70% or more, the model is accurate to an RMSD of less than 2–3 Å. If the sequence identity is above 50%, models tend to be reliable,

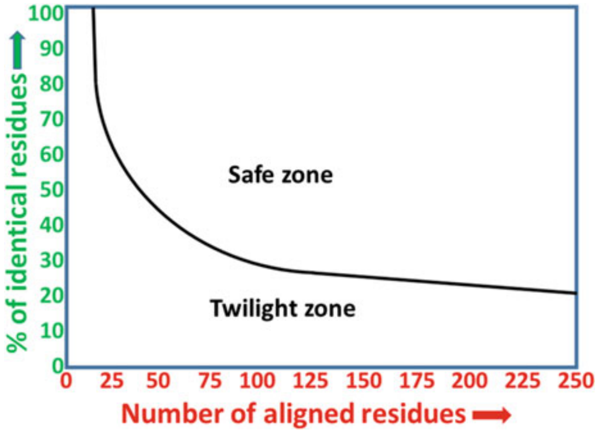


Fig. 10.4 Zones in sequence alignment – safe zone can provide reliable results for homology modeling, while fold prediction or threading is safer in the low-identity twilight zone

with only minor errors in side-chain packing and rotameric state. If the sequence identity is in the range of 30–50%, errors can be more severe and are often located in loops. The regions above 30% sequence identity fall in safe zone (Fig. 10.4) for homology modeling, while the regions below it fall in twilight zone (Rost 1999). In this low-identity region, fold recognition methods are preferred over homology modeling as serious errors can occur like wrong prediction of basic fold (Blake and Cohen 2001; Baker and Sali 2001). The primary source of error at high sequence identities (where homology modeling is done) can arrive due to wrong selection of the template or templates for model building, while at lower identities error can occur in sequence alignment inhibiting high-quality model generation (Venclovas and Margelevicius 2005).

10.6.1 Homology Modeling

Homology modeling, often called as template-based modeling, is a comparative modeling of protein, where 3D structure of the target protein is constructed from its amino acid sequence and an experimentally determined structure of a homolog called as template. It depends on identifying template/templates that might resemble the structure of the target/query and on the production of an alignment, which maps target-template sequence residues. It is advantageous to check alignment of conserved key structural and functional residues. During evolution, the protein structures are more stable and conserved and change much slower than protein sequences among homologous. Therefore, similar sequences might adopt identical structures, and distantly related sequences may still fold into similar structures (Chothia and Lesk 1986; Sander and Schneider 1991). The theoretical basis of this prediction method is that the sequences with more than 30% identity over an alignment of 80 residues or more may adopt the same basic structure. But sequences which have less than 30% sequence identity can have very different structures (Chothia and Lesk 1986). Homology modeling provides information about the spatial arrangement of residues in protein structure, which can serve as guide to design new experiments like site-directed mutagenesis.

Homology modeling is a multistep process and can be summarized in seven steps. Mostly in all the steps, choices have to be made. The best one has to be chosen from multiple seemingly similar choices:

1. Template identification and amino acid sequence alignment – it involves the alignment of the target sequence with unknown structure and the template or templates of known protein structure.
2. Alignment correction – alignment needs to be checked with caution before proceeding for structure prediction as the quality of the structure produced depends on the target-template alignment.
3. Backbone generation – it involves structure prediction of the core region comprising mainly the secondary structure elements (helices and strands) of the target. If more than one template structures are used, the atomic position framework having average position of atoms is calculated by superimposing all the structures in 3D. The template contribution in the process is weighted according to the similarity with the target sequence. If there is more similarity, more is the weight.
4. Loop modeling – it involves structure prediction of loop regions, which are usually not conserved, and thus requires more sophisticated prediction algorithms, the simplest being spare parts algorithm, which uses database of known loop structures from other proteins.
5. Side-chain modeling and optimization – it involves prediction of the side-chain atoms using side-chain rotamer library resulting in filling of available space in the interior of the protein without having internal clashes with other protein atoms.
6. Model optimization – it involves slight changes in the atomic position to produce a lower-energy model using energy minimization software.
7. Model validation – it involves validation of the predicted structure. It checks the accuracy of the predicted structure. The backbone dihedral angles of the predicted structure should fall in the allowed regions, and the hydrophobic core should be compactly packed. The structure should have minimum free energy.

There are various resources available for structure validation. Some of them are enumerated below:

1. MolProbity – It validates structure of the uploaded file, using all-atom contact analysis tools and updated geometrical criteria for ϕ-ψ, side-chain rotamer, and C-β deviations (Chen et al. 2010, 2015).
2. PDBsum – It summarizes all protein structures including validation checks (de Beer et al. 2014).
3. Procheck Structure validation suite – It is a program that checks the stereochemical quality of a protein structure (Laskowski et al. 1993, 1996).
4. CheckMyMetal – It checks for metal-binding site and validates it (Zheng et al. 2017).

Table 10.2 List of comparative modeling software

S. no.	Name	Method	Description
1.	MODELLER (Webb and Sali 2016)	Satisfaction of spatial restraints	Stand-alone program
2.	EasyModeller (Kuntal et al. 2010)	GUI to MODELLER	
3.	BHAGEERATH-H (Jayaram et al. 2014)	Combination of *ab initio* folding and homology methods	Automated web server
4.	SWISS-MODEL (Biasini et al. 2014)	Local similarity or fragment assembly	
5.	3D-JIGSAW (Bates et al. 2001)	Local similarity or fragment assembly	
6.	HHpred (Söding et al. 2005)	Template detection, alignment, 3D modeling	
7.	ESyPred3D (Lambert et al. 2002)	Template detection, alignment, 3D modeling	
8.	PROTINFO (Hung et al. 2005)	Minimum perturbation, loop building, 3D modeling	
9.	RaptorX (Kallberg et al. 2014)	Automated web server and downloadable program	
10.	PROTEUS2 (Montgomerie et al. 2008)	Comprehensive protein structure prediction and structure-based annotation	

5. ProSA-web – It gives quality scores of a protein in the context of all known protein structures, and problematic parts of a structure are shown in a 3D molecule viewer (Wiederstein and Sippl 2007).
6. NQ-Flipper – It recognizes unfavorable rotamers of asparagine and glutamine residues in protein structures obtained from X-ray crystallography, NMR, or modeling studies (Weichenberger and Sippl 2007).
7. Uppsala Electron Density Server – It generates density maps (Kleywegt et al. 2004).
8. SFCheck – It helps to validate the experimental structure factors associated with an X-ray diffraction experiment (Vaguine et al. 1999).
9. Verify3D – It is a structure evaluation server (Eisenberg et al. 1997).
10. PROSESS – It is a protein structure evaluation suite and server (Berjanskii et al. 2010).

10.6.2 Comparative Modeling Software

There are many comparative modeling software available (Table 10.2). Some are stand-alone, while others are automated web servers.

10.6.3 Fold Recognition Methods

Fold recognition is about searching the most compatible fold that the target protein might adopt from a library of known folds (known protein structures), using both sequence and structural information. Fold recognition uses alignment of the target sequence with one or more distantly related sequences of known structures and can be considered as extension of comparative modeling to discover distant relationships. Fold is detected even when there is no significant sequence similarity to any protein of known structure. Thus, the distant structural and evolutionary relationship is detected with separation from chance sequence similarities associated with the shared fold.

Fold recognition methods are effective because protein folds are limited in nature, mostly because of evolution but also due to constraints imposed by the polypeptide chain's chemistry. Hence, it is likely that a protein with similar fold to the target has already been experimentally studied and can be found in PDB.

10.6.4 Critical Components of Fold Recognition Techniques

1. Useful alignment between sequences and distantly related known structures
2. Selection criteria for identifying native like sequence-structure combinations
3. Sets of energy functions to provide a realistic description of protein-solvent systems

Fold recognition methods can be broadly classified into profile-based methods and threading. The profile-based fold recognition approach (Bowie et al. 1991) involves fitting of the physicochemical properties of the amino acids of the target protein with the environment in which they are placed in the modeled structure. In profile representation, each amino acid in the structure is labeled as either buried (protein core) or exposed (surface), whether it is part of α-helix or β-sheet (i.e., its local secondary structure) and/or its conservation (evolutionary information). The 3D representation describes a structure as a set of interatomic distances; although it is much richer and more flexible, it is harder for alignment calculation. The similarity in sequence detected by amino acid substitution matrices is added with structural information. For example, the three-dimensional position-specific scoring matrix (3D-PSSM; Kelley et al. 1999) uses both – the fold library structures which are described in terms of ordinary 1D sequence profiles generated by position-specific iterated basic local alignment search tool (PSI-BLAST; Altschul et al. 1997; Jones and Swindells 2002) and the 3D profiles holding secondary structure and solvation potential information. The secondary structure component describes the similarities between secondary structures of the predicted and of the member in fold library, while the solvation potential takes account of the tendency of hydrophobic amino acids to bury in hydrophobic core. Thus, this method requires a sequence-structure alignment. It can be done by using PSI-BLAST, which constructs a multiple sequence alignment followed by creation of a profile or a PSSM customized to the query to search matches in the database and estimation of statistical significance (E-values). PSI-BLAST detects weak but biologically meaningful relationships between proteins. Thus, this method is useful in detecting distant homologs.

The term threading coined in 1992 (Jones et al. 1992) is a fold recognition method to model proteins that have the same fold as proteins of known structures but do not have significant sequence similarity. It utilizes statistical information to draw relationship between existing structures in the PDB and the protein sequence to be modeled. Each amino acid in the target sequence is threaded (i.e., placed and aligned) to a position in the template structure, and fitting is evaluated. The best-fit template is selected and utilized for target's model building. Protein threading is grounded on two observations – first, the number of folds is limited in nature and secondly, in past few years most of the new structures deposited exhibited similar structural folds to ones already existing in the PDB. Sequences are fitted directly onto the backbone coordinates in 3D space including specific pair interactions explicitly from the library of protein folds derived from the database of known protein structures. Each fold can be considered as a chain tracing through space irrespective of the sequence. The fitting of the target with the template fold is optimized to allow for relative insertions and deletions in loop regions, and energy of each possible fit (threading) is calculated by summing the pairwise interactions and the solvation energy. The library of folds is then ranked in ascending order of total energy, and the lowest-energy fold is taken as the most probable match. Usually, protein threading consists of four steps:

1. Selection of template protein structure from the protein structure databases such as PDB (Rose et al. 2017), FSSP (Holm and Sander 1996), SCOP (Lo Conte et al. 2000), or CATH (Knudsen and Wiuf 2010), after removing protein structures with high sequence similarities.
2. Designing of a good scoring function to measure the fitness between target sequences and templates based on the knowledge of the known sequence-structure relationships. It should contain pairwise potential, mutation potential, secondary structure compatibilities, gap penalties, and environment fitness potential. The energy function quality relates to the alignment accuracy.
3. Threading alignment – it aligns target sequence and structure templates utilizing the designed scoring function. This is crucial for threading-based structure prediction programs that take pairwise contact potential into consideration. Alternatively, a dynamic programming algorithm is used.
4. Threading prediction – statistically the most probable threading alignment is selected for construction of target structure. The target sequence's backbone atoms are placed at the positions aligned with the backbone of structural template.

10.6.5 Comparison with Homology Modeling

Homology modeling and protein threading are template-based methods, which require knowledge from previously known structures of protein. In case of targets

Table 10.3 List of fold recognition software

S. no.	Name	Method	Description
1.	RaptorX (Kallberg et al. 2014)	Integer programming based fold recognition, probabilistic graphical models, statistical inference	Stand-alone program
2.	SUPERFAMILY (Madera et al. 2004)	Hidden Markov model	Automated web server/stand-alone program
3.	HHpred (Söding et al. 2005)	HHsearch, pairwise comparison of hidden Markov models	Automated web server
4.	Phyre and Phyre2 (Kelley et al. 2015)	Multi-templates, *ab initio* modeling	
5.	MUSTER (Wu and Zhang 2008)	Dynamic programming and sequence profile-profile alignment	
6.	SPARKS-X (Yang et al. 2011)	Probabilistic-based, fold recognition according to sequence profiles and structural profiles	
7.	BioShell Threader (Gniewek et al. 2014)	Profile-to-profile dynamic programming algorithm, sequence profiles and secondary structure profiles	
8.	I-TASSER (Yang and Zhang 2015)	Iterative Threading ASSEmbly Refinement – threading and *ab initio* method	
9.	pGenTHREADER (Lobley et al. 2009)	Sequence profile and predicted secondary structure	
10.	ORION (Ghouzam et al. 2016)	Fold recognition and structure prediction using evolutionary hybrid profiles	
11.	FALCON (Wang et al. 2016)	A position-specific hidden Markov model, iterative refining of dihedral angles	

with available homologous protein structure, homology modeling is used, but when only fold-level homology exists, threading is used for model generation. In other words, homology modeling handles easier targets, while protein threading handles harder targets.

Homology modeling utilizes sequence template and sequence homology in prediction, while protein threading utilizes structural template and extracts both sequence and structure information from the alignment. In the absence of significant homology, protein threading predicts based on the structural information.

In case of low sequence identity (<25%) in a sequence alignment, homology modeling may not produce reliable prediction. In such cases, protein threading could generate a good prediction if a distant homology is found for the target.

10.6.6 Fold Recognition Software

Many fold prediction software are now available (Table 10.3).

10.6.7 *Ab Initio* Structure Prediction

Ab initio modeling (Klepeis et al. 2005; Liwo et al. 2005) or *de novo* modeling (Bradley et al. 2005b), or physics-based modeling (Oldziej et al. 2005), or free modeling (Bradley et al. 2005a; Jauch et al. 2007) is a fundamental test of our knowledge of protein folding, how and why a protein adopts a specific structure out of many possibilities. *Ab initio* structure prediction uses the understanding of physicochemical principles of protein folding in nature and directly applies it to predict the native conformation of a protein from the amino acid sequence alone without the use of framework of earlier known structures, i.e., predicts from the scratch. It uses physical science theories like quantum mechanics and statistical thermodynamics.

Usually, the easiest way to predict the structure of a protein is to find a high-resolution structure of its homolog (analog in some cases) and use its framework to build model, which is the case of template-based modeling. This cannot be used many times because the corresponding protein structure might not be available as the protein structures lag far behind the protein sequences. Plausible, due to technical difficulties, intensive labor and time costs of the experimental structure determination, whereas an exponential increase in protein sequences can be attributed to the tremendous success of the genome sequencing projects. In such cases, computer-based algorithm efficient to predict 3D structures directly from sequences can be used to bridge the big gap between the number of protein sequences and the availability of their corresponding structures. A lot of advancement is needed in *ab initio* methods to handle the enormous system made of proteins in their natural solvation environment, which involves accurate calculations for thousands of atoms in 3D space.

Ab initio modeling is based on the consideration that all of the necessary information for a folding of protein into native conformation resides in its amino acid sequence. In the absence of large kinetic barriers in the free energy landscape, the protein's native conformation is the lowest free energy conformation for its sequences (Anfinsen 1973) with a few exceptions (Baker and Agard 1994). The protein folding is actually governed by the physical forces acting on the atoms of the protein, and thus the most accurate way of structure prediction is in consideration of all-atom model subjected to the physical forces. However, such a representation that contains all atoms of the protein and surrounding solvent molecules increases the complexity and makes the solution computationally expensive, which is beyond the current computational capacity. Moreover, the representation of huge number of atoms and the interactions between them might not be necessary during the initial phase of the search that is far from the native conformation. So, reduced representations of the polypeptide chain are used to reduce calculations and limit the conformational space to manageable size. This can be done in various ways:

1. Use of implicit solvent models instead of explicit solvent models.
2. Use of united atom representations where hydrogens are drawn into their base carbon, oxygen, and nitrogen atoms.

3. Representation of the side chains by limited set of conformations prevailing in PDB structures.
4. Replacement of the side-chain atoms completely by locating the side-chain properties either at the centroid of the side chain or at the β-carbon, which results in averaging of the side-chain degrees of freedom and enhances the performance at the loss of some degree of specificity.
5. The conformations available to the polypeptide backbone can be restricted to discrete values that are commonly observed in existing structures. It can be done either by using a small set of ϕ-ψ pairs by selecting pairs from an ideal set from predicted regular secondary structure or by using fragments from existing protein structures. The torsion angles can be restricted based on the knowledge that in particular local structures, amino acids prefer certain torsion angle pairs.

Thus, *ab initio* modeling requires a suitably defined protein representation with compatible energy functions that capture the most significant interactions that drive the folding of the protein sequence toward the native structures and efficient and reliable algorithms to search the conformational space in that protein representation to minimize the energy function. The conformations that minimize the energy function are considered likely structures of the protein in native conditions. Thus, all *ab initio* methods conduct a conformational search using an efficient energy function, generate a number of possible conformations, and select a final model from them. Therefore, success in *ab initio* modeling depends on three key features:

1. Accurate free energy function sufficiently close to the true potential for the native state that results from the native structure of a protein corresponds to the thermodynamically most stable state, i.e., lowest free energy minima among all possible conformations.
2. Efficient search method that swiftly does the conformational search to identify the low-energy states.
3. Efficient native-like model selection criteria from all the protein conformations.

10.6.8 Energy Functions

There are two kinds of energy functions – the physics-based energy functions and the knowledge-based energy functions. In the physics-based *ab initio* methods, all atoms are represented by their atom types, and only the number of electrons is significant. The interactions amid atoms are based on quantum mechanics, electron charge, and Planck constant (the fundamental parameters of the coulomb potential) (Hagler et al. 1974; Hagler and Lifson 1974; Weiner et al. 1984). However, even for small protein structure prediction, the complete use of quantum mechanics requires extensive computational resources. So in practice, the *ab initio* protein modeling uses a compromised force field with a huge number of selected atom types (Weiner et al. 1984; Hagler and Lifson 1974). The physics-based force fields which take all atoms into consideration include AMBER (Weiner et al. 1984; Cornell et al. 1995;

Duan and Kollman 1998; Kaus et al. 2013), CHARMM (Brooks et al. 1983; Neria et al. 1996; MacKerell et al. 1998; Hynninen and Crowley 2014), and OPLS (Jorgensen and Tiradorives 1988; Jorgensen and Tirado-Rives 1998; Jorgensen et al. 1996; Kaminski et al. 2001), with the major difference among them being the choice of atom types and interaction parameters. These potentials contain information about the bond lengths, angles, torsion angles, van der Waals, and electrostatic interactions, while the knowledge-based energy functions use the empirical energy terms obtained from the statistics of the existing 3D structure of proteins in PDB and can be divided into two categories (Skolnick 2006). One of them contains the generic and sequence-independent terms like the hydrogen bond and the local backbone rigidity of a polypeptide chain (Zhang et al. 2003), while the other contains amino acid or protein sequence-dependent terms, like pairwise residue contact potential (Skolnick et al. 1997), distance-dependent atomic contact potential (Samudrala and Moult 1998; Shen and Sali 2006; Lu and Skolnick 2001; Zhou and Zhou 2002), and secondary structure propensities (Zhang et al. 2003, 2006; Zhang and Skolnick 2005). The most successful *ab initio* methods using the knowledge-based energy functions are ROSETTA (Simons et al. 1997; Bender et al. 2016) and TASSER (Zhang and Skolnick 2004; Yang and Zhang 2015).

10.6.9 Conformational Search Methods

The success of *ab initio* modeling is dependent on conformational search method. It should be efficient enough to find the global minimum energy structure for a particular energy function in rugged energy landscape of protein conformational space (containing many energy barriers). The conformational search methods include the following:

1. Monte Carlo simulations – Simulated annealing (SA) is the most commonly used method (Kirkpatrick et al. 1983; Lee 1993).
2. Molecular dynamics simulations.
3. Genetic algorithm – Conformational space annealing (CSA) is one of the most widely used genetic algorithms (Lee et al. 1998).
4. Mathematical optimization (Klepeis et al. 2005; Klepeis and Floudas 2003).

Ab initio structure prediction is challenging because the current potential functions have limited accuracy, and the conformational space to be searched is vast. The successful modeling is limited to small proteins, less than 100 residues. Many *ab initio* methods have shown improvement in protein structure prediction by using reduced representations, coarse search strategies, and simplified potentials (Simons et al. 1997; Samudrala et al. 1999; Oldziej et al. 2005; Pillardy et al. 2001).

Table 10.4 List of *ab initio* modeling software

S. no.	Name	Method	Description
1.	UniCon3D (Bhattacharya et al. 2016)	De novo modeling, united-residue conformational search, stepwise probabilistic sampling	Stand-alone program
2.	QUARK (Xu and Zhang 2012)	Monte Carlo fragment assembly	Automated web server
3.	CABS-FOLD (Blaszczyk et al. 2013)	De novo modeling can also use alternative templates (consensus modeling)	
4.	PEP-FOLD (Lamiable et al. 2016)	De novo modeling, based on a HMM structural alphabet	
5.	BHAGEERATH (Jayaram et al. 2014)	Predicts protein structure using *ab initio* folding	
6.	ROBETTA (Kim et al. 2004)	Rosetta homology modeling and *ab initio* fragment assembly	
7.	I-TASSER (Yang and Zhang 2015)	Iterative Threading ASSEmbly Refinement – threading and *ab initio* method	
8.	Rosetta@home (Bender et al. 2016)	Distributed-computing implementation of Rosetta algorithm	Downloadable program

10.6.10 *Ab Initio* Modeling Software

Many software are available for *ab initio* modeling (Table 10.4).

10.7 Role of Structural Bioinformatics in Drug Discovery and Health Care

The recent advances in the sector of health care and disease prevention have come as a collimated effort of understanding disease biology and development of efficacious drug molecules to overcome the irregularity. The field of drug discovery dates back to the late 1800s when chemists at Bayer synthetically synthesized the first drug aspirin (Desborough and Keeling 2017; Sneader 2000). Since then the drug discovery pipeline has traversed from being highly dependent on identifying inhibitors of target molecule inferred from crystallographic structures (Beddell et al. 1976; Newman and Cragg 2012) to a paradigm of high-throughput format using computational as well as wet lab resources (Doman et al. 2002). The trend has arisen concurrently with the demand for new medicinal compounds for emerging diseases as well as the rising cost and the financial risks while introducing a drug into the market. The estimated value of introducing a new drug into the market has surged up from $400 million to $2.6 billion (DiMasi et al. 2003; Basak 2012) and has further

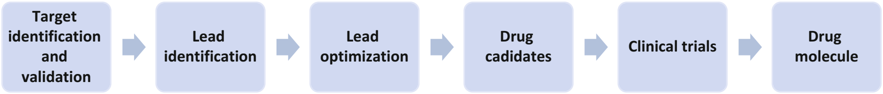


Fig. 10.5 The pipeline of rational drug design

risen. The issue is also thwarted by frequent failure of drugs at the clinical trial stages due to their insufficiency to meet the adsorption, distribution, metabolism, excretion, and toxicity (ADMET) criterions or even the withdrawal of marketed drugs due to unforeseen implications on their use. The current scenario calls for increasing the productivity of the pharma sector by screening for new drug targets or effector molecules that can elicit the desired effects as well as sustain the strict criterion laid by monitoring agency.

Efforts to integrate structural biology and drug discovery pipelines through computer-aided drug design (CADD) are underway. There are various steps involved in rational drug design (Fig. 10.5):

1. Target identification and validation – it involves understanding of disease biology and identification of potential drug target, followed by testing of the target molecule for therapeutic potential, i.e., assessing target druggability, obtaining structural information of target, and if not available predicting the structure or using ligand information.
2. Lead discovery – identification of drug candidate that interacts with the target. It involves generation (de novo ligand design) and screening of large chemical libraries to derive smaller sets of potential drug candidates (leads) that can be validated experimentally. Virtual high-throughput screening (HTS) is used for lead discovery.
3. Lead optimization – it focuses on improving efficacy of effector molecule by improving their drug metabolism and pharmacokinetics (DMPK) properties also called as ADMET properties. It uses either docking, side-chain modeling, or pharmacophore modeling depending on the particular requirement.
4. Clinical trials – the investigational new drug has to pass the clinical trials before it can come to market.

The approaches in the computational drug discovery can be divided into two categories:

1. *Structure-based drug design (SBDD)*

   SBDD relies on availability of structural information of a target molecule, which is used to design potential inhibitors. The protein structural data is used to predict the type of ligands that will interact with a given target. Considerations include the importance of protein in a disease, the involved pathway, availability of its structure or ease of prediction, and its ability to bind small molecules.

2. *Ligand-based drug design (LBDD)*

   LBDD uses information about the known drugs and compound libraries in cases when the structural information of target is not available. It is an indirect drug design; the knowledge of other molecules that bind to the biological target of interest is used. A pharmacophore model can be derived that defines the minimum necessary structural characteristics a molecule must possess in order to bind to the target. The quantitative structure-activity relationships (QSAR) are used to predict the activity of new analogs. These QSAR relationships derive a correlation between calculated properties of molecules and their experimentally determined biological activity.

The choice of method to be used for finding effector molecule depends on the availability of information – the structural knowledge of the target proteins or its homologs, existence of any previously known drugs or compound libraries, and the required computational resources. In both approaches, each step moves through numerous iterative cycles in order to present the best possible prediction of a target or the ligand molecule and their interaction.

Bioinformatics aids in the analysis of sequences and structure; in the development of algorithms and software for modeling the drug-target interaction, building the compound libraries, and easy retrieval system; and in the development of high-throughput screening (HTS) system (Matter et al. 2001; Scapin 2006; Edwards 2009; Cheng et al. 2013; Lagorce et al. 2015; Villoutreix 2016; Daina et al. 2017; Miteva and Villoutreix 2017; Lagorce et al. 2017).

10.7.1 Target Identification and Validation

The first step of SBDD methodology involves gathering all information on a target of interest: a thorough understanding of the mechanism of disease progression and the involvement of the target protein in particular stage/stages. The implicated proteins are identified, cloned, purified, and crystallized for solving their structure through X-ray crystallography, NMR, or a relevant structure prediction method, in case of experimental structure determination failure. The structure of the target molecule (usually a protein) is used to analyze its druggability. Not all proteins can act as valid drug targets. For being an effectual drug target, the protein must possess an active site that can be inhibited. In other words, protein should accommodate ligands – either analogues of the natural ligand or other small molecules in the active site by electrostatic interactions. The likelihood of finding suitable drug targets can be assessed using surface and active site properties like volume, charge, and shape that can be calculated using tools like CAST (Liang et al. 1998), CASTp (Dundas et al. 2006), GRASP (Nicholls et al. 1991), VICE (Tripathi and Kellogg 2010), POCKET (Levitt and Banaszak 1992), and TRAPP web server (Stank et al. 2017). The procedure of identifying targets also entails the possibility of having no functional overlap between the drug target and other host proteins, which is inferred using phylogenetic relationships between the target and host proteins. Structure

Table 10.5 General properties of lead compounds

Property	Definition/requirement
Potency	Ability to produce a desirable pharmacological response
Bioavailability	Ability to pass through multiple barriers like the gastrointestinal tract and liver and further get absorbed into the bloodstream
Stability or half-life	Capability of the compound to remain in the bloodstream for adequate time to elicit a significant pharmacological response
Safety	Specificity of the drug candidate to the target and minimal off-target response
Pharmaceutical acceptability	Chemical parameters relating to the cost of synthesis, stability at various temperatures and pH conditions, rate and level of solubility in an aqueous medium, etc.

activity relationship homology, SARAH (Frye 1999), based searches analyze and group proteins based on sequence similarity and their ability to bind a ligand in high throughput manner. The proposed drug targets must pass through a validation step in order to qualify for the next rounds of drug discovery process. Possible means to validate drug targets involve gene disruption by deletion or suppression of expression by RNA interference (RNAi) studies (Smith 2003; Ghosh et al. 2017) or site-directed mutagenesis (Zeng et al. 2010). The reverse (Eyers et al. 1998) and forward (Choi et al. 2014) chemical genetic screening is focused on creating or isolating mutants of target proteins sensitive to known inhibitors.

10.7.2 Lead Identification

The identification of the lead molecule involves the search for a substance with desirable biological activity, which may serve as drug (Di et al. 2009). The ligand molecule that binds only to the target molecule with medium or high potency is needed to ensure that only the safest and the most bioactive compounds pass through the trail cycles. This further reduces the risk of failures at later stages of the discovery process. The drug molecule should have some basic properties as listed in Table 10.5.

Appropriate assay systems to monitor the target-ligand binding should highlight the binding preferences of a particular target molecule and consider the physiological outcome expected for a living system and also pass well on criterion of cost and reproducibility and hold potential to assess the effects of drug. Counter-screening approaches using bioinformatics analysis rely on finding all possible targets (Davies et al. 2000). A vast pool of biochemical knowledge exists on protein-ligand interaction and protein-analog interaction. The existing knowledge of a target binding to a drug can be applied to a related target protein. Thus, focused set of library are required if the structure of the target is known. This will define particular set of ligands, i.e., focused on one region of the chemical space. Various chemical leads have been derived using structural similarity, which includes the development of enzyme inhibitors like angiotensin-converting enzyme, neutral endopeptidase, and thermolysin (Roques 1985; Oefner et al. 2000).

If the information about the binding properties of drug target is less or not available, diverse chemical libraries are required for efficient lead discovery. The diversity can be defined by comparing the lead molecules based on molecular descriptors (functional groups) and how the chemical space is filled. The initial screening of lead molecules requires computational approach to identify the most suitable lead amongst the vast databases. A high-throughput approach of virtual screening has been pioneered over the years to identify a suitable lead. It is divided into two categories – target-based virtual screening and ligand-based virtual screening.

10.7.3 Target-Based Virtual Screening

Once the target molecule is identified and biochemically characterized to detect its active site, its ligand-binding pocket is screened for finding a suitable ligand from a library of existing compounds. For this, docking tools like AutoDock (Osterberg et al. 2002), DOCK (Kuntz et al. 1982), FlexX (Rarey et al. 1996), Glide (Friesner et al. 2004), LigandFit (Venkatachalam et al. 2003), MOE-Dock (Corbeil et al. 2012), and UCSF Dock (Allen et al. 2015) are used (Pagadala et al. 2017; de Ruyck et al. 2016; Lohning et al. 2017). Boltzmann-weighted potentials of mean force are derived from the structural data of protein-ligand complex, and a scoring function is used to score and identify candidates. The approaches to analyze the empirical changes in the free energy and other changes in thermodynamic parameters on target binding to different ligand are taken into consideration. Finally, a Gaussian method to estimate the volume exclusion and solvent forces applies the Poisson-Boltzmann equation to small and larger molecules. Thus, if the target structural data is available, these algorithms can be applied to identify the interacting ligands that can serve as candidate drugs based on goodness of fit. Relenza and Captopril are the well-known drugs developed in this manner.

The important requirement of drug discovery is the availability of compound libraries with small drug-like molecules. For becoming a potential drug candidate, the ligand must follow the Lipinski's rule of five. It comprises set of physical parameters designed to predict the bioavailability of a molecule and other important pharmaceutical characteristics. To ensure maximal bioavailability, a compound must fulfill the following parameters of Lipinski's rule of five (Lipinski 2004; Oprea et al. 2001; Lipinski et al. 2001):

1. The molecular weight should be less than 500 daltons.
2. The compound's lipophilicity or the logP value (the logarithm of the partition coefficient between water and 1-octanol) is less than 5.
3. The number of groups in the molecule that can donate hydrogen atoms to hydrogen bonds is less than 5 (the total number of oxygen-hydrogen and nitrogen-hydrogen bonds).

4. The number of groups that can accept hydrogen atoms to form hydrogen bonds is less than 10 (all nitrogen and oxygen atoms).

Many variations in this rule have been introduced to increase the druglikeness (Ghose et al. 1999; Xu and Stevenson 2000; Avdeef 2001; Tice 2001, 2002; Veber et al. 2002; Congreve et al. 2003; Lovering et al. 2009; Meanwell 2011; Leeson 2012; Vallianatou et al. 2015; Meanwell 2016; Shekhawat and Pokharkar 2017). Nonetheless, the abovementioned measures form the basis of the well-established set of ADMET properties.

Recent developments in the field of pharmacokinetics have focused on creating alternative methods to design parameters that benchmark the properties a compound should possess for entering the lead discovery process. To quantify the druglikeness, the concept of desirability was implemented which provides a quantitative metric for assessing druglikeness called as quantitative estimate of druglikeness (QED; Harrington 1965; Derringer and Suich 1980; Bickerton et al. 2012). The QED approach assigns desirability values to a molecule for its assessment as drug based on categorical parameters built on desirable functions. Further, the functions are summed to provide a single numerical QED value ranging from 0 to 1 signifying an unfavorable to a highly favorable candidate. The desirability is simple but powerful approach for multi-criteria optimization. It can be implemented in numerous drug discovery applications like selection of compound, library design, molecular target prioritization, permeation of central nervous system, and reliability estimation of the screening data. It takes several numeric parameters measured on different scales and labels each by an individual desirability function, which are then combined into a single dimensionless score. A series of desirability functions (d) are derived for a particular compound, each of which corresponds to a different molecular descriptor. The individual desirability functions are combined into the QED by taking the geometric mean of the individual functions, as shown in the following QED equation:

$$\mathrm{QED} = \exp\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} \ln\ di\right)$$

For deriving the desirability, the eight widely used molecular properties include molecular weight, octanol-water partition coefficient, number of hydrogen bond donors, number of hydrogen bond acceptors, the number of aromatic rings, number of rotatable bonds, molecular polar surface area, and number of structural alerts (Bickerton et al. 2012). The selection is based on their relevance in determining druglikeness.

10.7.4 Ligand-Based Virtual Screening

The ligand-based screening approach uses the classification of existing or virtual ligands in a library based on 3D similarity or pharmacophore matching. It involves various software to query the chemical libraries. Once the lead is identified, it needs to be optimized for increasing its efficacy and specificity to the target.

10.7.5 Lead Optimization

Chemical leads that pass the initial screening process may still require further optimization to improve their potency. The inherent problem of solving complex crystal structures of target-ligand having variable side chains and problems in determining the kinetic parameters for target-ligand derivative binding makes it challenging to perform the task in high-throughput manner. The computational approaches for lead optimization depend on designing derivatives of lead compounds by addition of various side chains followed by prediction of 3D models for target-ligand complexes and their virtual ADMET profiles (Cheng et al. 2013; Honorio et al. 2013; Meanwell 2011).

The final optimized candidate drugs (CDs) are then passed through sets of clinical trials involving preclinical phase (animal model studies), phase I (studies on normal healthy human volunteers), phase II (selection of dose regime and the evaluation of safety and efficacy in patients), phase III (testing on large population of patients with potential drug and placebo – the commercial launch can be taken after this by regulatory authorities), and phase IV (monitoring the long-term effects or any adverse reactions reported by doctors). Thus, the drug discovery itself is a time taking and lengthy process, which therefore requires the aid of computational methods to cut down the time and cost at various steps in the process. This requires development of efficient prediction algorithms, methods to efficiently model the target-ligand interaction, efficient software, databases, and retrieval tools.

Thus, structural bioinformatics is not only an integral part of structural biology but is also indispensable in drug discovery and health care.

References

Adrian M, Heddi B, Phan AT (2012) NMR spectroscopy of G-quadruplexes. Methods (San Diego, Calif) 57(1):11–24. https://doi.org/10.1016/j.ymeth.2012.05.003

Agarwal T, Jayaraj G, Pandey SP, Agarwala P, Maiti S (2012) RNA G-quadruplexes: G-quadruplexes with "U" turns. Curr Pharm Des 18(14):2102–2111

Ahmed YL, Ficner R (2014) RNA synthesis and purification for structural studies. RNA Biol 11 (5):427–432. https://doi.org/10.4161/rna.28076

Allen WJ, Balius TE, Mukherjee S, Brozell SR, Moustakas DT, Lang PT, Case DA, Kuntz ID, Rizzo RC (2015) DOCK 6: Impact of new features and current docking performance. J Comput Chem 36(15):1132–1156. https://doi.org/10.1002/jcc.23905

Altschul SF, Madden TL, Schaffer 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

Amunts A, Brown A, Toots J, Scheres SH, Ramakrishnan V (2015) Ribosome. The structure of the human mitochondrial ribosome. Science 348(6230):95–98. https://doi.org/10.1126/science.aaa1193

Amzel LM, Poljak RJ (1979) Three-dimensional structure of immunoglobulins. Annu Rev Biochem 48:961–997. https://doi.org/10.1146/annurev.bi.48.070179.004525

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

Arcella A, Portella G, Ruiz ML, Eritja R, Vilaseca M, Gabelica V, Orozco M (2012) Structure of triplex DNA in the gas phase. J Am Chem Soc 134(15):6596–6606. https://doi.org/10.1021/ja209786t

Arieti F (2014) Structural studies of RNA-binding domains

Arnott S (1970) Crystallography of DNA: difference synthesis supports Watson-Crick base pairing. Science 167(3926):1694–1700

Arnott S, Chandrasekaran R, Hukins DW, Smith PJ, Watts L (1974a) Structural details of double-helix observed for DNAs containing alternating purine and pyrimidine sequences. J Mol Biol 88 (2):523–533

Arnott S, Chandrasekaran R, Marttila CM (1974b) Structures for polyinosinic acid and polyguanylic acid. Biochem J 141(2):537–543

Arnott S, Chandrasekaran R, Leslie AG (1976) Structure of the single-stranded polyribonucleotide polycytidylic acid. J Mol Biol 106(3):735–748

Artusi S, Perrone R, Lago S, Raffa P, Di Iorio E, Palu G, Richter SN (2016) Visualization of DNA G-quadruplexes in herpes simplex virus 1-infected cells. Nucleic Acids Res 44 (21):10343–10353. https://doi.org/10.1093/nar/gkw968

Asano S, Engel BD, Baumeister W (2016) In situ cryo-electron tomography: a post-reductionist approach to structural biology. J Mol Biol 428(2 Pt A):332–343. https://doi.org/10.1016/j.jmb.2015.09.030

Avdeef A (2001) Physicochemical profiling (solubility, permeability and charge state). Curr Top Med Chem 1(4):277–351

Bae S, Kim D, Kim KK, Kim YG, Hohng S (2011) Intrinsic Z-DNA is stabilized by the conformational selection mechanism of Z-DNA-binding proteins. J Am Chem Soc 133 (4):668–671. https://doi.org/10.1021/ja107498y

Bai X-C, McMullan G, Scheres SHW (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40(1):49–57. https://doi.org/10.1016/j.tibs.2014.10.005

Baker D, Agard DA (1994) Influenza hemagglutinin: kinetic control of protein function. Structure 2 (10):907–910

Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294 (5540):93–96

Basak SC (2012) Chemobioinformatics: the advancing frontier of computer-aided drug design in the post-genomic era. Curr Comput-Aided Drug Des 8(1):1–2

Basu HS, Feuerstein BG, Zarling DA, Shafer RH, Marton LJ (1988) Recognition of Z-RNA and Z-DNA determinants by polyamines in solution: experimental and theoretical studies. J Biomol Struct Dyn 6(2):299–309. https://doi.org/10.1080/07391102.1988.10507714

Bates PA, Kelley LA, MacCallum RM, Sternberg MJ (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins Suppl 5:39–46

Beddell CR, Goodford PJ, Norrington FE, Wilkinson S, Wootton R (1976) Compounds designed to fit a site of known structure in human haemoglobin. Br J Pharmacol 57(2):201–209

Bender BJ, Cisneros A 3rd, Duran AM, Finn JA, Fu D, Lokits AD, Mueller BK, Sangha AK, Sauer MF, Sevy AM, Sliwoski G, Sheehan JH, DiMaio F, Meiler J, Moretti R (2016) Protocols for molecular modeling with Rosetta3 and RosettaScripts. Biochemistry 55(34):4748–4763. https://doi.org/10.1021/acs.biochem.6b00444

Berjanskii M, Liang Y, Zhou J, Tang P, Stothard P, Zhou Y, Cruz J, MacDonell C, Lin G, Lu P, Wishart DS (2010) PROSESS: a protein structure evaluation suite and server. Nucleic Acids Res 38(suppl_2):W633–W640. https://doi.org/10.1093/nar/gkq375

Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J (2000a) The protein data bank and the challenge of structural genomics. Nat Struct Biol 7(Suppl):957–959. https://doi.org/10.1038/80734

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000b) The protein data bank. Nucleic Acids Res 28(1):235–242

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 112(3):535–542

Bharat TA, Scheres SH (2016) Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. Nat Protoc 11(11):2054–2065. https://doi.org/10.1038/nprot.2016.124

Bharat Tanmay A, Russo Christopher J, Löwe J, Passmore Lori A, Scheres Sjors H (2015) Advances in single-particle electron cryomicroscopy structure determination applied to sub-tomogram averaging. Structure (London, England:1993) 23(9):1743–1753. https://doi.org/10.1016/j.str.2015.06.026

Bhattacharya D, Cao R, Cheng J (2016) UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics (Oxford, England) 32(18):2791–2799. https://doi.org/10.1093/bioinformatics/btw316

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42(Web Server issue):W252–W258. https://doi.org/10.1093/nar/gku340

Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4(2):90–98. https://doi.org/10.1038/nchem.1243

Binkowski TA, Freeman P, Liang J (2004) pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. Nucleic Acids Res 32(Web Server issue): W555–W558. https://doi.org/10.1093/nar/gkh390

Blake JD, Cohen FE (2001) Pairwise sequence alignment below the twilight zone. J Mol Biol 307 (2):721–735

Blaszczyk M, Jamroz M, Kmiecik S, Kolinski A (2013) CABS-fold: Server for the de novo and consensus-based prediction of protein structure. Nucleic Acids Res 41(Web Server issue): W406–W411. https://doi.org/10.1093/nar/gkt462

Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016):164–170

Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D (2005a) Free modeling with Rosetta in CASP6. Proteins 61(Suppl 7):128–134

Bradley P, Misura KM, Baker D (2005b) Toward high-resolution de novo structure prediction for small proteins. Science 309(5742):1868–1871

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4(2):187–217

Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S (2006) Quadruplex DNA: sequence, topology and structure. Nucleic Acids Res 34(19):5402–5415. https://doi.org/10.1093/nar/gkl655

Campbell NH, Parkinson GN (2007) Crystallographic studies of quadruplex nucleic acids. Methods (San Diego, Calif) 43(4):252–263. https://doi.org/10.1016/j.ymeth.2007.08.005

Chen JL, Greider CW (2005) Functional analysis of the pseudoknot structure in human telomerase RNA. Proc Natl Acad Sci U S A 102(23):8080–8085; discussion 8077–8089. https://doi.org/10.1073/pnas.0502259102

Chen X, Ramakrishnan B, Sundaralingam M (1995) Crystal structures of B-form DNA-RNA chimers complexed with distamycin. Nat Struct Biol 2(9):733–735

Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D: Biol Crystallogr 66(Pt 1):12–21. https://doi.org/10.1107/S0907444909042073

Chen VB, Wedell JR, Wenger RK, Ulrich EL, Markley JL (2015) MolProbity for the masses-of data. J Biomol NMR 63(1):77–83. https://doi.org/10.1007/s10858-015-9969-9

Cheng YK, Pettitt BM (1992) Stabilities of double- and triple-strand helical nucleic acids. Prog Biophys Mol Biol 58(3):225–257

Cheng F, Li W, Liu G, Tang Y (2013) In silico ADMET prediction: recent advances, current challenges and future trends. Curr Top Med Chem 13(11):1273–1289

Choi J, Majima T (2011) Conformational changes of non-B DNA. Chem Soc Rev 40 (12):5893–5909. https://doi.org/10.1039/c1cs15153c

Choi H, Kim JY, Chang YT, Nam HG (2014) Forward chemical genetic screening. Methods Mol Biol 1062:393–404. https://doi.org/10.1007/978-1-62703-580-4_21

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826

Chou SH, Chin KH, Wang AH (2003) Unusual DNA duplex and hairpin motifs. Nucleic Acids Res 31(10):2461–2474

Coimbatore Narayanan B, Westbrook J, Ghosh S, Petrov AI, Sweeney B, Zirbel CL, Leontis NB, Berman HM (2014) The nucleic acid database: new features and capabilities. Nucleic Acids Res 42(Database issue):D114–D122. https://doi.org/10.1093/nar/gkt980

Congreve M, Carr R, Murray C, Jhoti H (2003) A 'rule of three' for fragment-based lead discovery? Drug Discov Today 8(19):876–877

Corbeil CR, Williams CI, Labute P (2012) Variability in docking success rates due to dataset preparation. J Comput-Aided Mol Des 26(6):775–786. https://doi.org/10.1007/s10822-012-9570-1

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J Am Chem Soc 117(19):5179–5197. https://doi.org/10.1021/Ja00124a002

Dahm R (2008) Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum Genet 122(6):565–581. https://doi.org/10.1007/s00439-007-0433-0

Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 7:42717. https://doi.org/10.1038/srep42717

Davies SP, Reddy H, Caivano M, Cohen P (2000) Specificity and mechanism of action of some commonly used protein kinase inhibitors. Biochem J 351(Pt 1):95–105

Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I (2017) CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res 45(Database issue):D289–D295. https://doi.org/10.1093/nar/gkw1098

de Beer TAP, Berka K, Thornton JM, Laskowski RA (2014) PDBsum additions. Nucleic Acids Res 42(D1):D292–D296. https://doi.org/10.1093/nar/gkt940

de Ruyck J, Brysbaert G, Blossey R, Lensink MF (2016) Molecular docking as a popular tool in drug design, an in silico travel. Adv Appl Bioinform Chem 9:1–11. https://doi.org/10.2147/aabc.s105289

Derringer G, Suich R (1980) Simultaneous-optimization of several response variables. J Qual Technol 12(4):214–219

Desai N, Brown A, Amunts A, Ramakrishnan V (2017) The structure of the yeast mitochondrial ribosome. Science 355(6324):528–531. https://doi.org/10.1126/science.aal2415

Desborough MJR, Keeling DM (2017) The aspirin story – from willow to wonder drug. Br J Haematol 177(5):674–683. https://doi.org/10.1111/bjh.14520

Devi G, Zhou Y, Zhong Z, Toh DF, Chen G (2015) RNA triplexes: from structural principles to biological and biotech applications. Wiley Interdiscip Rev RNA 6(1):111–128. https://doi.org/10.1002/wrna.1261

Di L, Kerns EH, Carter GT (2009) Drug-like property concepts in pharmaceutical design. Current Pharm Des 15(19):2184–2194

Dickerhoff J, Haase L, Langel W, Weisz K (2017) Tracing effects of fluorine substitutions on G-Quadruplex conformational changes. ACS Chem Biol 12(5):1308–1315. https://doi.org/10.1021/acschembio.6b01096

DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. J Health Econ 22(2):151–185. https://doi.org/10.1016/S0167-6296(02)00126-1

Doherty EA, Doudna JA (2000) Ribozyme structures and mechanisms. Annu Rev Biochem 69:597–615. https://doi.org/10.1146/annurev.biochem.69.1.597

Dolinnaya NG, Ogloblina AM, Yakubovskaya MG (2016) Structure, properties, and biological relevance of the DNA and RNA G-Quadruplexes: overview 50 years after their discovery. Biochem Biokhimiia 81(13):1602–1649. https://doi.org/10.1134/s0006297916130034

Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 45(11):2213–2221

Dorn M, E Silva MB, Buriol LS, Lamb LC (2014) Three-dimensional protein structure prediction: methods and computational strategies. Comput Biol Chem 53:251–276. https://doi.org/10.1016/j.compbiolchem.2014.10.001

Duan Y, Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 282(5389):740–744. https://doi.org/10.1126/science.282.5389.740

Dudek CA, Dannheim H, Schomburg D (2017) BrEPS 2.0: optimization of sequence pattern prediction for enzyme annotation. PloS One 12(7):e0182216. https://doi.org/10.1371/journal.pone.0182216

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34(Web Server issue):W116–W118. https://doi.org/10.1093/nar/gkl282

Edwards PJ (2009) Current parallel chemistry principles and practice: application to the discovery of biologically active molecules. Curr Opin Drug Discov Dev 12(6):899–914

Eisenberg D (2003) The discovery of the α-helix and β-sheet, the principal structural features of proteins. Proc Natl Acad Sci U S A 100(20):11207–11210. https://doi.org/10.1073/pnas.2034522100

Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 277:396–404

Eyers PA, Craxton M, Morrice N, Cohen P, Goedert M (1998) Conversion of SB 203580-insensitive MAP kinase family members to drug-sensitive forms by a single amino-acid substitution. Chem Biol 5(6):321–328

Fay MM, Lyons SM, Ivanov P (2017) RNA G-quadruplexes in biology: principles and molecular mechanisms. J Mol Biol 429(14):2127–2147. https://doi.org/10.1016/j.jmb.2017.05.017

Ferre-D'Amare AR, Doudna JA (1999) RNA folds: insights from recent crystal structures. Annu Rev Biophys Biomol Struct 28:57–73. https://doi.org/10.1146/annurev.biophys.28.1.57

Floudas CA (2007) Computational methods in protein structure prediction. Biotechnol Bioeng 97(2):207–213. https://doi.org/10.1002/bit.21411

Frank J (2017) Advances in the field of single-particle cryo-electron microscopy over the last decade. Nat Protoc 12(2):209–212. https://doi.org/10.1038/nprot.2017.004

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47(7):1739–1749. https://doi.org/10.1021/jm0306430

Frye SV (1999) Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. Chem Biol 6(1):R3–R7. https://doi.org/10.1016/S1074-5521(99)80013-1

Fukuhara M, Ma Y, Nagasawa K, Toyoshima F (2017) A G-quadruplex structure at the 5′ end of the H19 coding region regulates H19 transcription. Sci Rep 7:45815. https://doi.org/10.1038/srep45815 https://www.nature.com/articles/srep45815#supplementary-information

Furnham N, Holliday GL, de Beer TA, Jacobsen JO, Pearson WR, Thornton JM (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 42(Database issue):D485–D489. https://doi.org/10.1093/nar/gkt1243

Gajarsky M, Zivkovic ML, Stadlbauer P, Pagano B, Fiala R, Amato J, Tomaska L, Sponer J, Plavec J, Trantirek L (2017) Structure of a stable G-hairpin. J Am Chem Soc 139 (10):3591–3594. https://doi.org/10.1021/jacs.6b10786

Galaz-Montoya JG, Ludtke SJ (2017) The advent of structural biology in situ by single particle cryo-electron tomography. Biophys Rep 3(1):17–35. https://doi.org/10.1007/s41048-017-0040-0

Gebetsberger J, Micura R (2017) Unwinding the twister ribozyme: from structure to mechanism. Wiley Interdiscip Rev RNA 8(3). https://doi.org/10.1002/wrna.1402

Ghose AK, Viswanadhan VN, Wendoloski JJ (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J Comb Chem 1(1):55–68

Ghosh A, Bansal M (2003) A glossary of DNA structures from A to Z. Acta Crystallogr D Biol Crystallogr 59(Pt 4):620–626

Ghosh S, Kaushik A, Khurana S, Varshney A, Singh AK, Dahiya P, Thakur JK, Sarin SK, Gupta D, Malhotra P, Mukherjee SK, Bhatnagar RK (2017) An RNAi-based high-throughput screening assay to identify small molecule inhibitors of hepatitis B virus replication. J Biol Chem 292 (30):12577–12588. https://doi.org/10.1074/jbc.M117.775155

Ghouzam Y, Postic G, Guerin P-E, de Brevern AG, Gelly J-C (2016) ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. Sci Rep 6:28268. https://doi.org/10.1038/srep28268

Gniewek P, Kolinski A, Kloczkowski A, Gront D (2014) BioShell-Threading: versatile Monte Carlo package for protein 3D threading. BMC Bioinf 15:22. https://doi.org/10.1186/1471-2105-15-22

Greider CW, Blackburn EH (1985) Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. Cell 43(2 Pt 1):405–413

Griffith JD, Comeau L, Rosenfield S, Stansel RM, Bianchi A, Moss H, de Lange T (1999) Mammalian telomeres end in a large duplex loop. Cell 97(4):503–514

Groll M, Kim KB, Kairies N, Huber R, Crews CM (2000) Crystal structure of epoxomicin: 20S proteasome reveals a molecular basis for selectivity of α‘,β‘-epoxyketone proteasome inhibitors. J Am Chem Soc 122(6):1237–1238. https://doi.org/10.1021/ja993588m

Hagler AT, Lifson S (1974) Energy functions for peptides and proteins. II. The amide hydrogen bond and calculation of amide crystal properties. J Am Chem Soc 96(17):5327–5335

Hagler AT, Huler E, Lifson S (1974) Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. J Am Chem Soc 96 (17):5319–5327

Hall SR (1991) The star file – a new format for electronic data transfer and archiving. J Chem Inf Comp Sci 31(2):326–333. https://doi.org/10.1021/Ci00002a020

Hall SR, Allen FH, Brown ID (1991) The crystallographic information file (Cif) – a new standard archive file for crystallography. Acta Crystallogr A 47:655–685. https://doi.org/10.1107/S010876739101067x

Harrington EC (1965) The desirability function. Ind Qual Control 21:494–498

Hashi K, Ohki S, Matsumoto S, Nishijima G, Goto A, Deguchi K, Yamada K, Noguchi T, Sakai S, Takahashi M, Yanagisawa Y, Iguchi S, Yamazaki T, Maeda H, Tanaka R, Nemoto T, Suematsu H, Miki T, Saito K, Shimizu T (2015) Achievement of 1020MHz NMR. J Magn Reson 256:30–33. https://doi.org/10.1016/j.jmr.2015.04.009

Holley RW (1965) Structure of an alanine transfer ribonucleic acid. Jama 194(8):868–871

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A (1965) Structure of a ribonucleic acid. Science 147(3664):1462–1465

Holliday GL, Brown SD, Akiva E, Mischel D, Hicks MA, Morris JH, Huang CC, Meng EC, Pegg SC, Ferrin TE, Babbitt PC (2017) Biocuration in the structure-function linkage database: the anatomy of a superfamily. Database: J Biol Databases Curation 2017(1). https://doi.org/10.1093/database/bax006

Hollyfield JG, Besharse JC, Rayborn ME (1976) The effect of light on the quantity of phagosomes in the pigment epithelium. Exp Eye Res 23(6):623–635

Holm L, Laakso LM (2016) Dali server update. Nucleic Acids Res 44(W1):W351–W355. https://doi.org/10.1093/nar/gkw357

Holm L, Sander C (1996) The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Res 24(1):206–209

Honorio KM, Moda TL, Andricopulo AD (2013) Pharmacokinetic properties and in silico ADME modeling in drug discovery. Med Chem (Shariqah (United Arab Emirates)) 9(2):163–176

Hospital A, Goñi JR, Orozco M, Gelpí JL (2015) Molecular dynamics simulations: advances and applications. Adv Appl Bioinforma Chem 8:37–47. https://doi.org/10.2147/AABC.S70333

Hung L-H, Ngan S-C, Liu T, Samudrala R (2005) PROTINFO: new algorithms for enhanced protein structure predictions. Nucleic Acids Res 33(Web Server issue):W77–W80. https://doi.org/10.1093/nar/gki403

Huppert JL (2010) Structure, location and interactions of G-quadruplexes. FEBS J 277 (17):3452–3458. https://doi.org/10.1111/j.1742-4658.2010.07758.x

Hynninen AP, Crowley MF (2014) New faster CHARMM molecular dynamics engine. J Comput Chem 35(5):406–413. https://doi.org/10.1002/jcc.23501

Ilari A, Savino C (2008) Protein structure determination by x-ray crystallography. Methods Mol Biol 452:63–87. https://doi.org/10.1007/978-1-60327-159-2_3

Ilari A, Savino C (2017) A Practical Approach to Protein Crystallography. Methods in molecular biology 1525:47–78. https://doi.org/10.1007/978-1-4939-6622-6_3

Jauch R, Yeo HC, Kolatkar PR, Clarke ND (2007) Assessment of CASP7 structure predictions for template free targets. Proteins 69(Suppl 8):57–67

Jayaram B, Dhingra P, Mishra A, Kaushik R, Mukherjee G, Singh A, Shekhar S (2014) Bhageerath-H: a homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. BMC Bioinf 15(Suppl 16):S7–S7. https://doi.org/10.1186/1471-2105-15-S16-S7

Jones DT, Swindells MB (2002) Getting the most from PSI-BLAST. Trends Biochem Sci 27 (3):161–164

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358 (6381):86–89. https://doi.org/10.1038/358086a0

Jorgensen WL, Tiradorives J (1988) The Opls potential functions for proteins – energy minimizations for crystals of cyclic-peptides and crambin. J Am Chem Soc 110 (6):1657–1666. https://doi.org/10.1021/Ja00214a001

Jorgensen WL, Tirado-Rives J (1998) Development of the OPLS-AA force field for organic and biomolecular systems. Abstr Pap Am Chem S 216:U696–U696

Jorgensen WL, Maxwell DS, Tirado Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118 (45):11225–11236. https://doi.org/10.1021/Ja9621760

Kallberg M, Margaryan G, Wang S, Ma J, Xu J (2014) RaptorX server: a resource for template-based protein structure modeling. Methods Mol Biol 1137:17–27. https://doi.org/10.1007/978-1-4939-0366-5_2

Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105(28):6474–6487. https://doi.org/10.1021/jp003919d

Kaus JW, Pierce LT, Walker RC, McCammon JA (2013) Improving the efficiency of free energy calculations in the amber molecular dynamics package. J Chem Theory Comput 9 (9):4131–4139. https://doi.org/10.1021/ct400340s

Kelley LA, MacCallum RM, Sternberg MJE (1999) Recognition of remote protein homologies using three-dimensional information to generate a position specific scoring matrix in the program 3D-PSSM. In: Paper presented at the proceedings of the third annual international conference on computational molecular biology, Lyon, France

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10(6):845–858. https://doi.org/10.1038/nprot.2015.053

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 (4610):662–666

Kim SH, Suddath FL, Quigley GJ, McPherson A, Sussman JL, Wang AH, Seeman NC, Rich A (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. Science 185 (4149):435–440

Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32(Web Server issue):W526–W531. https://doi.org/10.1093/nar/gkh468

Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. Science 220 (4598):671–680. https://doi.org/10.1126/science.220.4598.671

Klepeis JL, Floudas CA (2003) ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. Biophys J 85(4):2119–2146. https://doi.org/10.1016/S0006-3495(03)74640-2

Klepeis JL, Wei Y, Hecht MH, Floudas CA (2005) Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double-blind case study. Proteins 58(3):560–570

Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, Jones TA (2004) The uppsala electron-density server. Acta Crystallogr D Biol Crystallogr 60(Pt 12 Pt 1):2240–2249. https://doi.org/10.1107/s0907444904013253

Knudsen M, Wiuf C (2010) The CATH database. Hum Genomics 4(3):207–212. https://doi.org/10.1186/1479-7364-4-3-207

Kocman V, Plavec J (2017) Tetrahelical structural family adopted by AGCGA-rich regulatory DNA regions. Nat Commun 8:15355. https://doi.org/10.1038/ncomms15355

Kriegel F, Ermann N, Forbes R, Dulin D, Dekker NH, Lipfert J (2017a) Probing the salt dependence of the torsional stiffness of DNA by multiplexed magnetic torque tweezers. Nucleic Acids Res 45(10):5920–5929. https://doi.org/10.1093/nar/gkx280

Kriegel F, Ermann N, Lipfert J (2017b) Probing the mechanical properties, conformational changes, and interactions of nucleic acids with magnetic tweezers. J Struct Biol 197(1):26–36. https://doi.org/10.1016/j.jsb.2016.06.022

Kroese DP, Brereton T, Taimre T, Botev ZI (2014) Why the Monte Carlo method is so important today. Wiley Interdiscip Rev: Comput Stat 6(6):386–392. https://doi.org/10.1002/wics.1314

Kuntal BK, Aparoy P, Reddanna P (2010) EasyModeller: a graphical interface to MODELLER. BMC Res Notes 3:226. https://doi.org/10.1186/1756-0500-3-226

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161(2):269–288

Kuryavyi V, Phan AT, Patel DJ (2010) Solution structures of all parallel-stranded monomeric and dimeric G-quadruplex scaffolds of the human c-kit2 promoter. Nucleic Acids Res 38 (19):6757–6773. https://doi.org/10.1093/nar/gkq558

Lagorce D, Sperandio O, Baell JB, Miteva MA, Villoutreix BO (2015) FAF-Drugs3: a web server for compound property calculation and chemical library design. Nucleic Acids Res 43(W1): W200–W207. https://doi.org/10.1093/nar/gkv353

Lagorce D, Douguet D, Miteva MA, Villoutreix BO (2017) Computational analysis of calculated physicochemical and ADMET properties of protein-protein interaction inhibitors. Sci Rep 7:46277. https://doi.org/10.1038/srep46277

Lambert C, Leonard N, De Bolle X, Depiereux E (2002) ESyPred3D: Prediction of proteins 3D structures. Bioinformatics (Oxford, England) 18(9):1250–1256

Lamiable A, Thevenet P, Rey J, Vavrusa M, Derreumaux P, Tuffery P (2016) PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. Nucleic Acids Res 44(W1):W449–W454. https://doi.org/10.1093/nar/gkw329

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26(2):283–291. https://doi.org/10.1107/S0021889892009944

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8(4):477–486

Lee J (1993) New Monte Carlo algorithm: Entropic sampling. Physical Rev Lett 71(2):211–214. https://doi.org/10.1103/PhysRevLett.71.211

Lee J, Scheraga HA, Rackovsky S (1998) Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. Biopolymers 46(2):103–116. https://doi.org/10.1002/(SICI)1097-0282(199808)46:2<103::AID-BIP5>3.0.CO;2-Q

Leeson P (2012) Drug discovery: chemical beauty contest. Nature 481(7382):455–456

Levitt DG, Banaszak LJ (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10(4):229–234

Li MH, Wang ZF, Kuo MH, Hsu ST, Chang TC (2014) Unfolding kinetics of human telomeric G-quadruplexes studied by NMR spectroscopy. J Phys Chem B 118(4):931–936. https://doi.org/10.1021/jp410034d

Li H, O'Donoghue AJ, van der Linden WA, Xie SC, Yoo E, Foe IT, Tilley L, Craik CS, da Fonseca PC, Bogyo M (2016) Structure- and function-based design of Plasmodium-selective proteasome inhibitors. Nature 530(7589):233–236. https://doi.org/10.1038/nature16936

Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci: Publ Protein Soc 7(9):1884–1897. https://doi.org/10.1002/pro.5560070905

Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today: Technol 1(4):337–341. https://doi.org/10.1016/j.ddtec.2004.11.007

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46(1-3):3–26

Liu Z, Gutierrez-Vargas C, Wei J, Grassucci RA, Sun M, Espina N, Madison-Antenucci S, Tong L, Frank J (2017) Determination of the ribosome structure to a resolution of 2.5 A by single-particle cryo-EM. Protein Sci: Publ Protein Soc 26(1):82–92. https://doi.org/10.1002/pro.3068

Liwo A, Khalili M, Scheraga HA (2005) Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc Natl Acad Sci U S A 102(7):2362–2367

Lo Conte L, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C (2000) SCOP: a structural classification of proteins database. Nucleic Acids Res 28(1):257–259

Lobley A, Sadowski MI, Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics (Oxford, England) 25(14):1761–1767. https://doi.org/10.1093/bioinformatics/btp302

Lohning AE, Levonis SM, Williams-Noonan B, Schweiker SS (2017) a practical guide to molecular docking and homology modelling for medicinal chemists. Curr Top Med Chem 17 (18):2023–2040. https://doi.org/10.2174/1568026617666170130110827

Lovering F, Bikker J, Humblet C (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. J Med Chem 52(21):6752–6756. https://doi.org/10.1021/jm901241e

Lu H, Skolnick J (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins-Struct Funct Genet 44(3):223–232. https://doi.org/10.1002/Prot.1087

MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102 (18):3586–3616

Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J (2004) The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res 32(Database issue):D235–D239. https://doi.org/10.1093/nar/gkh117

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 Throughput Screen 4(6):453–475

McClary B, Zinshteyn B, Meyer M, Jouanneau M, Pellegrino S, Yusupova G, Schuller A, Reyes JCP, Lu J, Guo Z, Ayinde S, Luo C, Dang Y, Romo D, Yusupov M, Green R, Liu JO (2017) Inhibition of eukaryotic translation by the antitumor natural product Agelastatin A. Cell Chem Biol 24(5):605–613 e605. https://doi.org/10.1016/j.chembiol.2017.04.006

Meanwell NA (2011) Improving drug candidates by design: a focus on physicochemical properties as a means of improving compound disposition and safety. Chem Res Toxicol 24 (9):1420–1456. https://doi.org/10.1021/tx200211v

Meanwell NA (2016) Improving drug design: an update on recent applications of efficiency metrics, strategies for replacing problematic elements, and compounds in nontraditional drug space. Chem Res Toxicol 29(4):564–616. https://doi.org/10.1021/acs.chemrestox.6b00043

Millevoi S, Moine H, Vagner S (2012) G-quadruplexes in RNA biology. Wiley Interdiscip Rev RNA 3(4):495–507. https://doi.org/10.1002/wrna.1113

Miteva MA, Villoutreix BO (2017) Computational biology and chemistry in MTi: emphasis on the prediction of some ADMET properties. Mol Inf 36. https://doi.org/10.1002/minf.201700008

Mixon MB, Lee E, Coleman DE, Berghuis AM, Gilman AG, Sprang SR (1995) Tertiary and quaternary structural changes in Gi alpha 1 induced by GTP hydrolysis. Science 270 (5238):954–960

Montgomerie S, Cruz JA, Shrivastava S, Arndt D, Berjanskii M, Wishart DS (2008) PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic Acids Res 36(Web Server issue):W202–W209. https://doi.org/10.1093/nar/gkn255

Murat P, Balasubramanian S (2014) Existence and consequences of G-quadruplex structures in DNA. Curr Opin Genet Dev 25:22–29. https://doi.org/10.1016/j.gde.2013.10.012

Myasnikov AG, Kundhavai Natchiar S, Nebout M, Hazemann I, Imbert V, Khatter H, Peyron JF, Klaholz BP (2016) Structure-function insights reveal the human ribosome as a cancer target for antibiotics. Nat Commun 7:12856. https://doi.org/10.1038/ncomms12856

Nagano N, Nakayama N, Ikeda K, Fukuie M, Yokota K, Doi T, Kato T, Tomii K (2015) EzCatDB: the enzyme reaction database, 2015 update. Nucleic Acids Res 43(Database issue):D453–D458. https://doi.org/10.1093/nar/gku946

Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. J Chem Phys 105(5):1902–1921. https://doi.org/10.1063/1.472061

Newman DJ, Cragg GM (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. J Nat Prod 75(3):311–335. https://doi.org/10.1021/np200906s

Nguyen LA, Wang J, Steitz TA (2017) Crystal structure of Pistol, a class of self-cleaving ribozyme. Proc Natl Acad Sci U S A 114(5):1021–1026. https://doi.org/10.1073/pnas.1611191114

Nicholls A, Sharp KA, Honig B (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 11(4):281–296. https://doi.org/10.1002/prot.340110407

Nugent CI, Lundblad V (1998) The telomerase reverse transcriptase: components and regulation. Genes Dev 12(8):1073–1085

Oefner C, D'Arcy A, Hennig M, Winkler FK, Dale GE (2000) Structure of human neutral endopeptidase (Neprilysin) complexed with phosphoramidon. J Mol Biol 296(2):341–349. https://doi.org/10.1006/jmbi.1999.3492

Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Arnautova YA, Jagielska A, Makowski M, Schafroth HD, Kazmierkiewicz R, Ripoll DR, Pillardy J, Saunders JA, Kang YK, Gibson KD, Scheraga HA (2005) Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proc Natl Acad Sci U S A 102(21):7547–7552

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(5):1308–1315

Orlov I, Myasnikov AG, Andronov L, Natchiar SK, Khatter H, Beinsteiner B, Menetret JF, Hazemann I, Mohideen K, Tazibt K, Tabaroni R, Kratzat H, Djabeur N, Bruxelles T, Raivoniaina F, Pompeo LD, Torchy M, Billas I, Urzhumtsev A, Klaholz BP (2017) The integrative role of cryo electron microscopy in molecular and cellular structural biology. Biol Cell 109(2):81–93. https://doi.org/10.1111/boc.201600042

Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS (2002) Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins 46(1):34–40

Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. Biophys Rev 9 (2):91–102. https://doi.org/10.1007/s12551-016-0247-1

Pandey RB, Jacobs DJ, Farmer BL (2017) Preferential binding effects on protein structure and dynamics revealed by coarse-grained Monte Carlo simulation. J Chem Phys 146(19):195101. https://doi.org/10.1063/1.4983222

Paquet E, Viktor HL (2015) Molecular dynamics, Monte Carlo simulations, and langevin dynamics: a computational review. BioMed Res Int 2015:183918. https://doi.org/10.1155/2015/183918

Parkinson GN, Lee MP, Neidle S (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. Nature 417(6891):876–880. https://doi.org/10.1038/nature755

Patel DJ, Phan AT, Kuryavyi V (2007) Human telomere, oncogenic promoter and 5'-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. Nucleic Acids Res 35(22):7429–7455. https://doi.org/10.1093/nar/gkm711

Patel TR, Chojnowski G, Astha, Koul A, McKenna SA, Bujnicki JM (2017) Structural studies of RNA-protein complexes: a hybrid approach involving hydrodynamics, scattering, and computational methods. Methods (San Diego, Calif) 118:146–162. https://doi.org/10.1016/j.ymeth.2016.12.002

Pauling L, Corey RB (1951) Configuration of polypeptide chains. Nature 168(4274):550–551

Pauling L, Corey RB, Branson HR (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 37(4):205–211

Perrone R, Lavezzo E, Palu G, Richter SN (2017) Conserved presence of G-quadruplex forming sequences in the Long Terminal Repeat Promoter of Lentiviruses. Sci Rep 7(1):2018. https://doi.org/10.1038/s41598-017-02291-1

Piccirilli JA, Koldobskaya Y (2011) Crystal structure of an RNA polymerase ribozyme in complex with an antibody fragment. Philos Trans R Soc Lond Ser B Biol Sci 366(1580):2918–2928. https://doi.org/10.1098/rstb.2011.0144

Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kazmierkiewicz R, Oldziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye YJ, Scheraga HA (2001) Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proc Natl Acad Sci U S A 98(5):2329–2333. https://doi.org/10.1073/pnas.041609598

Porrini M, Rosu F, Rabin C, Darre L, Gomez H, Orozco M, Gabelica V (2017) Compaction of duplex nucleic acids upon native electrospray mass spectrometry. ACS Cent Sci 3(5):454–461. https://doi.org/10.1021/acscentsci.7b00084

Quester S, Schomburg D (2011) EnzymeDetector: an integrated enzyme function prediction tool and database. BMC Bioinf 12:376. https://doi.org/10.1186/1471-2105-12-376

Ramachandran GN (1963) Protein structure and crystallography. Science 141(3577):288–291. https://doi.org/10.1126/science.141.3577.288

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99

Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261(3):470–489. https://doi.org/10.1006/jmbi.1996.0477

Razi A, Britton RA, Ortega J (2017) The impact of recent improvements in cryo-electron microscopy technology on the understanding of bacterial ribosome assembly. Nucleic Acids Res 45 (3):1027–1040. https://doi.org/10.1093/nar/gkw1231

Redfern OC, Harrison A, Dallman T, Pearl FMG, Orengo CA (2007) CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. PLOS Comput Biol 3(11):e232. https://doi.org/10.1371/journal.pcbi.0030232

Redfern OC, Dessailly BH, Dallman TJ, Sillitoe I, Orengo CA (2009) FLORA: a novel method to predict protein function from structure in diverse superfamilies. PLoS Comput Biol 5(8): e1000485. https://doi.org/10.1371/journal.pcbi.1000485

Rhodes D, Giraldo R (1995) Telomere structure and function. Curr Opin Struct Biol 5(3):311–322

Rhodes D, Lipps HJ (2015) G-quadruplexes and their regulatory roles in biology. Nucleic Acids Res 43(18):8627–8637. https://doi.org/10.1093/nar/gkv862

Rich A (1956) Recent studies on the structure of ribonucleic acid. Prog Neurobiol 1:114–121

Rich A (1960) A hybrid helix containing both deoxyribose and ribose polynucleotides and its relation to the transfer of information between the nucleic acids. Proc Natl Acad Sci U S A 46 (8):1044–1053

Rich A, Davies DR, Crick FH, Watson JD (1961) The molecular structure of polyadenylic acid. J Mol Biol 3:71–86

Rietveld K, Van Poelgeest R, Pleij CW, Van Boom JH, Bosch L (1982) The tRNA-like structure at the 3’ terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. Nucleic Acids Res 10(6):1929–1946

Robertus JD, Ladner JE, Finch JT, Rhodes D, Brown RS, Clark BF, Klug A (1974) Structure of yeast phenylalanine tRNA at 3 A resolution. Nature 250(467):546–551

Rodley GA, Scobie RS, Bates RH, Lewitt RM (1976) A possible conformation for double-stranded polynucleotides. Proc Natl Acad Sci U S A 73(9):2959–2963

Roques BP (1985) Enkephalinase inhibitors and molecular study of the differences between active sites of enkephalinase and angiotensin-converting enzyme. J Pharmacol 16(Suppl 1):5–31

Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, Green RK, Goodsell DS, Hudson B, Kalro T, Lowe R, Peisach E, Randle C, Rose AS, Shao C, Tao Y-P, Valasatava Y, Voigt M, Westbrook JD, Woo J, Yang H, Young JY, Zardecki C, Berman HM, Burley SK (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res 45(D1):D271–D281. https://doi.org/10.1093/nar/gkw1000

Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12(2):85–94

Ruiz-Blanco YB, Aguero-Chapin G (2017) Exploring general-purpose protein features for distinguishing enzymes and non-enzymes within the twilight zone. BMC Bioinf 18(1):349. https://doi.org/10.1186/s12859-017-1758-x

Sammito M, Millan C, Rodriguez DD, de Ilarduya IM, Meindl K, De Marino I, Petrillo G, Buey RM, de Pereda JM, Zeth K, Sheldrick GM, Uson I (2013) Exploiting tertiary structure through local folds for crystallographic phasing. Nat Methods 10(11):1099–1101. https://doi.org/10.1038/nmeth.2644

Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 275(5):895–916. https://doi.org/10.1006/jmbi.1997.1479

Samudrala R, Xia Y, Huang E, Levitt M (1999) Ab initio protein structure prediction using a combined hierarchical approach. Proteins Suppl 3:194–198

Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9(1):56–68

Sathyamoorthy B, Shi H, Zhou H, Xue Y, Rangadurai A, Merriman DK, Al-Hashimi HM (2017) Insights into Watson-Crick/Hoogsteen breathing dynamics and damage repair from the solution structure and dynamic ensemble of DNA duplexes containing m1A. Nucleic Acids Res 45 (9):5586–5601. https://doi.org/10.1093/nar/gkx186

Scapin G (2006) Structural biology and drug discovery. Current Pharm Des 12(17):2087–2097

Schlick T, Pyle AM (2017) Opportunities and challenges in RNA structural modeling and design. Biophys J 113(2):225–234. https://doi.org/10.1016/j.bpj.2016.12.037

Sedova A, Banavali NK (2015) RNA approaches the B-form in stacked single strand dinucleotide contexts. Biopolymers. https://doi.org/10.1002/bip.22750

Shekhawat PB, Pokharkar VB (2017) Understanding peroral absorption: regulatory aspects and contemporary approaches to tackling solubility and permeability hurdles. Acta Pharm Sin B 7 (3):260–280. https://doi.org/10.1016/j.apsb.2016.09.005

Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci: Publ Protein Soc 15(11):2507–2524. https://doi.org/10.1110/ps.062416606

Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268(1):209–225

Skiniotis G, Southworth DR (2016) Single-particle cryo-electron microscopy of macromolecular complexes. Microscopy (Oxford, England) 65(1):9–22. https://doi.org/10.1093/jmicro/dfv366

Skolnick J (2006) In quest of an empirical potential for protein structure prediction. Curr Opin Struct Biol 16(2):166–171. https://doi.org/10.1016/j.sbi.2006.02.004

Skolnick J, Jaroszewski L, Kolinski A, Godzik A (1997) Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci: Publ Protein Soc 6(3):676–688. https://doi.org/10.1002/pro.5560060317

Sleator RD, Walsh P (2010) An overview of in silico protein function prediction. Arch Microbiol 192(3):151–155. https://doi.org/10.1007/s00203-010-0549-9

Smith C (2003) Drug target validation: hitting the target. Nature 422(6929): 341, 343, 345 passim. https://doi.org/10.1038/422341a

Sneader W (2000) The discovery of aspirin: a reappraisal. BMJ: Br Med J 321(7276):1591–1594

Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33(Web Server issue):W244–W248. https://doi.org/10.1093/nar/gki408

Stahl K, Schneider M, Brock O (2017) EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction. BMC Bioinf 18:303. https://doi.org/10.1186/s12859-017-1713-x

Stank A, Kokh DB, Horn M, Sizikova E, Neil R, Panecka J, Richter S, Wade RC (2017) TRAPP webserver: predicting protein binding site flexibility and detecting transient binding pockets. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx277

Staple DW, Butcher SE (2005) Pseudoknots: RNA structures with diverse functions. PLoS Biol 3 (6):e213. https://doi.org/10.1371/journal.pbio.0030213

Subramaniam S, Earl LA, Falconieri V, Milne JLS, Egelman EH (2016) Resolution advances in cryo-EM enable application to drug discovery. Curr Opin Struct Biol 41:194–202. https://doi.org/10.1016/j.sbi.2016.07.009

Sugiki T, Kobayashi N, Fujiwara T (2017) Modern technologies of solution nuclear magnetic resonance spectroscopy for three-dimensional structure determination of proteins open avenues for life scientists. Comput Struct Biotechnol J 15:328–339. https://doi.org/10.1016/j.csbj.2017.04.001

Sun LZ, Zhang D, Chen SJ (2017) Theory and modeling of RNA structure and interactions with metal ions and small molecules. Annu Rev Biophys 46:227–246. https://doi.org/10.1146/annurev-biophys-070816-033920

Takahama K, Takada A, Tada S, Shimizu M, Sayama K, Kurokawa R, Oyoshi T (2013) Regulation of telomere length by G-quadruplex telomere DNA- and TERRA-binding protein TLS/FUS. Chem Biol 20(3):341–350. https://doi.org/10.1016/j.chembiol.2013.02.013

Tice CM (2001) Selecting the right compounds for screening: does Lipinski's Rule of 5 for pharmaceuticals apply to agrochemicals? Pest Manag Sci 57(1):3–16. https://doi.org/10.1002/1526-4998(200101)57:1<3::aid-ps269>3.0.co;2-6

Tice CM (2002) Selecting the right compounds for screening: use of surface-area parameters. Pest Manag Sci 58(3):219–233. https://doi.org/10.1002/ps.441

Tilton RF, Dewan JC, Petsko GA (1992) Effects of temperature on protein structure and dynamics: x-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320K. Biochemistry 31(9):2469–2481. https://doi.org/10.1021/bi00124a006

Tinoco I Jr, Bustamante C (1999) How RNA folds. J Mol Biol 293(2):271–281. https://doi.org/10.1006/jmbi.1999.3001

Tosatto SC, Toppo S (2006) Large-scale prediction of protein structure and function from sequence. Curr Pharm Des 12(17):2067–2086

Tripathi A, Kellogg GE (2010) A novel and efficient tool for locating and characterizing protein cavities and binding sites. Proteins 78(4):825–842. https://doi.org/10.1002/prot.22608

Vaguine AA, Richelle J, Wodak SJ (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr D Biol Crystallogr 55(Pt 1):191–205. https://doi.org/10.1107/s0907444998006684

Vallianatou T, Giaginis C, Tsantili-Kakoulidou A (2015) The impact of physicochemical and molecular properties in drug design: navigation in the "drug-like" chemical space. Adv Exp Med Biol 822:187–194. https://doi.org/10.1007/978-3-319-08927-0_21

Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45(12):2615–2623

Venclovas C, Margelevicius M (2005) Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. Proteins 61 (Suppl 7):99–105

Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 21 (4):289–307

Venko K, Roy Choudhury A, Novic M (2017) Computational approaches for revealing the structure of membrane transporters: case study on bilitranslocase. Comput Struct Biotechnol J 15:232–242. https://doi.org/10.1016/j.csbj.2017.01.008

Villoutreix BO (2016) Combining bioinformatics, chemoinformatics and experimental approaches to design chemical probes: applications in the field of blood coagulation. Ann Pharm Fr 74 (4):253–266. https://doi.org/10.1016/j.pharma.2016.03.006

Wahl MC, Sundaralingam M (1997) Crystal structures of A-DNA duplexes. Biopolymers 44 (1):45–63. https://doi.org/10.1002/(sici)1097-0282(1997)44:1<45::aid-bip4>3.0.co;2-#

Wan W, Briggs JA (2016) Cryo-electron tomography and subtomogram averaging. Methods Enzymol 579:329–367. https://doi.org/10.1016/bs.mie.2016.04.014

Wang G, Vasquez KM (2007) Z-DNA, an active element in the genome. Front Biosci 12:4424–4438

Wang AH, Quigley GJ, Kolpak FJ, Crawford JL, van Boom JH, van der Marel G, Rich A (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. Nature 282(5740):680–686

Wang Z, Yin P, Lee JS, Parasuram R, Somarowthu S, Ondrechen MJ (2013) Protein function annotation with Structurally Aligned Local Sites of Activity (SALSAs). BMC Bioinf 14(Suppl 3):S13–S13. https://doi.org/10.1186/1471-2105-14-S3-S13

Wang C, Zhang H, Zheng W-M, Xu D, Zhu J, Wang B, Ning K, Sun S, Li SC, Bu D (2016) FALCON@home: a high-throughput protein structure prediction server based on remote homologue recognition. Bioinformatics (Oxford, England) 32(3):462–464. https://doi.org/10.1093/bioinformatics/btv581

Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol 13(1):e1005324. https://doi.org/10.1371/journal.pcbi.1005324

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171(4356):737–738

Webb B, Sali A (2016) Comparative protein structure modeling using MODELLER. Current protocols in bioinformatics/editorial board, Andreas D Baxevanis [et al] 54:5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3

Weichenberger CX, Sippl MJ (2007) NQ-Flipper: recognition and correction of erroneous asparagine and glutamine side-chain rotamers in protein structures. Nucleic Acids Res 35(Web Server issue):W403–W406. https://doi.org/10.1093/nar/gkm263

Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P (1984) A new force-field for molecular mechanical simulation of nucleic-acids and proteins. J Am Chem Soc 106(3):765–784. https://doi.org/10.1021/Ja00315a051

Weisser M, Schafer T, Leibundgut M, Bohringer D, Aylett CHS, Ban N (2017) Structural and functional insights into human re-initiation complexes. Mol Cell 67(3):447–456.e447. https://doi.org/10.1016/j.molcel.2017.06.032

Weldon C, Eperon IC, Dominguez C (2016) Do we know whether potential G-quadruplexes actually form in long functional RNA molecules? Biochem Soc Trans 44(6):1761–1768. https://doi.org/10.1042/bst20160109

Weldon C, Behm-Ansmant I, Hurley LH, Burley GA, Branlant C, Eperon IC, Dominguez C (2017) Identification of G-quadruplexes in long functional RNAs using 7-deazaguanine RNA. Nat Chem Biol 13(1):18–20. https://doi.org/10.1038/nchembio.2228

Westbrook JD, Hall RS (1995) DDL. A dictionary description language for structure macromolecular, V. 2.1.1. Rutgers University NDB-110, New Brunswick

Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36(3):307–340

Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35(Web Server issue):W407–W410. https://doi.org/10.1093/nar/gkm290

Wilkins MH, Stokes AR, Wilson HR (1953) Molecular structure of deoxypentose nucleic acids. Nature 171(4356):738–740

Wing R, Drew H, Takano T, Broka C, Tanaka S, Itakura K, Dickerson RE (1980) Crystal structure analysis of a complete turn of B-DNA. Nature 287(5784):755–758

Wright WE, Tesmer VM, Huffman KE, Levene SD, Shay JW (1997) Normal human chromosomes have long G-rich telomeric overhangs at one end. Genes Dev 11(21):2801–2809

Wu S, Zhang Y (2008) MUSTER: Improving protein sequence profile–profile alignments by using multiple sources of structure information. Proteins 72(2):547–556. https://doi.org/10.1002/prot.21945

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(5):1177–1187

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. https://doi.org/10.1002/prot.24065

Yang J, Zhang Y (2015) I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Res 43(W1):W174–W181. https://doi.org/10.1093/nar/gkv342

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 D Biol Crystallogr 60(Pt 10):1833–1839. https://doi.org/10.1107/s0907444904019419

Yang Y, Faraggi E, Zhao H, Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics (Oxford, England) 27(15):2076–2082. https://doi.org/10.1093/bioinformatics/btr350

Yella VR, Bansal M (2017) DNA structural features of eukaryotic TATA-containing and TATA-less promoters. FEBS Open Bio 7(3):324–334. https://doi.org/10.1002/2211-5463.12166

Zamenhof S, Brawerman G, Chargaff E (1952) On the desoxypentose nucleic acids from several microorganisms. Biochim Biophys Acta 9(4):402–405

Zeng B, Wang H, Zou L, Zhang A, Yang X, Guan Z (2010) Evaluation and target validation of indole derivatives as inhibitors of the AcrAB-TolC efflux pump. Biosci Biotechnol Biochem 74 (11):2237–2241. https://doi.org/10.1271/bbb.100433

Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101(20):7594–7599. https://doi.org/10.1073/pnas.0305695101

Zhang Y, Skolnick J (2005) The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 102(4):1029–1034. https://doi.org/10.1073/pnas.0407152101

Zhang Y, Kolinski A, Skolnick J (2003) TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 85(2):1145–1164. https://doi.org/10.1016/S0006-3495(03)74551-2

Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci U S A 103 (8):2605–2610. https://doi.org/10.1073/pnas.0509379103

Zhao C, Pyle AM (2017) Structural insights into the mechanism of group II intron splicing. Trends Biochem Sci 42(6):470–482. https://doi.org/10.1016/j.tibs.2017.03.007

Zheng H, Cooper DR, Porebski PJ, Shabalin IG, Handing KB, Minor W (2017) CheckMyMetal: a macromolecular metal-binding validation tool. Acta Crystallogr D Struct Biol 73 (Pt 3):223–233. https://doi.org/10.1107/S2059798317001061

Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci: Publ Protein Soc 11(11):2714–2726. https://doi.org/10.1110/ps.0217002

Zhou H, Hintze BJ, Kimsey IJ, Sathyamoorthy B, Yang S, Richardson JS, Al-Hashimi HM (2015) New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. Nucleic Acids Res 43(7):3420–3433. https://doi.org/10.1093/nar/gkv241

Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. Proc Natl Acad Sci U S A 89(1):20–22

A Survey of the Structural Parameters Used for Computational Prediction of Protein Folding Process

11

Gulshan Khalique and Tambi Richa

11.1 Introduction

Proteins are one of the most important biomacromolecules, made up of a linear chain of amino acids held together by peptide bonds, having a unique 3D structure which gives them a distinguishing function (Campbell et al. 2008; Pace et al. 2004; Nelson et al. 2005). The function of a protein depends on its structure, such that any change at the structural level will be directly reflected in its biological activity. Structure of proteins can be ordered into four levels: primary, secondary, tertiary and quaternary structure. The complexity of the structure imparted through the covalent and non-covalent interactions increases with the level. Each structural level is outlined below.

11.1.1 Primary Structure

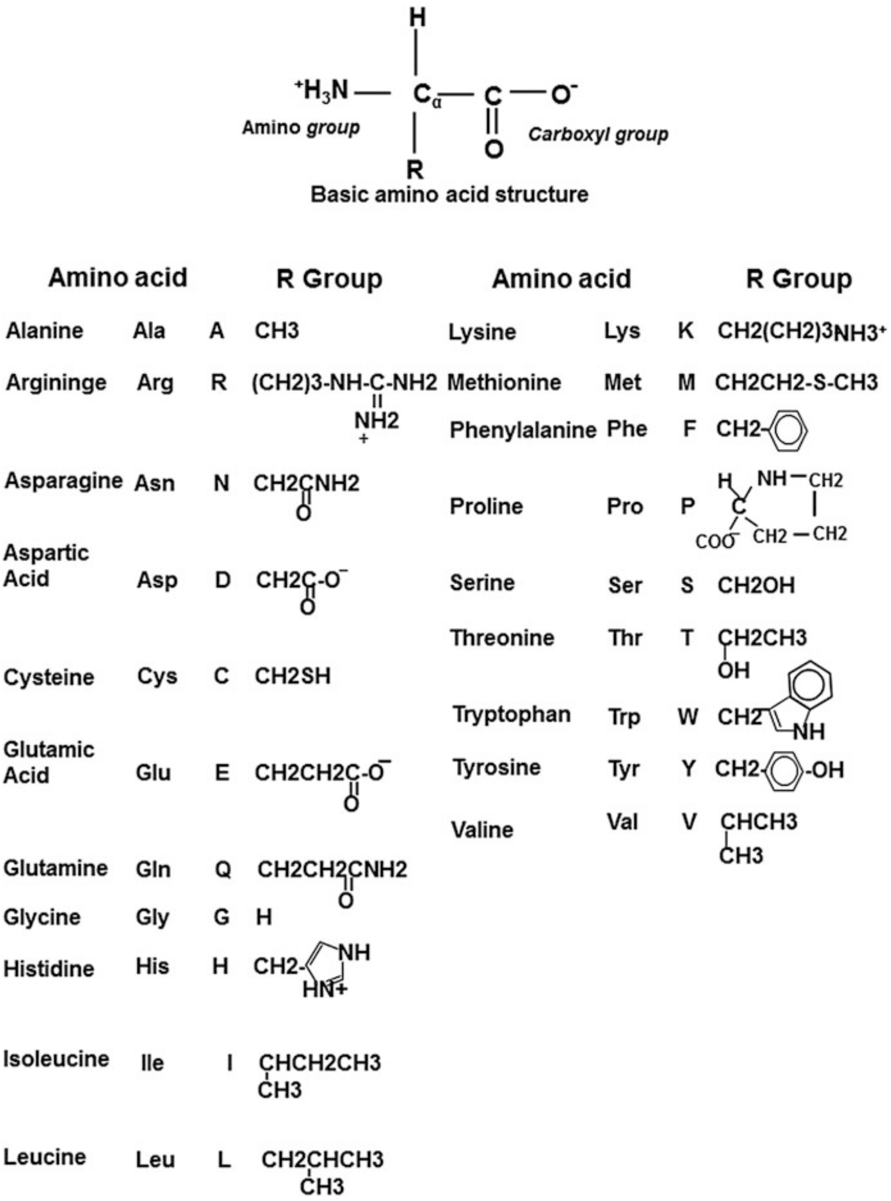
The linear sequence of amino acid connected by peptide bonds is the primary structure of a protein. There are 20 different L-α amino acids which commonly form the primary structure. These amino acids differ from each other in their side chain (known as the R-group), while all of them have the α-carbon atom to which a carboxylic (-COOH) group and an amino (-NH2) group are attached (Fig. 11.1). They are represented by a three-letter code as well as a one-letter code. Based on their R-group, they can be classified into hydrophobic, hydrophilic, acidic and basic

G. Khalique
Jain University, Bangalore, India

T. Richa (✉)
Tokyo University of Agriculture and Technology, Tokyo, Japan

Current Address: Banerjee Lab, Mohammed Bin Rashid University, Dubai, UAE

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_11



Amino acid			R Group	Amino acid			R Group
Alanine	Ala	A	CH3	Lysine	Lys	K	CH2(CH2)3NH3+
Argininge	Arg	R	(CH2)3-NH-C-NH2 NH2 +	Methionine	Met	M	CH2CH2-S-CH3
				Phenylalanine	Phe	F	CH2-
Asparagine	Asn	N	CH2CNH2 O	Proline	Pro	P	H NH–CH2 C COO– CH2–CH2
Aspartic Acid	Asp	D	CH2C-O– O	Serine	Ser	S	CH2OH
				Threonine	Thr	T	CH2CH3 OH
Cysteine	Cys	C	CH2SH	Tryptophan	Trp	W	CH2 NH
Glutamic Acid	Glu	E	CH2CH2C-O– O	Tyrosine	Tyr	Y	CH2- -OH
				Valine	Val	V	CHCH3 CH3
Glutamine	Gln	Q	CH2CH2CNH2 O				
Glycine	Gly	G	H				
Histidine	His	H	CH2- NH HN+				
Isoleucine	Ile	I	CHCH2CH3 CH3				
Leucine	Leu	L	CH2CHCH3 CH3				

Fig. 11.1 Structure of amino acids. Amino acids are represented by their three-letter and one-letter codes. Ala, Val, Leu, Ile, Pro, Phe, Trp and Met are hydrophobic amino acids. Gly, Ser, Thr, Cys, Tyr, Asn and Gln are hydrophilic uncharged amino acids. Acidic amino acids are Asp and Glu. Lys, Arg and His are the basic amino acids

amino acid. Apart from these 20 amino acids, certain uncommon amino acids such as hydroxyproline, thyroxine and selenocysteine are also found in some proteins. Moreover, a few non-proteinogenic amino acids are also present in cells performing specific functions such as histamines involved in allergic reactions, ornithine and citrulline as a part of the urea cycle, serotonin as a neurotransmitter and so on. Two proteinogenic amino acids form a peptide bond through a condensation reaction by removal of water from a carboxylic group of one amino acid and the amino group of another.

A polypeptide chain consists of an N/amino-terminal (end with the free amino group) and a C/carboxyl-terminal (end with free carboxyl group). These two free ends, as well as the ionisable R-groups of the residues constituting a polypeptide chain, are responsible for its acid-base behaviour. Some proteins consist of a single polypeptide chain, while a few others comprise of multiple identical or dissimilar polypeptide chains. For example, haemoglobin found in red blood cells is a tetrameric protein which has two identical α-chains made up of 141 amino acid residues and two identical β-chains made up of 146 amino acid residues, whereas other proteins such as ribonuclease A (124 residues) and lysozyme (129 residues) are made up of the single polypeptide chain. Another important detail regarding proteins is that many of them (known as conjugated proteins) are often associated with chemical groups other than amino acids (prosthetic group), for example, haem-containing proteins such as haemoglobin, myoglobin and cytochrome c and glycoproteins like ovalbumin, transferrin and mucin which contain different levels of carbohydrates, etc. The primary structure of a protein determines its sequence and lays the foundation for the next level in the structural hierarchy.

11.1.2 Secondary Structure

Secondary structure is the local substructure of proteins in which backbone hydrogen bonds play an important role. Secondary structures can be classified predominantly as regular and irregular. The regular secondary structure includes α-helices and β-sheets, whereas coils and turns constitute the most common type of irregular secondary structures. As their name suggests, helices and sheet have a regular pattern of hydrogen bonding which is missing in the irregular structures (Chan and Dill 1990; Lim 1974).

The commonly occurring α-helix is a right-handed coiled structure (Fig. 11.2). The side-chain substituent of the amino acid groups in an α-helix extends outside. A hydrogen bond is formed between the oxygen molecule of the backbone C=O group of the *i*th residue and the hydrogen molecule of the N-H group of the i+4th residue. It consists of 3.6 amino acid residues per turn and 1.5 Å rise per residue. It adopts a phi (φ) angle (torsion angle involving a carbonyl carbon, the connecting α-carbon, an amide nitrogen and the next carbonyl carbon defining rotation of polypeptide backbone around N-Cα bond) of −60° and psi (ψ) angle (torsion angle involving an amide nitrogen, a carbonyl carbon, an α-carbon and a second nitrogen defining rotation of polypeptide backbone around Cα-C bond) of −45° to −50°. A plot of ψ versus φ which explains the allowable regions and disfavoured regions (due to steric clashes) is known as the Ramachandran plot (Fig. 11.3), developed by

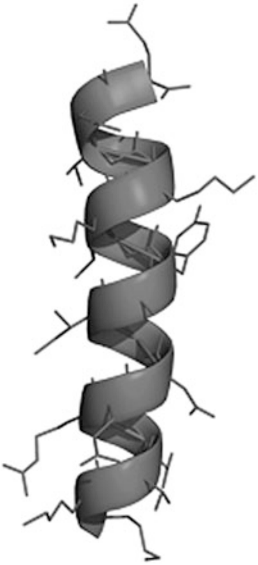


Fig. 11.2 α-Helical structure of the protein. The side chains are represented by stick model superimposed on the cartoon representation of helix, rendered using Pymol (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC). The side chain in α-helix extends outward

G.N. Ramachandran in 1963. Right-handed alpha helices are usually occupying the third quadrant of the Ramachandran plot. Amino acids have a varying propensity towards different secondary structures. Met, Ala, Leu, Glu and Lys favour α-helix, whereas Pro and Gly are helix breakers.

Another commonly occurring regular secondary structure is a β-sheet (Fig. 11.4). In β-sheet hydrogen bonds are found between rather than within the strands, and side chain mostly occupies a position above and below the sheet. The two strands can be either parallel or antiparallel depending on the directions of the strand (N-terminus to C-terminus). Parallel β-sheet adopts φ angle of −120° and ψ of −115°, and antiparallel adopts φ and ψ of −140° and − 135°, respectively. The sterically allowed conformations of β-strand occupy the second quadrant of the Ramachandran plot (Fig. 11.3)

Irregular secondary structures such as turn, loops and bends are mostly found joining the regular secondary structures. They are more flexible, and residues occupying irregular structures are usually exposed in comparison to the regular secondary structure residues. They confer a compact structure to the protein by connecting the regular secondary structure. Usually, they are made up of 2–16 residues and assist in the interaction of proteins with other biomolecules. β-Turns are the third most abundant secondary structure after helices and sheets and aid in reversing the direction of a polypeptide chain.

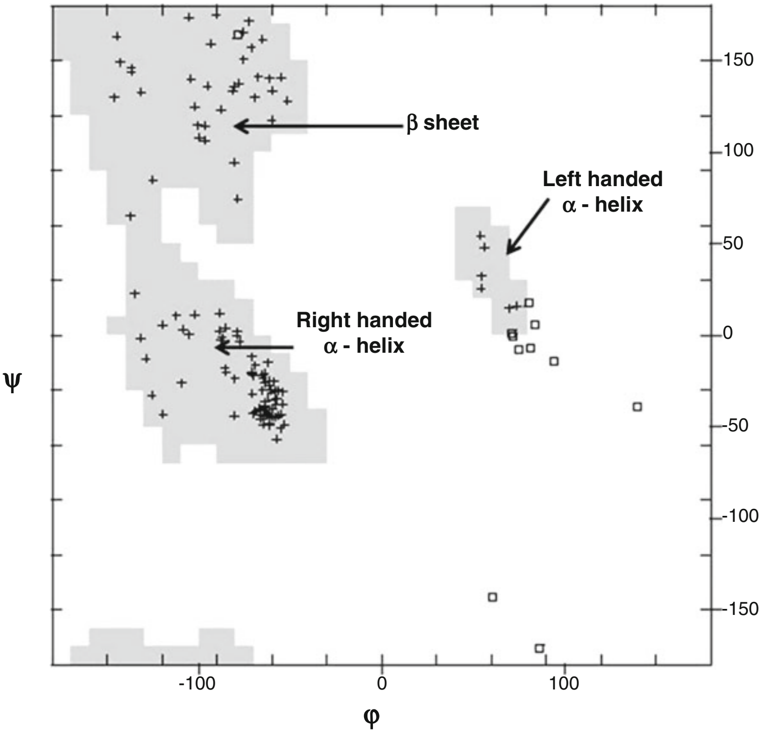
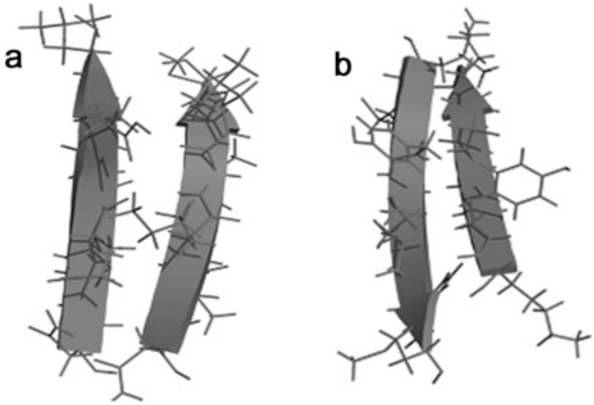


Fig. 11.3 Ramachandran plot for hen egg-white lysozyme (PDB ID: 5O6Q) generated by the Ramachandran server (http://eds.bmc.uu.se/ramachan.html - Kleywegt and Jones 1996). This two-dimensional ϕ,ψ-plot represents the energetically allowed region for this protein

Fig. 11.4 **β-Sheet:** (**a**) parallel and (**b**) antiparallel. The side chains are represented by stick model superimposed on the cartoon representation of sheets, rendered using Pymol (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC). β-Sheets are more extended than α-helices



11.1.3 Tertiary Structure

The third level of protein structural hierarchy is the tertiary structure which is the global 3D structure of proteins. Once secondary structures are formed, they spatially arrange themselves into domains which can evolve and function independently from the rest of the protein. Smaller proteins are usually made up of a single domain (such as cytochrome c and myoglobin), whereas large complex proteins (e.g. protein kinases) often consist of multiple domains. The side-chain bonding interactions which facilitate the tertiary structure of the protein are disulphide bonds, hydrogen bonds, salt bridges and hydrophobic interactions. Disulphide bond also known as S-S bond is a covalent bond formed between two cysteine (containing thiol –SH group). It may be formed within the same chain or between two polypeptide chains in a protein. These bonds provide stability to the protein structure by keeping the structure intact. Hydrogen bond is the interaction between two electronegative atoms through hydrogen bound covalently to one of the atom. Side chain of Ser, Thr and Tyr, or Asp and Tyr, or Asp and Glu or Ser and Lys or that of Ser and Asn may be hydrogen bonded. Salt bridge is the ionic bond between a positively and negatively charged side chains. The most important stabilizing interaction is the non-polar hydrophobic interaction. The hydrophobic residues will be protected from the aqueous medium and remain buried forming the core of protein structure. On the contrary, the hydrophilic residues will be usually present on the surface of proteins.

The precise demonstration of tertiary structure of a protein is a tedious task as it takes a very long time. Based on the information of primary and secondary structure of a protein, various softwares are used to predict its tertiary structure.

11.1.4 Quaternary Structure

All proteins share three levels of structural hierarchy and a few of them also entail a quaternary structure. Proteins are often made up of several polypeptide chains, which are termed as protein subunits which may be the same (as in a homodimer) or different (as in a heterodimer). Interaction of these protein subunits with each other in order to organize them to create a larger protein complex is termed as the quaternary structure. Some proteins with quaternary structure are dimeric creatine kinase, tetrameric haemoglobin and octomeric tryptophanase. The same kind of covalent as well as non-covalent interactions which facilitate the tertiary structure also stabilizes the quaternary structure of a protein.

11.2 Folding Process of Proteins

The ‘protein folding problem’ is defined as the quest to understand the mechanism by which a protein spontaneously adapts its native structure from its primary sequence within the biologically relevant timescale (Creighton 1995; Dill and MacCallum 2012; Dill et al. 2008; Ivarsson et al. 2008; Rose et al. 2006). The protein folding problem has intrigued the researchers since decades (Dill and MacCallum 2012; Dill et al. 2007; Gianni and Jemth 2016; Ivarsson et al. 2008). The in vivo protein folding process is often facilitated by molecular chaperones (Balchin et al. 2016; Mogk et al. 2002). Molecular chaperones are protein molecules which help other proteins to attain their error-free biologically active conformation. A decrease in these ubiquitous protein molecules in the cell is often associated with an increase in typically folded proteins (Chaudhuri and Paul 2006; Welch 2004; Hartl et al. 2011). Errors in the protein folding process vitiate the biological function of proteins. This leads to an array of diseases either due to accumulation of toxic aggregates constituted by the misfolded proteins or due to the inactivity of the protein at their site of action (Balchin et al. 2016; Betts and King 1999; Bross et al. 1999; Chaudhuri and Paul 2006; Chi and Liberles 2016; Cohen 1999; Dobson 1999; Ivarsson et al. 2008; Jaenicke 1995; Ogen-Shtern et al. 2016). Alzheimer's disease, Parkinson's disease, Huntington's disease, prion disease, type II diabetes, amyloidosis, Creutzfeldt-Jakob disease, cataracts and cystic fibrosis are some of the lethal diseases caused by protein misfolding (Walker et al. 2006; Fadiel et al. 2007; Harrison et al. 2007; Eftekharzadeh et al. 2016; Santucci et al. 2008; Winklhofer et al. 2008; Luheshi et al. 2008). Resolving the protein folding problem will help us in finding a possible cure to these fatal diseases. Once we are aware of the rules of protein folding process, structure prediction from the sequence will be much easier. Moreover, the rules dictating the folding-unfolding mechanism will also help in interpreting 'the inverse folding problem', allowing us to navigate through the structure to sequence (Khoury et al. 2014; Godzik et al. 1993; Park et al. 2004). Therefore, unravelling the protein folding puzzle will help in stimulating research in the field of protein misfolding/aggregation as well as protein designing.

In the early 1970s, Christian Anfinsen showed that proteins can fold reversibly and the knowledge about the mechanism by which a linear polypeptide chain acquires its native, biologically active structure is stored in its primary structure (Anfinsen 1973). The protein should fold back from its unfolded state to its native conformation within a reasonable time frame. In principle, a polypeptide chain can adopt numerous conformations. Hence, if the protein needs to sample out all the possible conformation to decide on its global minimum, then for a protein consisting of 100 amino acids, it would take more than the age of universe (about 1066 years, if one assumes that only 10^{-13} s is required to convert from one conformation to other) to fold into its native conformation (Tompa and Rose 2011; Karplus 1997; Dill and Chan 1997). However, the protein folds within seconds, in general. This puzzle is popularly called as the 'Levinthal paradox'.

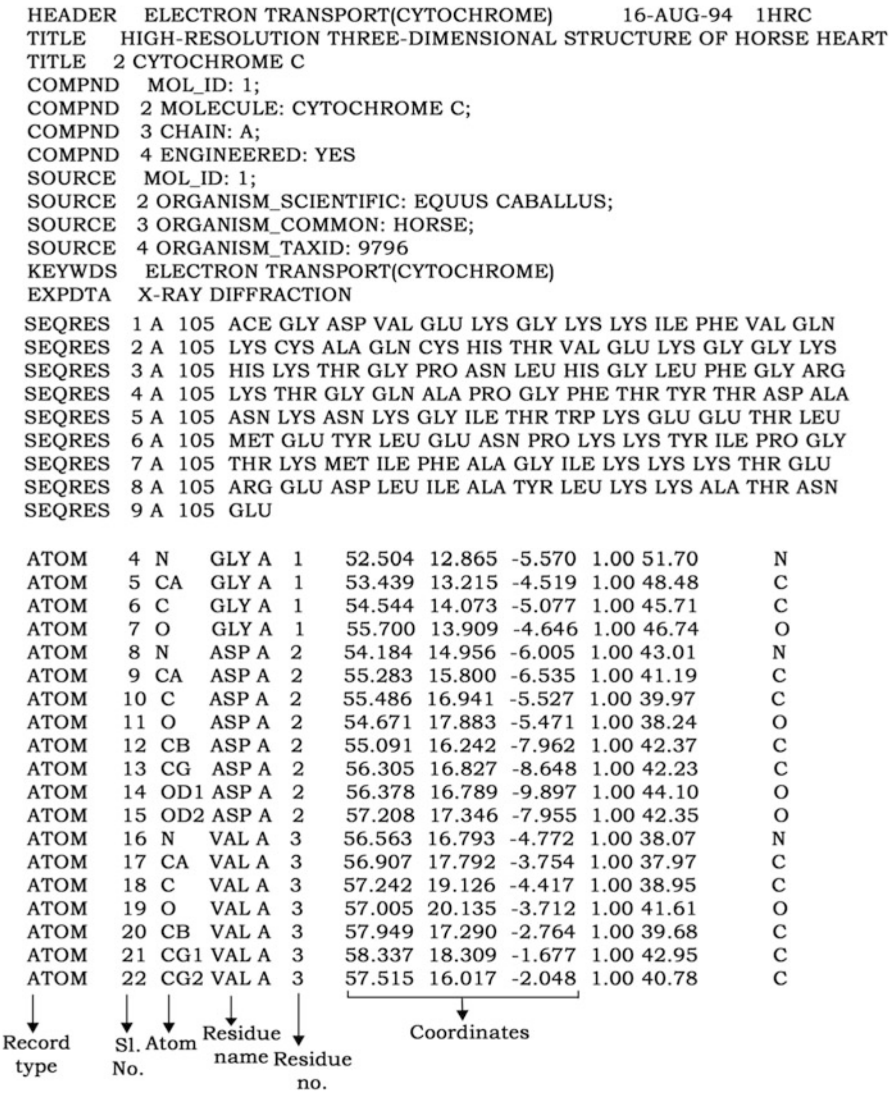
The different aspects of the protein folding mechanism are studied using a variety of experimental techniques such as nuclear magnetic resonance spectroscopy (NMR), circular dichroism (CD), fluorescence spectroscopy, dual polarization interferometry and mass spectrometry (Sikder and Zomaya 2005; Miles and Wallace 2016; Kuwajima and Schmid 1984; Nguyen et al. 1995; Sheu et al. 2010; Santucci et al. 2008). The mechanism is also assessed utilizing the statistical methods and molecular dynamics approaches (Gipson et al. 2012; Carugo and Pongor 2002; Lazaridis and Karplus 2003; Richa and Sivaraman 2012).

One of the most important facets of the protein folding problem being extensively studied using both experimental and theoretical methods is 'analysis of the protein folding rate'. Experimentally, it is equally time-consuming and expensive; therefore more and more computational algorithms have been developed (Richa and Sivaraman 2014; Chaudhary et al. 2015; Schafer et al. 2012; Gromiha and Huang 2011). These algorithms are based on the principle that the native state topology of the small two-state folding (folding by 'all or none' mechanism without the accumulation of intermediate states) proteins is directly linked to their folding rates (Gromiha 2003; Riddle and Grantcharova 1999). Usually, these theoretical methods test their accuracy using the data generated from the experimental methods. One of the most important requirements for computational investigation of proteins is their 3D structure. Protein structures are usually solved experimentally and deposited in the Protein Data Bank (PDB), which forms the basis of the in silico protein research, explained concisely in the next section.

11.3 The Protein Data Bank

The atomic coordinates of many of the macromolecules have been experimentally determined and deposited in the *Protein Data Bank* (www.rcsb.org). PDB was formed in 1971 at the Brookhaven National Laboratories (BNL) and initially consisted of only the following structures: carboxypeptidase, chymotrypsin, cytochrome b5, haemoglobin, lactate dehydrogenase, myoglobin, rubredoxin, subtilisin and trypsin inhibitor. In order to make PDB global and uniform, wwPDB was created in 2003 by Research Collaboratory for Structural Bioinformatics (RCSB) PDB in the USA, PDB in Europe (http://pdbe.org) and PDB in Japan (http://pdbj.org). In 2006 BioMagResBank (http://bmrb.wisc.edu) also joined wwPDB (Berman et al. 2012; Berman 2008). Currently (Feb 2018), the PDB contains 137,322 biological macromolecular structures which involve 127,490 distinct protein sequences. X-ray crystallography, NMR spectroscopy and electron microscopy are widely used for solving the biomolecular structures. PDB is weekly updated and is freely available to the public domain. In PDB each biomolecule is denoted by a four-letter alphanumeric code, and they have a uniform format as shown in Fig. 11.5. The PDB format consists of a line of information known as records described in a text file. Some of the important record types are SEQRES, polypeptide sequence represented using three-letter coding of amino acids; ATOM, atomic coordinate record for the standard residues; HETATM, non-standard residue's atomic coordinate records; TER, end of a chain; HELIX, position of helices; SHEET, position and type of sheet; and SSBOND, position of disulphide bonds formed by cysteine residues. The structure of proteins deposited in PDB is the primary source for starting most of the in silico analysis of proteins.

The alignment of the regular secondary structure of proteins, also known as protein topology, plays an important role in understanding the relationship between protein structure and its folding mechanism (Dokholyan et al. 2002; Baker 2000; Martin 2000). Protein topology can be directly inferred in terms of number and type



```
HEADER    ELECTRON TRANSPORT(CYTOCHROME)        16-AUG-94   1HRC
TITLE    HIGH-RESOLUTION THREE-DIMENSIONAL STRUCTURE OF HORSE HEART
TITLE    2 CYTOCHROME C
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CYTOCHROME C;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: EQUUS CABALLUS;
SOURCE   3 ORGANISM_COMMON: HORSE;
SOURCE   4 ORGANISM_TAXID: 9796
KEYWDS    ELECTRON TRANSPORT(CYTOCHROME)
EXPDTA   X-RAY DIFFRACTION
SEQRES   1 A  105  ACE GLY ASP VAL GLU LYS GLY LYS LYS ILE PHE VAL GLN
SEQRES   2 A  105  LYS CYS ALA GLN CYS HIS THR VAL GLU LYS GLY GLY LYS
SEQRES   3 A  105  HIS LYS THR GLY PRO ASN LEU HIS GLY LEU PHE GLY ARG
SEQRES   4 A  105  LYS THR GLY GLN ALA PRO GLY PHE THR TYR THR ASP ALA
SEQRES   5 A  105  ASN LYS ASN LYS GLY ILE THR TRP LYS GLU GLU THR LEU
SEQRES   6 A  105  MET GLU TYR LEU GLU ASN PRO LYS LYS TYR ILE PRO GLY
SEQRES   7 A  105  THR LYS MET ILE PHE ALA GLY ILE LYS LYS LYS THR GLU
SEQRES   8 A  105  ARG GLU ASP LEU ILE ALA TYR LEU LYS LYS ALA THR ASN
SEQRES   9 A  105  GLU

ATOM      4  N    GLY A  1    52.504  12.865  -5.570  1.00 51.70        N
ATOM      5  CA   GLY A  1    53.439  13.215  -4.519  1.00 48.48        C
ATOM      6  C    GLY A  1    54.544  14.073  -5.077  1.00 45.71        C
ATOM      7  O    GLY A  1    55.700  13.909  -4.646  1.00 46.74        O
ATOM      8  N    ASP A  2    54.184  14.956  -6.005  1.00 43.01        N
ATOM      9  CA   ASP A  2    55.283  15.800  -6.535  1.00 41.19        C
ATOM     10  C    ASP A  2    55.486  16.941  -5.527  1.00 39.97        C
ATOM     11  O    ASP A  2    54.671  17.883  -5.471  1.00 38.24        O
ATOM     12  CB   ASP A  2    55.091  16.242  -7.962  1.00 42.37        C
ATOM     13  CG   ASP A  2    56.305  16.827  -8.648  1.00 42.23        C
ATOM     14  OD1  ASP A  2    56.378  16.789  -9.897  1.00 44.10        O
ATOM     15  OD2  ASP A  2    57.208  17.346  -7.955  1.00 42.35        O
ATOM     16  N    VAL A  3    56.563  16.793  -4.772  1.00 38.07        N
ATOM     17  CA   VAL A  3    56.907  17.792  -3.754  1.00 37.97        C
ATOM     18  C    VAL A  3    57.242  19.126  -4.417  1.00 38.95        C
ATOM     19  O    VAL A  3    57.005  20.135  -3.712  1.00 41.61        O
ATOM     20  CB   VAL A  3    57.949  17.290  -2.764  1.00 39.68        C
ATOM     21  CG1  VAL A  3    58.337  18.309  -1.677  1.00 42.95        C
ATOM     22  CG2  VAL A  3    57.515  16.017  -2.048  1.00 40.78        C
```

Fig. 11.5 PDB format of cytochrome c (PDBID: 1HRC). SEQRES field contains the protein sequence, and 3D coordinates are recorded under ATOM field

of contacts formed by its residues. Protein with simple native topology will have a number of local contacts, and complex protein will have a higher number of long-range contacts. A protein with higher local contacts usually folds faster when compared to the one with non-local contacts. Based on these observations, many structural descriptors have been formulated in the past years (Plaxco et al. 1998; Fersht 2000; Ivankov et al. 2003; Gromiha and Selvaraj 2001; Nolting et al. 2003).

In the following section, we discuss few of the most important structural descriptors which have been widely used to predict the protein folding rates.

11.4 Structural Descriptors for Computational Prediction of Protein Folding Rate

11.4.1 Contact Order

Contact order is a topological descriptor of protein proposed by Plaxco et al. (1998). It was then correlated to the logarithm of folding rate for 12 two-state folding proteins. A correlation coefficient of 0.81 was observed. The test set proteins were 28–70% homologous and shared similar topologies. 'Relative contact order' as it was typically called by Plaxco et al. (1998) is calculated using the following formula:

$$\mathrm{CO} = \frac{1}{L \bullet N} \sum^{N} \Delta S_{ij} \tag{11.1}$$

where L is the length of proteins (i.e. the total number of amino acid residues constituting the protein), N is the total number of contacts and S_{ij} is the sequence separation between the ith and jth residue, which are in contact. Two residues were considered to be in contact if the interatomic distance between any two non-hydrogen atoms was within 6 Å. Therefore, contact order is basically the average sequence separation of protein normalized over the protein length. It has been shown that slow-folding proteins, which usually involve more of non-local interactions, have larger contact order, whereas comparatively fast-folding proteins have a number of local networks and a smaller value for a contact order (Fersht 2000). One more modified form of contact order is absolute contact order computed using the following equation:

$$\mathrm{Abs\_CO} = \mathrm{CO} * L \tag{11.2}$$

However, absolute contact order showed weaker correlation with respect to folding rates of two-state proteins (Grantcharova et al. 2001; Ivankov et al. 2003). A direct correlation between contact order and protein folding rate indicates that native interaction of protein plays a dominant role in kneading the folding mechanism of proteins. Contact order of a protein can be easily calculated by uploading the structural coordinate file of protein at http://www.bakerlab.org/contact_order/.

11.4.2 Long-Range Order

Long-range order (Gromiha and Selvaraj 2001) is a statistical measure of the number of long-range contacts constituted by a protein structure under study. Long-range contacts are contacts which are far off in the polypeptide sequence and imminent in space. Long-range order often abbreviated as 'LRO' is computed using the following formula:

$$\mathrm{LRO} = \sum N_{ij}/L\, N_{ij} \begin{matrix} = 1 & \text{if } \mid i - j \mid > 12 \\ = 0 & \text{otherwise} \end{matrix} \tag{11.3}$$

where '*i*' and '*j*' are the residues for which contact is being computed. Two residues are considered to be in long-range contact when the distance between their C_α atoms is within 8 Å, and they are at least separated by 12 residues. '*L*' is the length of the polypeptide chain. As per the formula, LRO is the average number of long-range contacts for each residue of a protein structure. LRO and folding rates of 23 small two-state proteins were correlated, and an inverse correlation of −0.78 was observed. This data set consisted of four all-α proteins, ten all-β proteins and nine mixed-class proteins. LRO also had a strong correlation (−0.82) with CO of mixed-class proteins and weaker correlation of −0.56 and − 0.46 for all-α and all-β proteins, respectively. This shows that the structural classification of protein predominantly affects its folding rates. Grohima's group has also developed a web server, Fold-Rate (Gromiha et al. 2006), which predicts the folding rates of proteins from its sequence using the amino acid property of each structure.

11.4.3 Total Contact Distance

Total contact distance (TCD; Zhou and Zhou 2002) combines the concept of both contact order and long-range order. The effect of TCD on folding rates was measured using a database which contained experimental data for 28 proteins belonging to all three classes (Dinner and Karplus 2001). It consisted of 4 all-α proteins, 13 all-β proteins and 11 mixed proteins. TCD can be enumerated using Eq. 11.4 explained below:

$$\mathrm{TCD} = \frac{1}{\mathrm{L}^2} \sum \mathrm{S}_{ij} \tag{11.4}$$

where two residues '*i*' and '*j*' were considered to be in contact when a heavy atom of these residues was within 6 Å and they were separated by $\mathrm{l}_{\mathrm{cut}}$ (cut-off for sequence separation – $2 \leq \mathrm{l}_{\mathrm{cut}} \leq 14$). S_{ij} is the sequence separation between the contacting *i*th and *j*th residue. When same distance cut-off and sequence separation limits are used, the product of CO and LRO yields TCD. Jackknife correlation was used to test the efficacy of the method. A correlation coefficient of 0.89 between experimentally obtained folding rates and predicted (using TCD) folding rates was observed. When the same dataset was used to examine the prediction efficiency of CO and LRO methods, a correlation coefficient of 0.71 and 0.80, respectively, was observed.

11.4.4 Chain Topology Parameter

Chain topology parameter (CTP; Nolting et al. 2003) is very similar to the previously explained contact order. CTP is calculated using Eq. 11.5:

$$\mathrm{CTP} = \frac{1}{L \bullet N} \sum \Delta \mathrm{S}_{i,j}^{2} \tag{11.5}$$

where 'L' is the number of residues in proteins, 'N' is the total number of contacts and '$\mathrm{S}_{i,j}$' is the sequence separation for the contacting residues 'i' and 'j'. This equation was formulated in order to find out the best exponential power defining the curvature observed for the plot of CO versus $\ln k_f$, which indicated a non-linear relationship between the sequence separation and free energy change. The dataset used to study the relationship between k_f and CTP consisted of 20 proteins and 2 small peptides (16-residue β-hairpin and a 10-residue helical polyalanine peptide) structures. A correlation coefficient of 0.86 was observed between CTP and folding rate of these proteins. The prediction of folding rates using CO was weaker when compared to CTP which worked preferably well for small peptides. It was concluded from these studies that protein having k_f in the range of $10^{-1}\ \mathrm{s}^{-1} - 10^{8}\ \mathrm{s}^{-1}$ could be efficiently predicted using CTP.

11.4.5 Cliquishness

Based on contact order, another parameter known as cliquishness was developed (Micheletti 2003). Cliquishness/clustering coefficient for the ith residue is calculated using Eq. 11.6 given below:

$$\text{Cliquishness}\ (i) = \frac{\sum\limits_{j<l} \Delta_{ij}\Delta_{il}\Delta_{jl}}{\mathrm{N_c}(\mathrm{N_c}-1)/2} \tag{11.6}$$

where 'Δ_{ij}', 'Δ_{il}' and 'Δ_{jl}' is 1 when 'i' and 'j', 'i' and 'l' and 'j' and 'l' are in contact, respectively. Else, it will be zero. 'N_c' is the total number of contacts of residue 'i'. It defines the cross-interaction of the residues interacting with the ith residue. A correlation coefficient of around 0.6 was obtained between cliquishness and folding rates of two-state folding proteins.

Protein folding is a complex process – this puzzle is a combination of many questions. What is the protein folding code? How is the native state of protein acquired? What is the mechanism of this process? How fast the primary sequence folds to its native state? How many intermediates are involved in this process? What are the structures of these partially unfolded states and so on? A combination of experimental and computational techniques has been used to understand this process. A lot of progress has been made in providing theoretical insights to the energy landscape of proteins, protein designing and protein folding simulations. These solutions indeed help towards finding a cure or treating fatal diseases associated with the misfolding of proteins. Still a larger part of this problem remains unsolved and a lot is yet to be learned and discovered. The freely accessible archive of protein structures – the PDB – acts as a goldmine for the theoretical scientists. Here we introduced our readers to the field of protein folding and hope that it will aid in developing their interest in this captivating research arena.

References

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

Baker D (2000) A surprising simplicity to protein folding. Nature 405(6782):39–42. https://doi.org/10.1038/35011000

Balchin D, Hayer-Hartl M, Hartl FU (2016) In vivo aspects of protein folding and quality control. Science 353(6294):aac4354. https://doi.org/10.1126/science.aac4354

Berman HM (2008) The protein data bank: a historical perspective. Acta Crystallogr A: Found Crystallogr 64(Pt 1):88–95. https://doi.org/10.1107/s0108767307035623

Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2012) The protein data bank at 40: reflecting on the past to prepare for the future. Structure 20(3):391–396. https://doi.org/10.1016/j.str.2012.01.010

Betts S, King J (1999) There's a right way and a wrong way: in vivo and in vitro folding, misfolding and subunit assembly of the P22 tailspike. Structure 7(6):R131–R139

Bross P, Corydon TJ, Andresen BS, Jorgensen MM, Bolund L, Gregersen N (1999) Protein misfolding and degradation in genetic diseases. Hum Mutat 14(3):186–198. https://doi.org/10.1002/(sici)1098-1004(1999)14:3<186::aid-humu2>3.0.co;2-j

Campbell NA, Reece JB, Urry LA, Cain LM, Wasserman SA, Minorsky PV, Jackson RB (2008) Biology, 8th edn. Pearson, San Francisco

Carugo O, Pongor S (2002) Recent progress in protein 3D structure comparison. Curr Protein Pept Sci 3(4):441–449

Chan HS, Dill KA (1990) Origins of structure in globular proteins. Proc Natl Acad Sci 87 (16):6388–6392. https://doi.org/10.1073/pnas.87.16.6388

Chaudhary P, Naganathan AN, Gromiha MM (2015) Folding RaCe: a robust method for predicting changes in protein folding rates upon point mutations. Bioinformatics 31(13):2091–2097. https://doi.org/10.1093/bioinformatics/btv091

Chaudhuri TK, Paul S (2006) Protein-misfolding diseases and chaperone-based therapeutic approaches. FEBS J 273(7):1331–1349. https://doi.org/10.1111/j.1742-4658.2006.05181.x

Chi PB, Liberles DA (2016) Selection on protein structure, interaction, and sequence. Protein Sci 25 (7):1168–1178. https://doi.org/10.1002/pro.2886

Cohen FE (1999) Protein misfolding and prion diseases. J Mol Biol 293(2):313–320. https://doi.org/10.1006/jmbi.1999.2990

Creighton TE (1995) Protein folding. An unfolding story. Curr Biol 5(4):353–356

Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4(1):10–19

Dill KA, MacCallum JL (2012) The protein-folding problem, 50 years on. Science 338 (6110):1042–1046. https://doi.org/10.1126/science.1219021

Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA (2007) The protein folding problem: when will it be solved? Curr Opin Struct Biol 17(3):342–346. https://doi.org/10.1016/j.sbi.2007.06.001

Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37:289–316. https://doi.org/10.1146/annurev.biophys.37.092707.153558

Dinner AR, Karplus M (2001) The roles of stability and contact order in determining protein folding rates. Nat Struct Biol 8(1):21–22. https://doi.org/10.1038/83003

Dobson CM (1999) Protein misfolding, evolution and disease. Trends Biochem Sci 24(9):329–332

Dokholyan NV, Li L, Ding F, Shakhnovich EI (2002) Topological determinants of protein folding. Proc Natl Acad Sci 99(13):8637–8641. https://doi.org/10.1073/pnas.122076099

Eftekharzadeh B, Hyman BT, Wegmann S (2016) Structural studies on the mechanism of protein aggregation in age related neurodegenerative diseases. Mech Ageing Dev 156:1–13. https://doi.org/10.1016/j.mad.2016.03.001

Fadiel A, Eichenbaum KD, Hamza A, Tan O, Lee HH, Naftolin F (2007) Modern pathology: protein mis-folding and mis-processing in complex disease. Curr Protein Pept Sci 8(1):29–37

Fersht AR (2000) Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. Proc Natl Acad Sci 97(4):1525–1529

Gianni S, Jemth P (2016) Protein folding: vexing debates on a fundamental problem. Biophys Chem 212:17–21. https://doi.org/10.1016/j.bpc.2016.03.001

Gipson B, Hsu D, Kavraki LE, Latombe JC (2012) Computational models of protein kinematics and dynamics: beyond simulation. Annu Rev Anal Chem 5:273–291. https://doi.org/10.1146/annurev-anchem-062011-143024

Godzik A, Kolinski A, Skolnick J (1993) De novo and inverse folding predictions of protein structure and dynamics. J Comput Aided Mol Des 7(4):397–438

Grantcharova V, Alm EJ, Baker D, Horwich AL (2001) Mechanisms of protein folding. Curr Opin Struct Biol 11(1):70–82

Gromiha MM (2003) Importance of native-state topology for determining the folding rate of two-state proteins. J Chem Inf Model 43(5):1481–1485. https://doi.org/10.1021/ci0340308

Gromiha MM, Huang LT (2011) Machine learning algorithms for predicting protein folding rates and stability of mutant proteins: comparison with statistical methods. Curr Protein Pept Sci 12 (6):490–502

Gromiha MM, Selvaraj S (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J Mol Biol 310(1):27–32. https://doi.org/10.1006/jmbi.2001.4775

Gromiha MM, Thangakani AM, Selvaraj S (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucleic Acids Res 34:W70–W74. https://doi.org/10.1093/nar/gkl043

Harrison RS, Sharpe PC, Singh Y, Fairlie DP (2007) Amyloid peptides and proteins in review. Rev Physiol Biochem Pharmacol 159:1–77. https://doi.org/10.1007/112_2007_0701

Hartl FU, Bracher A, Hayer-Hartl M (2011) Molecular chaperones in protein folding and proteostasis. Nature 475(7356):324–332. https://doi.org/10.1038/nature10317

Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV (2003) Contact order revisited: influence of protein size on the folding rate. Protein Sci 12(9):2057–2062. https://doi.org/10.1110/ps.0302503

Ivarsson Y, Travaglini-Allocatelli C, Brunori M, Gianni S (2008) Mechanisms of protein folding. Eur Biophys J 37(6):721–728. https://doi.org/10.1007/s00249-007-0256-x

Jaenicke R (1995) Folding and association versus misfolding and aggregation of proteins. Philos Trans R Soc Lond Ser B Biol Sci 348(1323):97–105. https://doi.org/10.1098/rstb.1995.0050

Karplus M (1997) The Levinthal paradox: yesterday and today. Fold Des 2(4):S69–S75

Khoury GA, Smadbeck J, Kieslich CA, Floudas CA (2014) Protein folding and de novo protein design for biotechnological applications. Trends Biotechnol 32(2):99–109. https://doi.org/10.1016/j.tibtech.2013.10.008

Kleywegt GJ, Jones TA (1996) Phi/psi-chology: Ramachandran revisited. Structure 4 (12):1395–1400. https://doi.org/10.1016/S0969-2126(96)00147-5

Kuwajima K, Schmid FX (1984) Experimental studies of folding kinetics and structural dynamics of small proteins. Adv Biophys 18:43–74

Lazaridis T, Karplus M (2003) Thermodynamics of protein folding: a microscopic view. Biophys Chem 100(1–3):367–395

Lim VI (1974) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J Mol Biol 88(4):857–862. https://doi.org/10.1016/0022-2836(74)90404-5

Luheshi LM, Crowther DC, Dobson CM (2008) Protein misfolding and disease: from the test tube to the organism. Curr Opin Chem Biol 12(1):25–31. https://doi.org/10.1016/j.cbpa.2008.02.011

Martin ACR (2000) The ups and downs of protein topology; rapid comparison of protein structure. Protein Eng 13(12):829–837. https://doi.org/10.1093/protein/13.12.829

Micheletti C (2003) Prediction of folding rates and transition-state placement from native-state geometry. Proteins 51(1):74–84. https://doi.org/10.1002/prot.10342

Miles AJ, Wallace BA (2016) Circular dichroism spectroscopy of membrane proteins. Chem Soc Rev 45:4859. https://doi.org/10.1039/c5cs00084j

Mogk A, Mayer MP, Deuerling E (2002) Mechanisms of protein folding: molecular chaperones and their application in biotechnology. Chembiochem 3(9):807–814. https://doi.org/10.1002/1439-7633(20020902)3:9<807::aid-cbic807>3.0.co;2-a

Nelson DL, Lehninger AL, Cox MM (2005) Lehninger principles of biochemistry. W.H. Freeman, New York

Nguyen DN, Becker GW, Riggin RM (1995) Protein mass spectrometry: applications to analytical biotechnology. J Chromatogr A 705(1):21–45

Nolting B, Schalike W, Hampel P, Grundig F, Gantert S, Sips N, Bandlow W, Qi PX (2003) Structural determinants of the rate of protein folding. J Theor Biol 223(3):299–307

Ogen-Shtern N, Ben David T, Lederkremer GZ (2016) Protein aggregation and ER stress. Brain Res 1648:658–666. https://doi.org/10.1016/j.brainres.2016.03.044

Pace CN, Treviño S, Prabhakaran E, Scholtz JM (2004) Protein structure, stability and solubility in water and other solvents. Philos Trans R Soc Lond Ser B Biol Sci 359(1448):1225–1235. https://doi.org/10.1098/rstb.2004.1500

Park S, Yang X, Saven JG (2004) Advances in computational protein design. Curr Opin Struct Biol 14(4):487–494. https://doi.org/10.1016/j.sbi.2004.06.002

Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277(4):985–994. https://doi.org/10.1006/jmbi.1998.1645

Richa T, Sivaraman T (2012) OneG: a computational tool for predicting cryptic intermediates in the unfolding kinetics of proteins under native conditions. PLoS One 7(3):e32465. https://doi.org/10.1371/journal.pone.0032465

Richa T, Sivaraman T (2014) OneG-Vali: a computational tool for detecting, estimating and validating cryptic intermediates of proteins under native conditions. RSC Adv 4(68):36325–36335. https://doi.org/10.1039/C4RA04642K

Riddle DS, Grantcharova V (1999) Experiment and theory highlight role of native state topology in SH3 folding. Nat Struct Biol 6(11):1016–1024. https://doi.org/10.1038/14901

Rose GD, Fleming PJ, Banavar JR, Maritan A (2006) A backbone-based theory of protein folding. Proc Natl Acad Sci 103(45):16623–16633. https://doi.org/10.1073/pnas.0606843103

Santucci R, Sinibaldi F, Fiorucci L (2008) Protein folding, unfolding and misfolding: role played by intermediate states. Mini Rev Med Chem 8(1):57–62

Schafer NP, Hoffman RM, Burger A, Craig PO, Komives EA, Wolynes PG (2012) Discrete kinetic models from funneled energy landscape simulations. PLoS One 7(12):e50635. https://doi.org/10.1371/journal.pone.0050635

Sheu BC, Lin YH, Lin CC, Lee AS, Chang WC, Wu JH, Tsai JC, Lin S (2010) Significance of the pH-induced conformational changes in the structure of C-reactive protein by dual polarization interferometry. Biosens Bioelectron 26(2):822–827. https://doi.org/10.1016/j.bios.2010.06.001

Sikder AR, Zomaya AY (2005) An overview of protein-folding techniques: issues and perspectives. Int J Bioinforma Res Appl 1(1):121–143. https://doi.org/10.1504/ijbra.2005.006911

Tompa P, Rose GD (2011) The Levinthal paradox of the interactome. Protein Sci 20(12):2074–2079. https://doi.org/10.1002/pro.747

Walker LC, Levine H 3rd, Mattson MP, Jucker M (2006) Inducible proteopathies. Trends Neurosci 29(8):438–443. https://doi.org/10.1016/j.tins.2006.06.010

Welch WJ (2004) Role of quality control pathways in human diseases involving protein misfolding. Semin Cell Dev Biol 15(1):31–38. https://doi.org/10.1016/j.semcdb.2003.12.011

Winklhofer KF, Tatzelt J, Haass C (2008) The two faces of protein misfolding: gain- and loss-of-function in neurodegenerative diseases. EMBO J 27(2):336–349. https://doi.org/10.1038/sj.emboj.7601930

Zhou H, Zhou Y (2002) Folding rate prediction using total contact distance. Biophys J 82(1 Pt 1):458–463. https://doi.org/10.1016/s0006-3495(02)75410-6

Quality Assessment of Protein Tertiary Structures: Past, Present, and Future

12

Ankita Singh, Rahul Kaushik, and B. Jayaram

12.1 Introduction

Unavailability of protein tertiary structures is a foremost bottleneck in structural biology for gaining better insights into biological functions of proteins (Jayaram et al. 2006; Cheng 2008; Jayaram et al. 2014). Cost and time efficiency involved in experimental methods of structure elucidation via X-ray crystallography and NMR spectroscopy restrict it further. On the other hand, owing to worldwide genome projects, the relatively faster rate of new protein sequences being explored is increasing the gap between known protein sequences and experimentally solved structures (Fise 2010; Kaushik and Jayaram 2016). Despite continual improvements in the experimental methods of structural determination of proteins, the number of available protein structures is limited to ~22% of known protein sequences (Fig. 12.1). However, when only unique proteins are accounted, this percentage availability of protein structures declines to ~11% of the known protein sequences. In the present scenario, the yearly growth of Protein Data Bank (PDB) indicates insufficiency of experimental methods in bridging this gap, thus necessitating

A. Singh
Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi, New Delhi, India

Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan, India

R. Kaushik
Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi, New Delhi, India

Kusuma School of Biological Sciences, Indian Institute of Technology, New Delhi, India

B. Jayaram (✉)
Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi, New Delhi, India

Kusuma School of Biological Sciences, Indian Institute of Technology, New Delhi, India

Department of Chemistry, Indian Institute of Technology, New Delhi, India
e-mail: bjayaram@scfbio-iitd.res.in

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_12

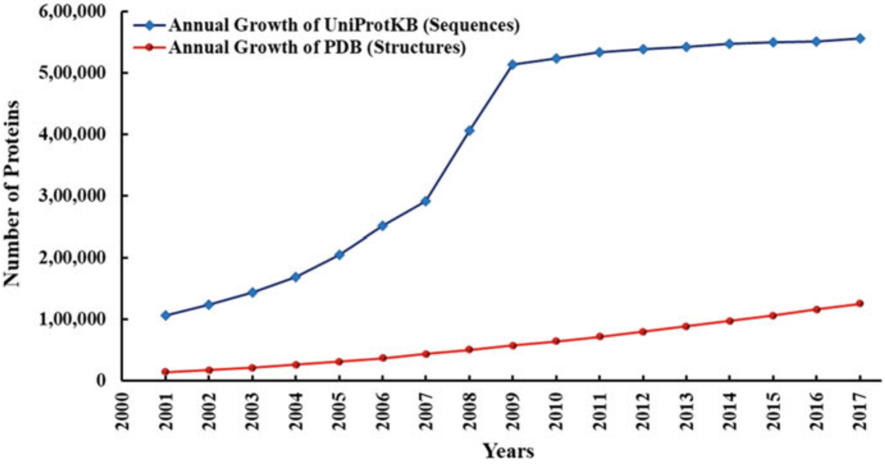


Fig. 12.1 Comparison of annual growth of PDB and UniProtKB (as of 13th November, 2017)

development of reliable computational approaches for structure prediction and their quality assessment (Samudrala and Levitt 2000; Kryshtafovych and Fidelis 2009; Wang et al. 2011).

In recent years, the protein structure prediction community has dedicated huge endeavor to predict more accurate structural models of proteins, and consistent improvements have been reported through Critical Assessment of Protein Structure Prediction (CASP) experiments (Zemla et al. 2001; Bourne 2003; Venclovas et al. 2003). The CASP experiments emphasize on chronological assessment of the developments in the field of protein structure prediction biennially via its blind prediction. The interested research group can register for the CASP experiment in different categories including tertiary structure prediction, protein structure quality assessment, contact prediction, data-assisted modeling, and protein structure refinement category. The participant state-of-the-art methods for different categories are subjected to perform for the given targets, and the submitted predictions are evaluated by the organizers. After every CASP experiment, the rankings of the participants are released for different categories. There have been 12 previous CASP experiments. The details pertaining to any individual CASP experiment can be accessed at http://predictioncenter.org. The lately introduced quality assessment (QA) category of these experiments ensures better quality of predicted model structures (Melo and Sali 2007). With highly efficient quality assessment methods, the model structures resulting from continuous improvements in structure prediction approaches can be optimally utilized for ligand-binding studies, drug designing, and biological function annotations of proteins. The fields of protein structure prediction and protein structure quality assessment have evolved together, leading to some of the best state-of-the-art methods in respective fields (Zhang and Skolnick 2004; Narang et al. 2005; Cozzetto et al. 2007).

In this chapter, we focus on the various quality assessment methods along with their merits and limitations.

12.2 Classification of Quality Assessment of Protein Structures

The quality assessment of protein structures can be performed mainly in two ways, namely, single-model-based QA (Cao et al. 2014) which evaluates the quality of individual model without using the information of other models and multiple-model-based QA (McGuffin and Roche 2010) which uses the mutual structural similarity among the models of the same protein to assess their quality. The various methods performing single-model-based QA and multiple-model-based QA are discussed in detail, further in this chapter. The single-model-based QA methods execute better with models having diverse quality range. Most of the currently available single-model-based QA methods implement evolutionary statistics, residue environment, structural features, and physics-based information while performing quality assessment. The multiple-model-based QA methods often perform better when the decoy set of model structures is derived from different structure prediction methods. The performance of these QA methods is hugely affected by the proportion of good quality model structures and poor quality model structures in the decoy set. The relative ranking among the diverse model structures for the same protein can also be performed by implementation of structural topology-based clustering methods. The main assumption for clustering-based approaches is the tendency of native-like structures to get clustered in a large free energy basin. Therefore, these usually end up with an average model rather than the best model (Cao and Jianlin 2016; Jing and Dong 2017).

12.3 Algorithms in Quality Assessment of Protein Structures

Generation of a large number of protein structural decoys for a given sequence is made possible with the recent developments in protein structure prediction field which has necessitated development of highly efficient quality assessment methods to precisely discriminate good model structures from bad ones. The initial efforts in the field were focused on the detection of the erroneous experimentally solved protein structures. With the passage of time, the field has advanced to evaluating model structures. The various approaches implemented in such methods can be broadly classified into three categories. These approaches and their implementation in different softwares/tools are discussed.

12.3.1 Physics-Based Approaches

This approach implemented mainly in the form of energy functions accounting for various bonded and nonbonded interactions among all the atoms of protein structures for use in energy minimization/geometry optimization and molecular dynamics simulations. The bonded interactions include distances, angles, and torsional angles, while nonbonded interactions account for electrostatic and van der Waals interactions. AMBER (http://amber.scripps.edu; Pearlman et al. 1995), CHARMM (http://www.charmm.org; Brooks et al. 1983), and GROMOS (http://www.igc.ethz.ch/gromos/; Soares et al. 2005) are among the most widely used packages which implement physics-based energy functions that incorporate diverse parameters derived from experiment and quantum mechanical calculations. Various versions of these functions are proposed which have succeeded in discriminating native-like structures from non-native-like structures efficiently, and integration of solvation terms has played a vital role in the success achieved by these approaches (Melo and Feytmans 1998). The various energies calculated via these energy functions directly give the estimation of quality of protein structures. Ideally, these energies should be highly negative for a good protein structure depending upon the size of the protein under consideration.

12.3.2 Knowledge-Based Approaches

The knowledge-based approaches implement rules of evolution and statistical preferences of different amino acid residues for performing quality assessment of protein structures. Usually these preferences are derived from a benchmark dataset consisting of selected experimentally solved protein structures. Since 1970, attempts have been made to derive some statistical rules from protein structures (Sippl 1995). The inverse Boltzmann distribution is most widely implemented to calculate pseudo-energies from nonredundant set of protein structures. Further, several other parameters are also used in evaluation of structural features of protein model structures which include contact-based features (Olechnovic and Venclovas 2017), distance-based features, solvent accessibility-based features, and pairwise interactions.

The concept of statistical preferences was initially successfully utilized in 1990 via PROSA (Wiederstein and Sippl 2007) which accounted for C_α/C_β distances and benchmarked it with the data taken from 163 protein structures. It was claimed to detect native conformation with a considerably high success rate. The random interaction approximation with lately developed alternative reference state has further boosted the precision of protein structure quality assessment. DFIRE (Yang and Zhou 2008) scoring function used an ideal gas reference state and showed 84% success rate in identifying native structures from 32 decoy sets. Likewise, discrete optimized protein energy (DOPE) scoring function (Shen and Sali 2006) employed noninteracting atom-dependent reference state in a homogeneous sphere and segregated 87% native structures in the 32 decoy sets. MolProbity uses a variety of physics-based and knowledge-based algorithms to assess a structure. More recently, some new scoring functions utilizing relative orientation of different residues have been developed and shown promising prospects.

Overall knowledge-based approaches proved advantageous in the evaluation of protein structures. The implementation of various reference states may deliver the desired precision for large-scale protein structure quality assessment.

12.3.3 Consensus-Based Approaches

The accuracy achieved through "physics-based approaches" and "knowledge-based approaches" can be boosted by implementation of "consensus-based approach". Various scoring functions have been established via weighted integration of discrete functions from physics- and knowledge-based approaches. The collective effect of weighted individual scores has been shown to outperform their individual accuracies.

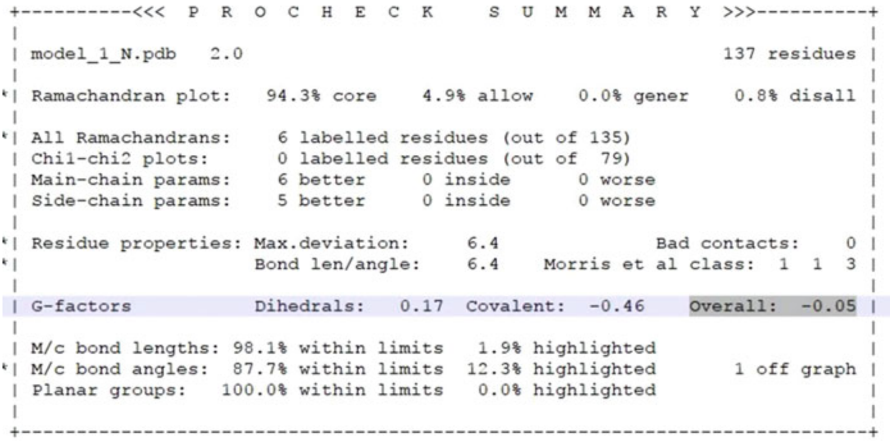
For instance, ProQ (Wallner and Elofsson 2003) uses a number of structural features, viz., atom- and residue-based potentials, secondary structure information, solvent accessibility, Cα–Cα distances, and globularity, implemented via a neural network. QMEAN (Benkert et al. 2009) is one of the more recently developed scoring functions which employs linear combination discrete scores from torsional angle potentials, secondary structure pairwise potentials, solvent accessibility, etc. Among the latest developments in the field, pcSM (Mishra et al. 2013) and D2N (Mishra et al. 2014) also implement consensus-based approach for efficient quality assessment. Very recently metaserver-based approaches have been introduced in the field of protein structure quality assessment and have shown improved results over individual servers.

12.4 Individual Servers/Tools for Quality Assessment

The servers/tools which integrate the features directly derived from protein structures and perform the quality assessment in completely independent manner are categorized in individual servers/tools for quality assessment. Some of the thoroughly validated and highly accurate tools are described.

12.4.1 PROCHECK

PROCHECK (Laskowski 1993) evaluates "stereo chemical quality" of a given protein structure by calculating deviation in the geometry of residues from their standard values which are derived from well-refined, high-resolution native structures. It accounts for bond lengths, angles, main-chain and side-chain parameters, residue contacts, geometry, and distribution of backbone torsion angles (Φ and Ψ) in Ramachandran plot while evaluating the stereochemical quality of a



```
+----------<<<  P  R  O  C  H  E  C  K     S  U  M  M  A  R  Y  >>>----------+
|                                                                            |
| model_1_N.pdb   2.0                                           137 residues |
|                                                                            |
*| Ramachandran plot:   94.3% core    4.9% allow    0.0% gener    0.8% disall |
|                                                                            |
*| All Ramachandrans:    6 labelled residues (out of 135)                    |
| Chi1-chi2 plots:      0 labelled residues (out of  79)                    |
| Main-chain params:    6 better     0 inside       0 worse                 |
| Side-chain params:    5 better     0 inside       0 worse                 |
|                                                                            |
*| Residue properties: Max.deviation:     6.4              Bad contacts:    0 |
*|                     Bond len/angle:    6.4    Morris et al class:  1  1  3 |
|                                                                            |
| G-factors           Dihedrals:   0.17  Covalent:  -0.46     Overall:  -0.05 |
|                                                                            |
| M/c bond lengths: 98.1% within limits   1.9% highlighted                  |
*| M/c bond angles:  87.7% within limits  12.3% highlighted       1 off graph |
| Planar groups:   100.0% within limits   0.0% highlighted                  |
|                                                                            |
+----------------------------------------------------------------------------+
```

Fig. 12.2 Summary of overall quality assessment performed via PROCHECK

protein structure. For a good protein structure, most of the residues (represented by solid blue squares) should be falling in red-shaded regions (most favorable), followed by orange-shaded region (favorable regions) and yellow-shaded region (generously allowed regions).

Protein structure of interest in pdb file format is required as input. A summary of overall quality assessment using PROCHECK is shown in Fig. 12.2**.**

12.4.2 ProSA

ProSA (Wiederstein and Sippl 2007) quality assessment tool calculates overall protein structure quality score in terms of Z-score and performs statistical comparison with experimental protein structures. It exploits knowledge-based Cα potentials of mean force to estimate model accuracy. The Z-score lying within a defined range in the plot differentiates native-like protein structures from erroneous structures. The position dark black circle represents the quality of input structure, and ideally it should be falling in blue or light blue region for a good model structure. Protein structure of interest in pdb file format is required, and ProSA generates a graphical interpretation of assessment as shown in Fig. 12.3.

12.4.3 ProQ

ProQ (Wallner and Elofsson 2003) is a neural-network-based method for quality assessment of protein structure which utilizes various structural features, including frequency of atom–atom contacts, solvent-accessible surface area, and residue–

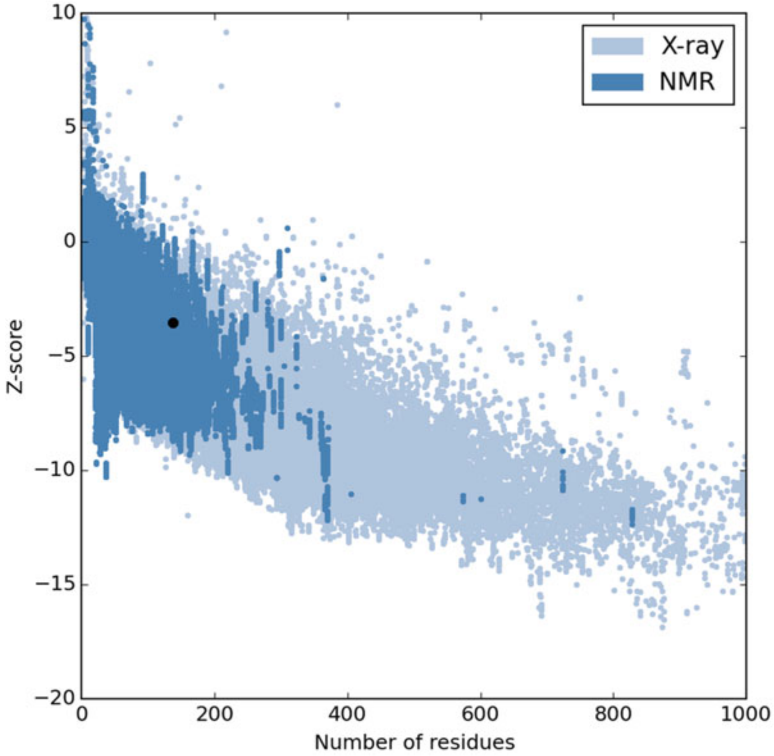


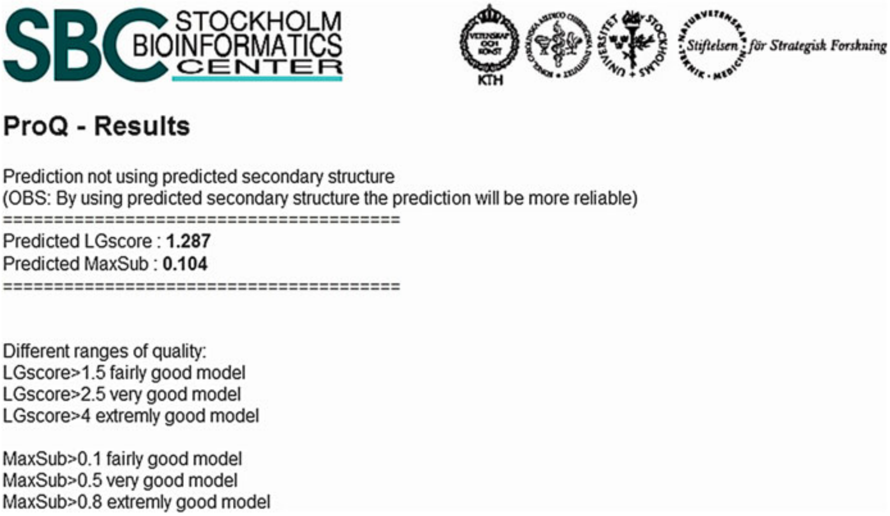
Fig. 12.3 A graphical interpretation of quality assessment performed using ProSA

residue contacts. Models are evaluated on the bases of both the LG score and the MaxSub score. The LG score is a P-value for the importance of a structural similarity, and higher values represent better quality of model structure. Similarly, the MaxSub score identifies the largest subset of correctly predicted Cα atoms of a model and furnishes a single normalized score which characterizes the quality of the model structures. The MaxSub score varies from 0 to 1, where 0 is insignificant and 1 most significant. Correctly predicted model structures should have at least 1.5 LG score and 0.1 MaxSub score. ProQ also requires protein structure of interest in pdb file format and predicts

LG score and MaxSub score as shown in Fig. 12.4.

12.4.4 Verify-3D

Verify-3D (Luthy et al. 1992) evaluates protein models by comparing 3D-1D profiles. The match of an atomic model (3D) is assessed with its own amino acid sequence (1D), and a quality score is predicted with parameters derived from


SBC STOCKHOLM BIOINFORMATICS CENTER
KTH
Stiftelsen för Strategisk Forskning

ProQ - Results

```
Prediction not using predicted secondary structure
(OBS: By using predicted secondary structure the prediction will be more reliable)
=======================================
Predicted LGscore : 1.287
Predicted MaxSub : 0.104
=======================================


Different ranges of quality:
LGscore>1.5 fairly good model
LGscore>2.5 very good model
LGscore>4 extremly good model

MaxSub>0.1 fairly good model
MaxSub>0.5 very good model
MaxSub>0.8 extremly good model
```

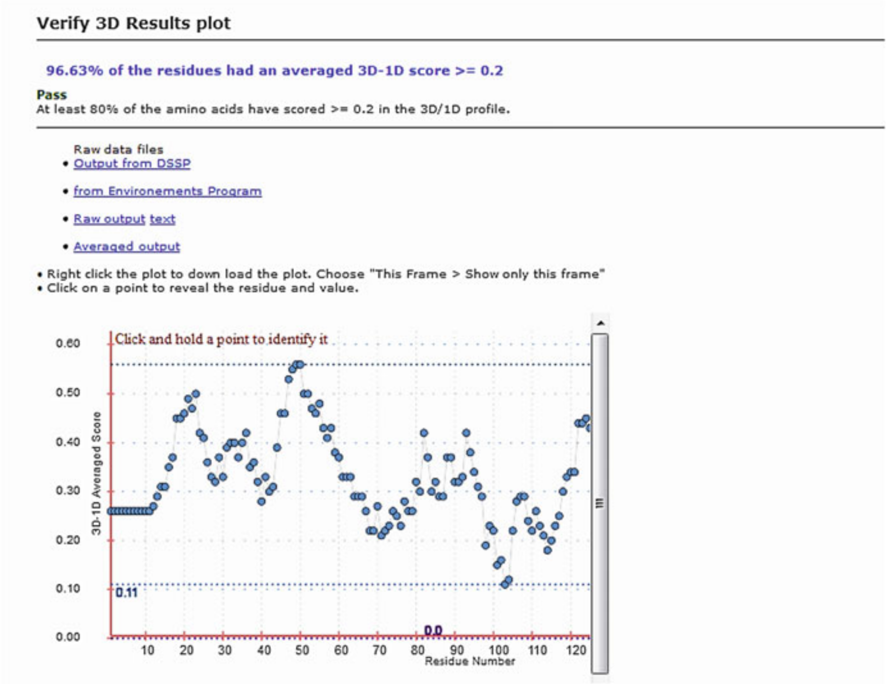
Fig. 12.4 A quality assessment performed using ProQ server

database of experimental structures. The 3D profile of a protein structure is calculated from the atomic coordinates of the experimentally solved structures. The 3D profiles derived from a protein structure should counterpart its sequence with a high score. An incorrectly modeled structure can be recognized by inspecting the profile score in a moving-window scan. Average 3D-1D profile score for each residue in a 21-residue sliding window is represented in the form of a plot. The scores vary from −1 (bad score) to +1 (good score). A protein structure of interest in pdb file format is required, and Verify-3D generates 3D-1D profile score for individual residue as shown in Fig. 12.5.

12.4.5 Naccess

Naccess (Lee and Richards 1971) calculates solvent-exposed surface area of all atoms and residues with defined probe size. The probe is moved around the van der Waal's surface of protein structure provided in pdb file format. Usually, the probe considered is of same radius as water (1.4 Å). Naccess helps in better insights on structure by comparing residue-wise surface area with experimental structure. Considering globular nature of monomeric proteins, lower overall exposed surface area indicates better quality of a protein structure.

Naccess generates two files having accessible surface area at atomic level (.asa file) and at residue level (.rsa file). An overall accessible surface area for the input structure is also provided in the output.



Verify 3D Results plot

96.63% of the residues had an averaged 3D-1D score >= 0.2

Pass
At least 80% of the amino acids have scored >= 0.2 in the 3D/1D profile.

Raw data files
- Output from DSSP
- from Environements Program
- Raw output text
- Averaged output

- Right click the plot to down load the plot. Choose "This Frame > Show only this frame"
- Click on a point to reveal the residue and value.

Fig. 12.5 A quality assessment plot generated by Verify-3D for an input protein structure.

12.4.6 QMEAN

Qualitative Model Energy Analysis (QMEAN) is a composite scoring function to evaluate the major geometrical aspects of protein structures (Benkert et al. 2009). It utilizes various structural descriptors such as analysis of local geometry of dihedral angles with a window size of three amino acids, secondary structure-specific pairwise long-range interactions, solvation potential derived from solvent-accessible surface area, etc. The long-range interactions among secondary structural elements and solvation potential take account of protein structure stability. The overall QMEAN quality score ranges from 0 to 1 where 0 represents badly modeled and 1 represents accurately modeled structures. A protein structure in pdb file format is required, and the overall quality score in the form of QMEAN score is calculated as shown in Fig. 12.6.

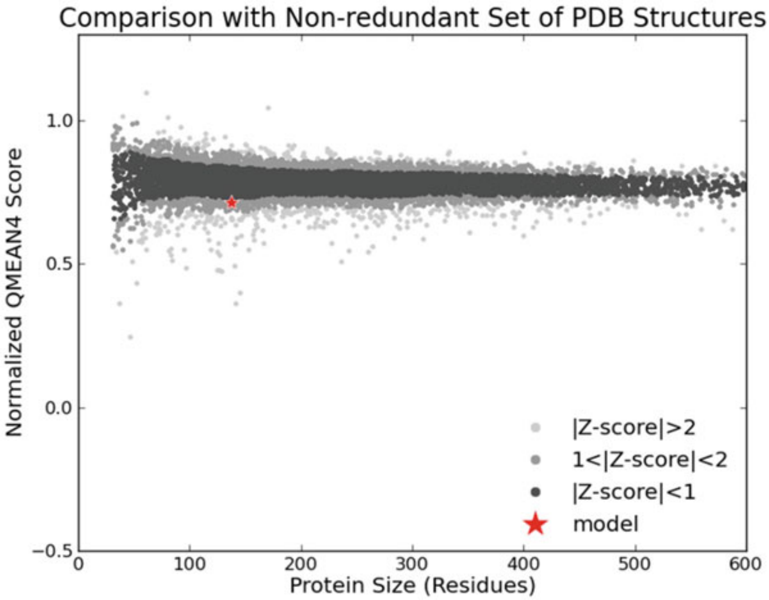


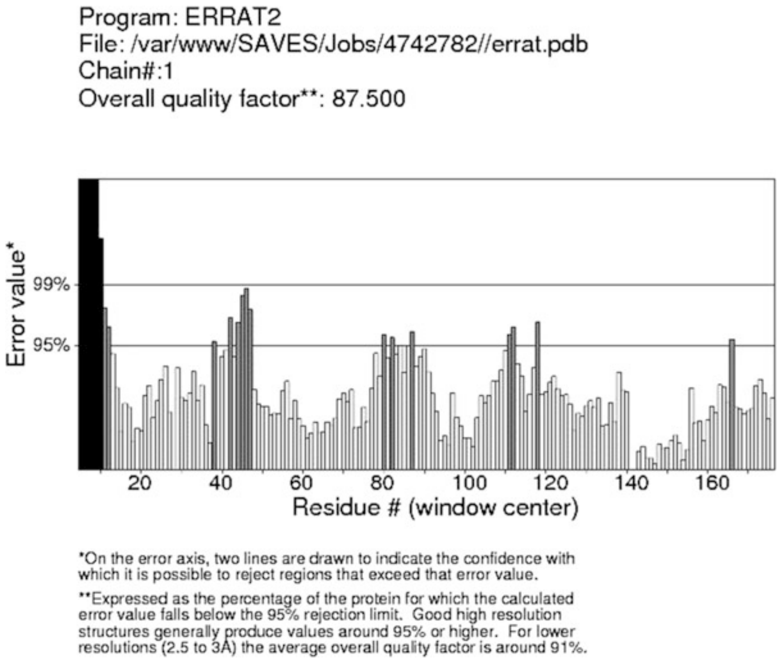
Fig. 12.6 Graphical representation of QMEAN quality assessment where the model structure is showing an overall QMEAN score of 0.71

12.4.7 Errat

Errat (Colovos and Yeates 1993) distinguishes correctly modeled regions from incorrectly modeled regions in the protein structures by evaluating certain characteristic atomic interactions. It also provides an overall quality factor for a protein structure expressed as the percentage of protein with error value falling under 95% limit. The proposed threshold overall quality factor is 91% for medium resolution structures and 95% or above for high-resolution (good) structures which is derived from experimental structures. The overall quality factor represents the percentage number of residues predicted accurately in the given protein. Moreover, the plot provides the residue-wise insight in terms of percentage error value per residue. The quality assessment performed via Errat helps to identify the error prone regions in the predicted model structures. Errat requires a protein structure in pdb file format as an input and generates a quality factor and residue-level assessment in the form of plot (Fig. 12.7).

12.4.8 PSN-Ensemble

PSN-Ensemble (Ghosh and Vishveshwara 2014) employs networks of experimental and modeled protein structures, integrated with support vector machines for performing quality assessment. These networks mainly take account of



```
Program: ERRAT2
File: /var/www/SAVES/Jobs/4742782//errat.pdb
Chain#:1
Overall quality factor**: 87.500
```

*On the error axis, two lines are drawn to indicate the confidence with which it is possible to reject regions that exceed that error value.

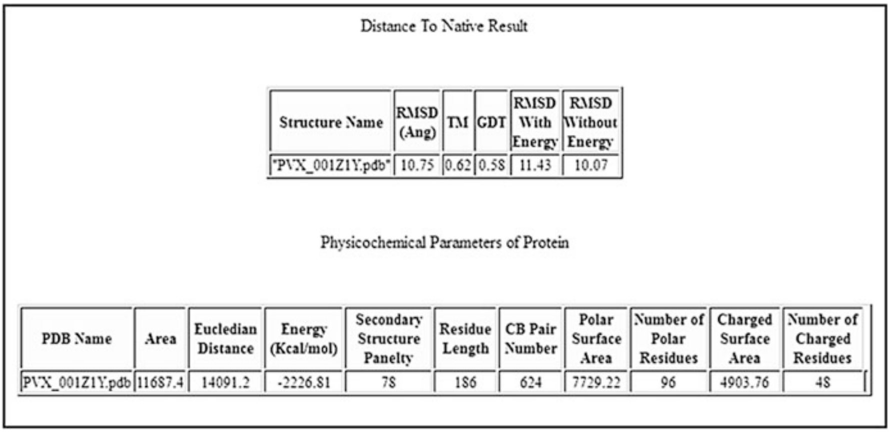
**Expressed as the percentage of the protein for which the calculated error value falls below the 95% rejection limit. Good high resolution structures generally produce values around 95% or higher. For lower resolutions (2.5 to 3A) the average overall quality factor is around 91%.

Fig. 12.7 A graphical illustration of quality assessment for an input protein structure via Errat

non-covalent interactions among side chains in protein structures. A PSN-QA score above 16 reflects features of native-like conformation, and a score below 10 is indicative of non-native-like conformation. A pdb file is required to generate overall quality assessment scores and relative ranking of protein structure.

12.4.9 D2N (Distance 2 Native)

D2N (Mishra et al. 2014) is a random forest approach-based machine learning tool which predicts the quality of a protein structure derived from different physicochemical features of native protein structures. It accounts for all atom nonbonded energy, solvent-accessible surface area of polar and charged residues, C_β geometrical constraint, and secondary structure penalties. It predicts root-mean-square deviation (RMSD), template modeling (TM) score, and global distance test (GDT) score for a given protein structure. The predicted parameters for the input protein structure are direct indicative of quality of structure. The RMSD is measured in angstroms (Å) and should be as low as possible. The TM score assesses the local accuracy of model structure and varies from 0 to 1 where 0 represents randomly modeled structure and 1 represents the most accurately modeled structure. The GDT score is another measure used for calculating the structural similarity. Similar to TM score, the GDT score also varies from 0 to 1 where 0 represents worse model and 1 represents



Distance To Native Result

Structure Name	RMSD (Ang)	TM	GDT	RMSD With Energy	RMSD Without Energy
"PVX_001Z1Y.pdb"	10.75	0.62	0.58	11.43	10.07

Physicochemical Parameters of Protein

PDB Name	Area	Eucledian Distance	Energy (Kcal/mol)	Secondary Structure Panelty	Residue Length	CB Pair Number	Polar Surface Area	Number of Polar Residues	Charged Surface Area	Number of Charged Residues
PVX_001Z1Y.pdb	11687.4	14091.2	-2226.81	78	186	624	7729.22	96	4903.76	48

Fig. 12.8 Quality assessment performed using D2N for input protein structure

best model structure. D2N requires protein structure of interest in pdb file format and generates predicted RMSD, TM score, and GDT score (Fig. 12.8).

12.4.10 dDFIRE

It is an energy function which reports for pairwise atomic interactions and dipole–dipole interactions among the amino acid residues. For a given protein structure, it provides a set of free energy scores. The energy score provided in the first column represents dDFIRE total energy which needs to be highly negative for a good model structure. The other energy terms give rise to the total energy via their algebraic addition. dDFIRE predicts the free energy scores of a given pdb file of a protein structure (Yang and Zhou 2008).

12.4.11 MolProbity

MolProbity (Chen et al. 2010) uses a variety of physics-based and knowledge-based algorithms to assess a structure. All-atom contacts, side-chain clashes, and Ramachandran distribution of backbone dihedral angles (ϕ and Ψ) are the major parameters for MolProbity-dependent protein structure quality assessment. An overall MolProbity score, derived from these parameters, with lower values reflects good quality of the model structures and vice versa.

MolProbity takes protein structure of interest in pdb file format as an input and predicts MolProbity score (in case of single-model structure) and MolProbity rank (in case of multiple-model structures).

12.5 Metaserver Approaches

The metaserver approaches integrate some of the already existing individual servers or the features calculated by individual servers to perform the quality assessment of protein structures. Since metaservers utilize the quality assessment performed by individual servers, therefore, these servers are expected to perform better than individual server. A metaserver may either provide a combined score which is indicative of quality of protein structure or provide a common platform to utilize the individual servers without any combined score. Different types of metaservers including MetaMQAP (Pawlowski et al. 2008), ProTSAV (Singh et al. 2015), and SAVES (http://services.mbi.ucla.edu/SAVES) are discussed.

12.5.1 MetaMQAP

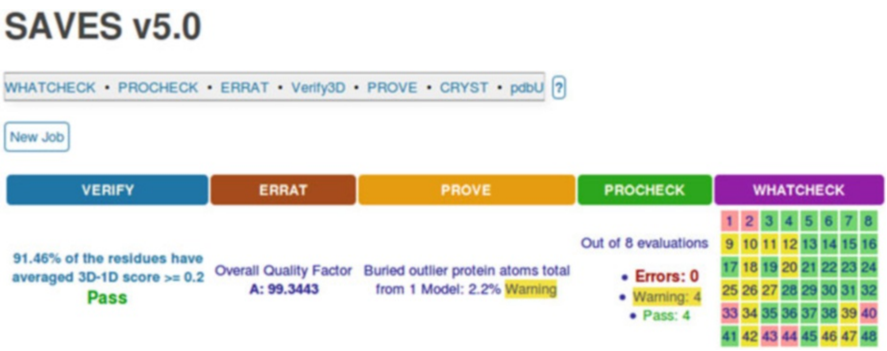
It is based on a multivariate regression model, which implements scores generated via eight different individual tools, namely, Verify-3D, ProSA, BALA, ANOLEA, PROVE, TUNE, REFINER, and PROQRES. MetaMQAP predicts the C-alpha atoms' absolute deviation in terms of RMSD for the given model structure along with its GDT score. The predicted local model accuracy has a correlation coefficient of 0.7, and the global score has a correlation coefficient of 0.9 with true deviations from native structures. The predicted scores have a significant improvement over all constituent individual tools. A pdb file of protein structure is required, and MetaMQAP predicts RMSD and GDT score.

12.5.2 SAVES

Structural Analysis and Verification Server (SAVES) is a metaserver which provides a common platform for performing quality assessment with five different individual servers/tools, namely, PROCHECK, What-Check, ERRAT, Verify-3D, and PROVE. It neither integrates these individual servers for predicting overall quality nor does it perform its own analysis for quality assessment of given structure. SAVES provides an interactive user interface for the analysis of results of individual servers. A protein structure of interest in pdb file format is required, and SAVES produces results of individual servers (Fig. 12.9).

12.5.3 ProTSAV

Protein tertiary structure analysis and validation (ProTSAV) is a metaserver which efficiently integrates ten different quality assessment individual tools (modules) to evaluate predicted model structures which necessarily take care of most of the quality assessment parameters used in previously discussed approach (Table 12.1). It outperforms the individual accuracies of constituent modules and provides an



91.46% of the residues have averaged 3D-1D score >= 0.2
Pass

Overall Quality Factor
A: 99.3443

Buried outlier protein atoms total from 1 Model: 2.2% Warning

Out of 8 evaluations
- Errors: 0
- Warning: 4
- Pass: 4

Fig. 12.9 A snapshot of SAVES quality assessment metaserver

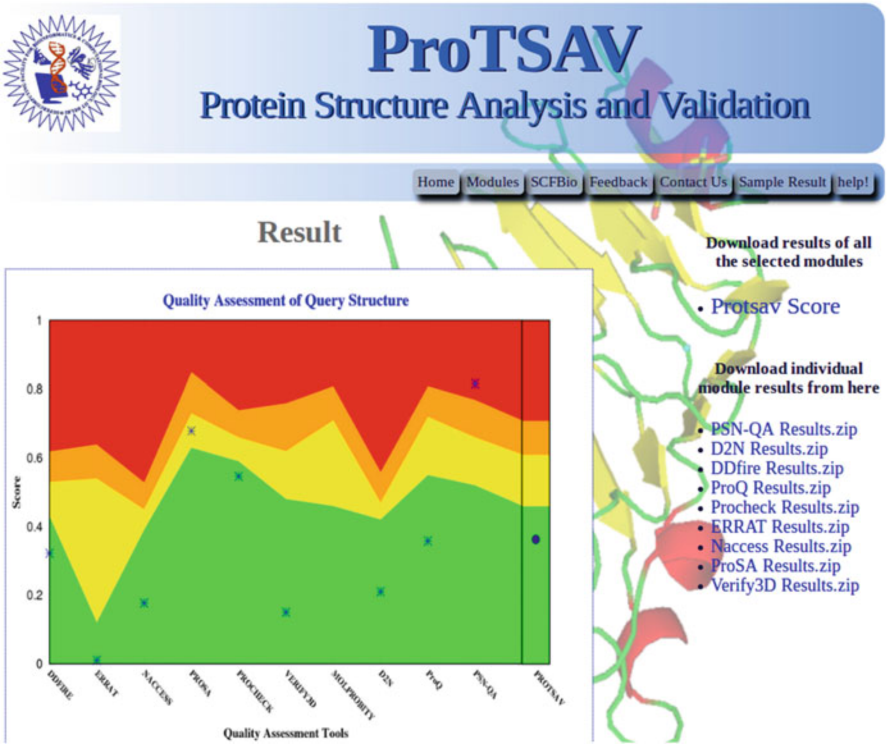
Table 12.1 Different modules of ProTSAV metaserver and their quality assessment parameter. The solid green-filled boxes represent the implemented parameter in respective module

The solid green filled boxes represent the implemented parameter in respective module

Assessment Features	DFIRE	Errat	Naccess	ProSA	Pro-check	Verify-3D	Mol-Probity	D2N	ProQ	PSN-QA
Van der Waals Clashes				🟩			🟩	🟩		
Contact Potential	🟩	🟩				🟩			🟩	🟩
Burial Preferences				🟩						
Accessible Surface Area			🟩						🟩	
Residue Packing		🟩								
Globularity			🟩						🟩	
Secondary Structure				🟩		🟩		🟩	🟩	
Φ and ψ Distribution					🟩		🟩			
Energy Based Scorings	🟩							🟩		
Side Chain Packing					🟩		🟩			🟩

overall score (ProTSAV score). The quality assessment via ProTSAV classifies a protein structure in any of the predefined four classes on the basis of its structural features. The predefined classes are indicated by green color (structures with RMSD 0–2 Å), yellow color (structures with RMSD 2–5 Å), orange color (structures with RMSD 5–8 Å), and red color (structures with RMSD beyond 8 Å). The ProTSAV score with lower value reflects better quality of protein structure and vice versa. ProTSAV is also capable of performing relative ranking in case of multiple model structures.

Coordinates of the structure of interest in PDB file format in case of individual model and zipped file (.zip) in case of multiple models are required as an input



Download results of all the selected modules

- Protsav Score

Download individual module results from here

- PSN-QA Results.zip
- D2N Results.zip
- DDfire Results.zip
- ProQ Results.zip
- Procheck Results.zip
- ERRAT Results.zip
- Naccess Results.zip
- ProSA Results.zip
- Verify3D Results.zip

Fig. 12.10 A snapshot of ProTSAV metaserver for protein tertiary structure quality assessment

ProTSAV generates a graphical representation of quality assessment for input protein structure into any of the predefined classes in case of individual model (Fig. 12.10) and relative rankings in case of multiple models.

12.6 Case Studies on CASP11 (T0760) and CASP12 (T0860) Targets

In this section the performance of different quality assessment tools/metaservers on protein tertiary structures modeled by Zhang-Server for a CASP11 target T0760 (http://www.predictioncenter.org/casp11/targetlist.cgi) and modeled by BhageerathH+ for a CASP12 target T0860 (http://predictioncenter.org/casp12/targetlist.cgi) is presented. These models were predicted by respective servers in the absence of their native experimental structures and post-CASP compared with native experimental structures to calculate their RMSD.

In the case study, we assessed the quality of the modeled structures via various quality assessment tools as shown in Table 12.2 and compared the quality assessment prediction with actual RMSD values.

Table 12.2 Quality assessment prediction scores via different tools/metaservers and their interpretations for structure (T0760, Zhang-Server Model 1, and T0860, BhageerathH+ Model 1). The actual RMSDs for T0760 and T0860 are 2.9 Å and 3.3 Å, respectively

Servers	T0760 scores	T0860 scores	Threshold values
Errat	91.4%	77.9%	>95% considered as good
Verify-3D	90.1%	92.0%	>95% considered as good
ProSA	−6.6	−3.5	> −5 considered as good
PSN-QA	5.3	5.5	>16 considered as good
QMEAN	0.45	0.71	Range 0 (bad) to 1 (good)
Naccess	12,819	8462	Lower values preferred
ProQ	LG score = 4.7	LG score = 4.0	>3 considered as good
	MaxSub score = 0.2	MaxSub score = 0.2	>0.5 considered as good
PROCHECK	−0.69 (overall G-score)	−0.05 (overall G-score)	> −0.5 considered as good
D2N	5.5 Å (RMSD)	6.4 Å (RMSD)	<5 Å RMSD
dDFire	−477.9	−274.8	Lower value is considered as good
MolProbity	3.3	2.3	>2 considered as good
MetaMQAP	4.2 Å (RMSD)	5.1 Å (RMSD)	<5 Å RMSD
ProTSAV	Yellow region	Yellow region	Green and yellow as good

The tertiary structure quality assessment performed via different methods (Table 12.2) indicates that the consensus approaches are able to perform better prediction as compared to individual method approaches. The individual servers, while capturing some of the quality assessment feature, miss out the other. On the other hand, the consensus approaches, implemented in metaservers, account for all the quality assessment features simultaneously and perform better predictions.

In recent years, methodological developments in the field of protein structure prediction and their availability to scientific community have raised the necessity of highly accurate protein structure quality assessment for better understanding of structural features and their further application, viz., drug designing, functional characterization, and annotation.

All approaches described in the chapter utilize the structural features and perform with considerable accuracy within their limitations. However, all QA server scores predict model quality differently according to their individual parameters which sometimes further complicate the assessment to select the best model. Thus, the metaserver approaches, where various combinations of multiple QA tools are used to address the issue, perform much better quality assessment of structures. The metaserver approaches, while combining the individual servers, also overcome their individual limitations with their efficient combination. Post-protein structure prediction and selection of best model structure are very critical and can be performed by implementing suitable quality assessment tools/servers/metaservers.

Acknowledgments Support to the Supercomputing Facility for Bioinformatics & Computational Biology (SCFBio), IIT Delhi, from the Department of Biotechnology, Govt. of India, is gratefully acknowledged.

References

Benkert P, Kuenzli M, Schwede T (2009) QMEAN server for protein model quality estimation. Nucleic Acids Res 37:W510–W514

Bourne PE (2003) CASP and CAFASP experiments and their findings. Methods Biochem Anal 44:501–507

Brooks BR, Bruccoleri RE, Olafson BD, States DJ et al (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217

Cao R, Jianlin C (2016) Protein single-model quality assessment by feature-based probability density function. Sci Rep 6:23990

Cao R, Wang Z, Wang Y, Cheng J (2014) SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC Bioinformatics 15:120

Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA et al (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D66:12–21

Cheng J (2008) A multi-template combination algorithm for protein comparative modeling. BMC Struct Biol 8:18

Colovos C, Yeates TO (1993) Verification of protein structures, patterns of non-bonded atomic interactions. Protein Sci 2:1511–1519

Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A (2007) Assessment of predictions in the model quality assessment category. Proteins 69(8):175–183

Fise A (2010) Template-based protein structure modeling. Methods Mol Biol 673:73–94

Ghosh S, Vishveshwara S (2014) Ranking the quality of protein structure models using side chain based network properties. F1000Res 3:17

Jayaram B, Bhushan K, Shenoy SR et al (2006) Bhageerath: an energy based web enabled computer soft-ware suite for limiting the search space of tertiary structures of small globular proteins. Nucleic Acids Res 34:6195–6204

Jayaram B, Dhingra P, Mishra A et al (2014) Bhageerath-H: a homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. BMC Bioinformatics 15(S16):S7

Jing X, Dong Q (2017) MQAPRank: improved global protein model quality assessment by learning-to-rank. BMC Bioinformatics 18:275

Kaushik R, Jayaram B (2016) Structural difficulty index: a reliable measure for modelability of protein tertiary structures. Protein Eng Des Sel 29(9):391–397

Kryshtafovych A, Fidelis K (2009) Protein structure prediction and model quality assessment. Drug Discov Today 14:386–390

Laskowski RA (1993) PROCHECK: a program to check the stereo chemical quality of protein structures. J Appl Crystallogr 26:283–291

Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55:379–400

Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356:83–85

McGuffin L, Roche D (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics 26:182–188

Melo F, Feytmans E (1998) Assessing protein structures with non-local atomic interaction energy. J Mol Biol 17:1141–1152

Melo F, Sali A (2007) Fold assessment for comparative protein structure modeling. Protein Sci 16:2412–2426

Mishra A, Rao S, Mittal A, Jayaram B (2013) Capturing native/native like structures with a Physico-chemical metric (pcSM) in protein folding. Biochim Biophys Acta 1834:1520–1531

Mishra A, Rana PS, Mittal A, Jayaram B (2014) D2N: distance to the native. Biochim Biophys Acta 10:1798–1807

Narang P, Bhushan K, Bose S, Jayaram B (2005) A computational pathway for bracketing native-like structures for small alpha helical globular proteins. Phys Chem Chem Phys 7:2364–2375

Olechnovic K, Venclovas C (2017) VoroMQA: assessment of protein structure quality using interatomic contact areas. Proteins 85:1131–1145

Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM (2008) MetaMQAP: a meta-server for the quality assessment of protein models. BMC Bioinformatics 9:403

Pearlman DA, Case DA, Caldwell JW, Ross WS et al (1995) AMBER 4.1. University of San Francisco, San Francisco

Samudrala R, Levitt M (2000) Decoys 'R' Us: a database of incorrect protein conformations to improve protein structure prediction. Protein Sci 9:1399–1401

Shen M, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15:2507–2524

Singh A, Kaushik R, Mishra A, Shanker A, Jayaram B (2015) ProTSAV: a protein tertiary structure analysis and validation server. Biochim Biophys Acta 1864:11–19

Sippl MJ (1995) Knowledge-based potentials for proteins. Curr Opin Struct Biol 5:229–235

Soares TA, Hünenberger PH, Kastenholz MA, Kräutle V et al (2005) An improved nucleic acid parameter set for the GROMOS force field. J Comput Chem 26:725–737

Venclovas C, Zemla A, Fidelis K, Moult J (2003) Assessment of progress over the CASP experiments. Proteins 53:585–595

Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073–1086

Wang Q, Shang Y, Xu D (2011) Improving a consensus approach for protein structure selection by removing redundancy. IEEE/ACM Trans Comput Biol Bioinform 8:1708–1715

Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucl Acids Res 35:W407–W410

Yang Y, Zhou Y (2008) Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 72:793–803

Zemla A, Venclovas C, Fidelis K, Moult J (2001) Processing and evaluation of predictions in CASP4. Proteins 5:13–21

Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710

Predicting Protein Function Using Homology-Based Methods

13

Swati Sinha, Birgit Eisenhaber, and Andrew M. Lynn

13.1 Introduction

In general, it is very important to define 'function' of a protein. There are two main aspects of protein function, namely, 'molecular function' and 'cellular function'. The molecular function of a protein is defined by various actions including binding, activation, inhibition or catalysis, while the cellular function tells us more about the context in which a protein operates within a cell (Marcotte et al. 1999, 2000). Homology that depicts common evolutionary ancestry among different organisms plays a crucial role in the prediction of protein function. Traditional homology-based methods transfer a function to an unknown protein based on the sequence similarity between the known and unknown protein. In other words, the function of a protein can be deciphered by analysing similarity of the protein with other proteins of well-characterized functions. In case of significant sequence similarity, the annotations of protein with known function are transferred to the protein with unknown function.

Here, in this chapter, we explain the advantages and limitations of different methods for protein function prediction and emphasize on improvements that were implemented for better prediction of protein function. Towards the end, we highlight a list of tools which are helpful to perform deep sequence analysis of a protein sequence with unknown function.

S. Sinha (✉) · B. Eisenhaber
Bioinformatics Institute (BII) Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
e-mail: swatis@bii.a-star.edu.sg

A. M. Lynn
School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_13

13.2 Sequence-Sequence Comparison Methods

One of the first and most informative steps to analyse an unknown protein involves sequence similarity-based search in order to identify homologous sequences. It is effective because sequences that share significant similarity can be predicted as homologous (Pearson 2013). The traditional and the most commonly used approach for sequence-sequence comparison between two proteins is BLAST.

13.2.1 BLAST

- **B**asic **L**ocal **A**lignment **S**earch **T**ool (**BLAST**) was developed in 1990 and is one of the major contributions in the area of sequence homology detection indicated by the high number of citations (64,388 as of 11 August 2017) of the original article (Altschul et al. 1990). The method identifies regions of local similarities between the sequences by comparing DNA or protein sequences to well-characterized sequences. The method performs 'local' alignments and uses a heuristic approach to accelerate the process of producing these alignments. It is known that most of the proteins are modular in nature, possessing one or more domains or motifs which can be uniquely correlated to functionality. The local alignments help to find sequence signatures representative of these motifs. BLAST has many different flavours including:
- **blastn** searches nucleotide databases with a nucleotide query.
- **blastp** searches a protein query against protein databases.
- **blastx** searches protein databases with a translated nucleotide query.
- **tblastn** searches translated nucleotide databases with a protein query.
- **tblastx** compares all the six-frame translations of a nucleotide query against translated nucleotide sequence database.

BLAST is one of the most widely used methods for the first line of preliminary analysis to predict the function of a protein, but the method is known to have a variety of limitations which are discussed subsequently.

13.2.1.1 Limitations of BLAST

Since its development, BLAST is one of the best tools for sequence homology detection; however, there are limitations of this method like the following:

- Sequences having low-complexity regions often give artificially high scores.
- BLAST-based annotation can fail to distinguish members of subfamilies which have significant variation or in other words lower level of sequence similarity among the different members of a protein family, for example, membrane proteins.
- Presence of a small number of substrate specificity-determining residues, for example, protein kinase subfamilies where high levels of interfamily sequence similarity allow the selection of false positives.

- The local alignments are not able to find similarity when there are discontinuous conserved patterns in a sequence which can only be captured by global alignment algorithms.

13.3 Sequence-Profile Comparison Methods

The limitations of sequence-sequence comparison methods necessitate the development of more sensitive methods using sequence-profile comparisons. The alignment of a set of homologous sequences is used to build a sequence profile which includes information about the probability of occurrence of all the amino acids along each of the columns in the multiple sequence alignment. This profile is more sensitive than a single sequence because it has information of all the sequences from the entire family (Söding 2005). PSI-BLAST is one such method (Altschul et al. 1997).

13.3.1 PSI-BLAST

Position-**S**pecific **I**terative BLAST (**PSI-BLAST**) is a flavour of BLAST which performs multiple iterations to search for distant homologs in a sequence database. The first step in PSI-BLAST is similar to BLAST which provides a number of hits for the query proteins. In the next step, it creates an alignment of all the high-scoring top hits from the first step and converts the alignment into a position-specific scoring matrix (PSSM) which is used in the subsequent steps. In a PSSM, the highly conserved residues in the alignment are assigned a comparatively higher score than the less conserved residues. The following step utilizes this PSSM to perform the similarity search between the profile and the sequence database. These steps are repeated iteratively until convergence, that is, no new sequences appear further in the iteration. This type of iterative search helps in the identification of remote homologs which could possibly be missed by methods like BLAST. Therefore, these methods are said to increase the sensitivity of identification of distant homologs. PSI-BLAST is faster than BLAST and could extensively find divergent homologs with low percent identity.

13.3.1.1 Limitations of PSI-BLAST

Since PSI-BLAST works on the principle of iterative search, one has to be very careful about selecting the homologous sequences in each step which are used for construction of a PSSM. Addition of non-homologous sequences in any iteration will further enable inclusion of more such sequences reducing the overall sensitivity of the method. Moreover, this method might provide incorrect results when the database includes sequences having low-complexity regions or transmembrane regions or coiled-coil regions which usually have a high ratio of biased amino acid residues. These regions are prone to reflect significant sequence similarity even in the absence of homology (Altschul et al. 1997).

13.4 Sequence-Profile HMM Comparison Methods

The simple sequence profiles (PSSMs) only contain the frequency of each of the amino acid in an alignment but do not have information for the insertions and deletions present in an alignment. Hence, these profiles are replaced by profile hidden Markov models (HMMs) (Eddy 1998) which contain position-specific probabilities for deletions and insertions along the alignment (Söding 2005). These profile HMMs tend to perform better than the sequence profiles for the detection of remote homologs. 'HMMER' (Eddy 2009) is one such method that illustrates biological sequence analysis using profile HMMs which are very useful for the identification of homologs.

13.4.1 HMMER

HMMER is one of the most commonly used methods to search sequence databases for distant homologs of protein sequences (Stein 2001; Yoon 2009) and also is the basis for the popular Pfam database (Finn et al. 2014). It implements methods based on probabilistic models called profile HMMs (Eddy 1998). In a profile HMM, each column is modelled by three states: a match state, an insert state and a delete state along with state transitions. A typical profile HMM has different types of probabilities like the transition probability which defines the transition from one state to another and the emission probability where each match state in a HMM emits a symbol with a defined probability of a residue at a particular position in an alignment.

In comparison with traditional sequence-sequence comparison methods (BLAST, FASTA), HMMER is known to be more accurate and detects distant homologs due to its strength based on underlying mathematical models. Earlier, the method was computationally very expensive, but the newer version of HMMER (HMMER3) is essentially equally fast as BLAST (Eddy 2011).

13.4.1.1 Limitations of HMMER

The conservation pattern in an alignment of protein families arises from 'fold'--specific signals shared across the entire family and 'function'-specific signals unique to the subfamily level. A profile HMM built from such an alignment will have both fold- and function-specific signals; therefore, it is prone to detect a large number of false positive (FP) sequences. For example, the multiple sequence alignment (MSA) in Fig. 13.1 shows an example of profile HMMs picking up a large number of FPs. This alignment is built for the six subfamilies: the cAMP-dependent protein kinase (PKA), protein kinase C (PKC), protein kinases related to PKA and PKC (RAC), G protein-coupled receptor kinase (GRK), ribosomal S6 PK and PVPK protein kinase of AGC kinase family (Srivastava et al. 2007). It is coloured on the basis of conservation of amino acid residues where pink and red columns are conserved and identical, respectively, across all families corresponding to fold signals, while green and blue are conserved and identical within a family. The columns highlighted

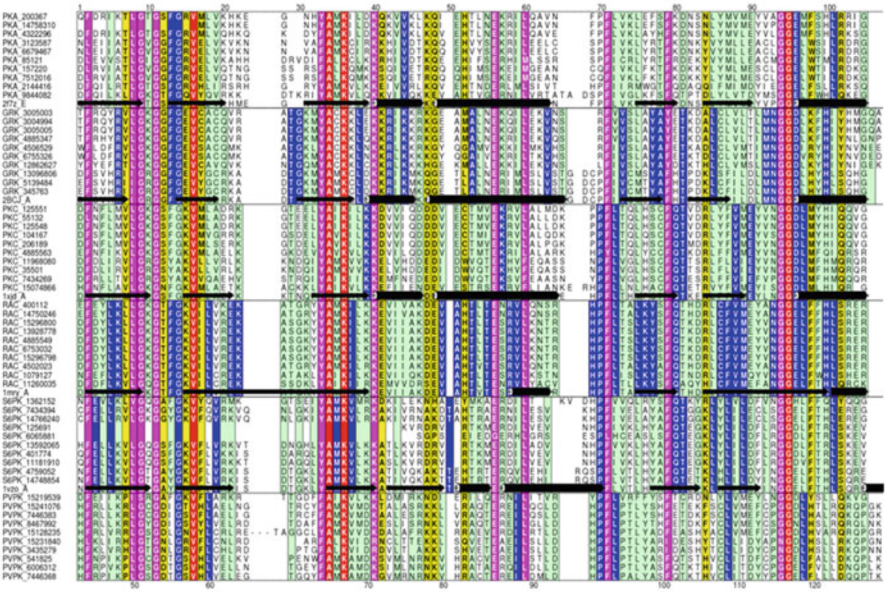


Fig. 13.1 Multiple sequence alignment showing the common fold- and function-specific signals. The alignment is only a part of the full alignment of six protein kinase families discussed in the text. It is coloured based on residue conservation: pink and red, conserved and identical across all families correspond to fold-specific signals; green and blue, conserved and identical within a family; and yellow, positions which are predicted to confer specificity for the family (Liu et al. 2004). Deleted regions are marked by dashes (- - -). The figure is only a part of the full alignment that has been published. (Srivastava et al. 2007)

in yellow signify positions which are predicted to confer specificity for the family. A profile HMM built using such an alignment will pick up sequences from other subfamilies due to shared fold- and function-specific signals.

13.5 Sequence-Modified Profile HMM Comparison Methods

A profile HMM needs to be modified in order to reduce the fold-specific signals and maximize the function-specific signals. A profile HMM modified in such a way should be able to differentiate protein sequences based on their function even when they share conserved common fold. Methods utilizing pre-classified data that make use of positive as well as negative training sequences to refine transition and emission probabilities using Viterbi algorithm (Mamitsuka 1996; Wistrand and Sonnhammer 2004) have been used earlier. In addition, methods based on positional entropy (Hannenhalli and Russell 2000) and support vector machines (SVMs) have been developed (Jaakkola et al. 2000; Karchin et al. 2002) to discriminate between function- and fold-specific signals. A similar method, HMM-ModE (Srivastava et al. 2007; Sinha and Lynn 2014), was developed that uses negative training sequences to modify the emission probabilities in the profile HMM in order to improve the specificity of prediction.

13.5.1 HMM-ModE

HMM-ModE (Srivastava et al. 2007; Sinha and Lynn 2014) creates family-specific profile HMMs by optimizing the discrimination threshold based on the mode of average Matthews correlation coefficient (MCC) distribution from tenfold cross validation. It also modifies the emission probabilities using negative training sequences. Modification of emission probabilities provides increased discrimination by identifying the differentiating alignment positions from the profile-profile alignment of positive and negative sequences using relative entropy (RE). The RE is calculated using the probability distributions of positive (p) and negative (q) sets for a particular position ‘i’ through the following equations:

$$\begin{aligned}
&\mathrm{RE}_i = \sum p_{i,x} \log \frac{p_{i,x}}{q_{i,x}}\\
&\mathrm{RE}_{i\mathrm{Neg}} = \sum\nolimits_{x=1..20} p_{i,x} \log \frac{p_{i,x}}{q_{i,x}}\\
&\mathrm{RE}_{i\mathrm{Null}} = \sum\nolimits_{x=1..20} p_{i,x} \log \frac{p_{i,x}}{P_{x\mathrm{Null}}}
\end{aligned}$$

where

$p_{i,x}$ and $q_{i,x}$ are the probabilities of the amino acid ‘x’ at a position ‘i’ in the positive and negative sets of sequences, respectively.

The log-odds score is then calculated as

$$S = \sum_{i=1..n} \begin{cases} \sum_{x=1..20} \log \frac{p_{i,x}}{q_{i,x}}, & \mathrm{RE}_{i\mathrm{Neg}} > \mathrm{RE}_{i\mathrm{Null}} \\ \sum_{x=1..20} \log \frac{p_{i,x}}{P_{x\mathrm{Null}}}, & \text{otherwise} \end{cases}$$

This score is calculated from the emission probabilities of the model, and a heuristic method that modifies the emission probabilities of the model was used to implement this score.

The modified profile HMMs in terms of their emission probabilities represent the method HMM-ModE. In this protocol, only the sequences selected as false positives by the subfamily HMM are used to modify model parameters and optimize the discrimination threshold; therefore, the training of the model is much faster. The method provides a significant improvement over all the other existing methods for differentiation and classification of function- and fold-specific signals. In general, HMM-ModE protocol is a stepwise procedure to construct such modified profiles and uses the following steps to prepare modified profiles:

- *Clustering of pre-classified datasets:* The protocol uses Markov chain clustering (MCL) (Dongen 2000) for clustering of sequences. It is a fast and scalable unsupervised clustering algorithm for graphs based on simulation of flow in graphs. In order to cluster sequences using MCL, BLAST is used for performing all-against-all sequence comparisons to identify highly similar sequences within a given family of sequences. Once the sequences are scored based on similarity, they are being clustered using MCL.
- *Generation of TP and FP profile:* Each cluster of sequences is aligned using MUSCLE, and a profile HMM is built using 'hmmbuild' (HMMER). These are known as true positive (TP) profiles which are scanned across all the training sequences using 'hmmsearch' (HMMER) to identify the FP sequences. These FP sequences are then aligned in similar manner using MUSCLE to generate FP profile HMMs.
- *Modification of emission probability:* The emission probability of the TP profile is modified by identifying the discriminating alignment positions between the TP and FP alignment using relative entropy as described earlier (Srivastava et al. 2007; Sinha and Lynn 2014).
- *Tenfold cross validation:* Tenfold cross validation is performed to estimate the accuracy of the modified profiles to provide a discrimination threshold that helps to separate the TPs from FPs efficiently as compared to the default threshold (zero) of the profile HMM built using HMMER. These thresholds can be used with the profiles using the '-T' option of 'hmmsearch' to search against the desired sequence database.

13.5.1.1 Limitations of HMM-ModE

HMM-ModE profiles provide significant improvement in specificity of prediction with minimal loss in sensitivity for the classification of sequences based on their function as compared to methods like PSI-BLAST or HMMER which are highly sensitive but much less specific for the identification of protein function. However, even BLAST, PSI-BLAST, HMMER or HMM-ModE fails to detect remote homologs where the sequence identity is very low (less than 20%). More sensitive methods based on HMM-HMM comparisons are used in these cases.

13.6 HMM-HMM Comparison Methods

The HMM-HMM comparison methods generalize the protein sequence alignment with a profile HMM to the case of pairwise alignment of two profile HMMs to detect distant homologs. The HMM-HMM comparison increases the sensitivity of prediction. One such method is 'HHsearch'.

13.6.1 HHsearch

HHsearch is a powerful tool to find remote homology by performing a HMM-HMM comparison and by incorporating the information from protein's secondary structure. Previously it was shown that involvement of protein secondary structure improves homology detection (Kabsch and Sander 1983; Hargbo and Elofsson 1999; Kelley et al. 2000; Kawabata and Nishikawa 2000). The method has been benchmarked on a dataset which is below the twilight zone (20% sequence identity) with other methods of homology detection like BLAST (Altschul et al. 1990), PSI-BLAST (Altschul et al. 1997) and HMMER (Eddy 2009). HHsearch outperforms all these methods (Söding 2005). When compared to similar profile-based method HMMER, the standard Viterbi algorithm used by HMMER is replaced by more accurate maximum accuracy (MAC) algorithm in HHsearch.

13.6.1.1 Case Study to Identify Putative Z-Ring-Associated Cell Division Proteins in *Helicobacter pylori* (*H. pylori*) Using HHsearch

In this case study, HHsearch was used to identify unknown candidate proteins involved in the process of cell division in *H. pylori*. The known set of these cell division proteins include FtsZ, FtsA, FtsW, FtsK, MurD, MinD and MinC, while some of the unknown ones are ZapA and ZapB. Traditional methods (like BLAST, PSI-BLAST, HMMER) scored only the already known cell division proteins but do not score any unknown candidate protein. Therefore, HHsearch which is a highly sensitive method was used to predict candidate proteins for ZapA and ZapB. As expected, some high-confidence hits were observed for the unknown proteins as listed in Table 13.1. One of the predicted homologs for ZapA, P64659, was tested experimentally using techniques like genetic complementation, biochemical analysis and immuno-colocalization (Kamran et al. 2016). The take-home message from this case study is to use sensitive methods based on sequence-profile comparisons when there is a low sequence similarity. These methods though have a lower specificity as they may score many false positives for the query protein. This problem was handled later on by the development of HMM-based comparison methods like HMM-ModE which aims to improve specificity of prediction without any significant loss in the sensitivity. In cases where all of these methods fail to detect any homology between the proteins of unknown function and well-characterized sequences, methods like HHsearch could serve as a reasonable alternative (Kamran et al. 2016).

13.7 Web-Based Tool for Homology-Based Sequence Analysis: ANNOTATOR

In addition to these very basic methods, there exist many other concepts in protein sequence analysis and function prediction like the presence of globular and non-globular segments typically in all the protein sequences (Eisenhaber et al. 2004; Eisenhaber and Eisenhaber 2007). In simple terms, globular segments have balanced amino acid composition, while the non-globular segments tend to have

Table 13.1 Various tools used to predict distant homologs of cell division proteins *in H. pylori*

Protein name/ UniProt ID	Presence in *H. pylori*/ Essentiality	BLAST	HMMER/ HMM-ModE	HHsearch	Probability of prediction (HHsearch)
FtsZ/ P56097	Y/Y	√	√	√	1
FtsA/ O25629	Y/Y	√	√	√	1
FtsW/ P56096	Y/Y	√	√	√	1
FtsK/ O25722	Y/Y	√	√	√	1
MurD/ O25236	Y/Y	√	√	√	1
MinD/ O25098	Y/N	√	√	√	1
MinC/ O25693	Y/N	√	√	√	1
ZapA/ P64659	N/N	–	–	√	0.96
ZapB/ O25147	N/N	–	–	√	0.99

HMMER/HMM-ModE and BLAST detected already known proteins in *H. pylori*; however, only HHsearch predicts the unknown proteins. Prediction made by the methods is presented as '√', while '–' represents that there is no prediction made. The last columns represent the probability of scoring the protein as a putative homolog using HHsearch

more biased composition with high number of repetitive patterns (low-complexity regions). Therefore, these two types of regions require a substantially different kind of algorithms for function prediction. Most of the methods discussed so far facilitate to annotate the globular part of the protein sequences. The non-globular segments usually require the assistance of numerous tools to predict their location in the sequence.

One of the very useful tools to perform automated sequence analysis for a query protein to annotate both kind of segments, i.e. globular and non-globular, is ANNOTATOR (Eisenhaber et al. 2016). It is an open-source tool available at http://annotator.bii.a-star.edu.sg/. The method is a web-based platform integrating various tools for protein sequence analysis using a plugin-style mechanism. The various algorithms integrated in ANNOTATOR are listed in Table 13.2.

13.8 Limitations of Homology-Based Methods for Function Prediction

As discussed in the previous sections, homology-based methods are widely used to transfer functions to novel sequences. In general, these methods fail to infer the function of a protein in case of absence of annotation of a homolog. There are still

Table 13.2 List of tools along with their description available in the ANNOTATOR system

Tools available in ANNOTATOR (Eisenhaber et al. 2016)	Description of the method
CAST (Promponas et al. 2000)	Detects low-complexity regions in a protein sequence
DisEMBL (Linding et al. 2003a)	Detects disordered or unstructured regions in a protein sequence
GlobPlot (Linding et al. 2003b)	Detects if a protein sequence is ordered or disordered
IUPred (Dosztanyi et al. 2005; Dosztányi 2018)	Detects intrinsically disordered and ordered regions in a protein sequence
SAPS (Brendel et al. 1992)	Performs statistical analysis of various properties of protein sequences
XNU (Claverie and States 1993)	Detects internal and intrinsic repeats in a protein sequence
DISOPRED (Ward et al. 2004)	Tool to predict protein disorder
SEG (Wootton 1994)	Tool to identify low-complexity regions
big-Pi (Eisenhaber et al. 1999)	Tool to predict suitable candidate for GPI-lipid anchoring
MyrPS/NMT (Maurer-Stroh and Eisenhaber 2004; Eisenhaber et al. 2004)	Tool to predict the myristoylation site in a protein sequence
PrePS/Prenylation-FT (Maurer-Stroh et al. 2003a, b; Neuberger et al. 2003)	Tool to predict the prenylation site and aim to model the substrate enzyme interaction based on refinement of the recognition motif of the eukaryotic enzyme farnesyltransferase (FT)
PeroxyPS/PTS1 (Neuberger et al. 2003)	Tool to predict whether a protein has the C-terminal peroxisomal targeting signal PTS1 or not
SIGCLEAVE (von Heijne 1986)	Tool to identify the cleavage site between a signal sequence and the mature exported protein
SignalP (Nielsen 2017)	Tool to identify the cleavage site and a signal peptide/non-signal peptide prediction based on artificial neural networks
DAS-Tmfilter (Cserzö et al. 2002; Cserzo et al. 2004)	Tool to predict the TM regions and differentiate them from non-TM regions
HMMTOP (Tusnády and Simon 2001)	Tool to implement a HMM model to predict TM protein topology
PHOBIUS (Käll et al. 2004)	Tool to predict and differentiate between the signal regions and transmembrane regions
TMHMM (Krogh et al. 2001)	Tool to predict membrane topology based on HMM model
TopPred (Claros and von Heijne 1994)	Tool to predict the location of the TM segments
TM-complexity (Wong et al. 2010, 2012)	Tool to predict the complexity of the TM regions as simple, twilight or complex
ImpCOIL (Frishman and Argos 1996)	Tool to predict the coiled-coil regions in proteins
Predator (Frishman and Argos 1997)	Tool to predict the secondary structural elements

(continued)

Table 13.2 (continued)

Tools available in ANNOTATOR (Eisenhaber et al. 2016)	Description of the method
SSCP (Eisenhaber et al. 1996)	Tool to predict the secondary structural elements, i.e. alpha helices, beta sheets and coil state
HMMER (Eddy 2009)	Tool to search HMM profiles against the HMM libraries of known domains and motifs
ModEnzA (Desai et al. 2011)	Tool to identify the enzyme class
IMPALA (Schäffer et al. 1999)	Tool to compare a query sequence against a library of position-specific scoring matrices (PSSMs)
HHpred (Soding et al. 2005)	Tool to query a HMM against a database of HMMs
HHblits (Remmert et al. 2012)	Tool to generate HMM-HMM-based alignments
HHsearch (Hargbo and Elofsson 1999; Kelley et al. 2000; Kawabata and Nishikawa 2000)	Tool to detect remote homology using HMM-HMM comparisons
PROSITE (Sigrist et al. 2002)	Tool to detect protein's domain
RPS-BLAST (Marchler-Bauer et al. 2011)	Tool to compare query against a library of PSSMs
ELM-patterns (Puntervoll et al. 2003)	Tool to predict functional sites in eukaryotic proteins
PROSITE-patterns (Sigrist et al. 2002)	Tool to search patterns from a collection of annotated protein motifs
EF-Patterns (Berezovsky et al. 2000)	Tool to predict the function of a protein based on a combination of elementary functional patterns
PROSPERO (Mott 2000)	Tool to analyse repeats within a sequence by comparing a sequence to itself, another sequence or a profile and print all local alignments
NCBI-BLAST (Altschul et al. 1990)	Tool to perform local alignments to detect sequence homology
OMA-BLAST (Altenhoff et al. 2011)	Tool to find orthologs of query protein
PSI-BLAST (Altschul et al. 1997)	Tools to identify remote homologs using iterative blast searches
CSI-BLAST (Biegert and Söding 2009)	Tool to derive context-specific amino acid similarities
GLSearch (Pearson 2000)	Tool to search a query sequence against a sequence database using an optimal algorithm that requires an entire query to match at least part of the database sequences
Prim-Seq-An (Schneider et al. 2010; Eisenhaber et al. 2016)	It runs a standard set of algorithms on a sequence of interest
Orphan-Search (Schneider et al. 2010; Eisenhaber et al. 2016)	Tool to determine whether a sequence is an orphan within a specific sequence database
Family Searcher (Schneider et al. 2010; Eisenhaber et al. 2016)	Tool to trace distant evolutionary relationships involving large protein families

(continued)

Table 13.2 (continued)

Tools available in ANNOTATOR (Eisenhaber et al. 2016)	Description of the method
Orthologue Search (Schneider et al. 2010; Eisenhaber et al. 2016)	Tool to identify the orthologs of a protein
Disan (Sirota et al. 2010)	Tool to run a set of disorder predictors to allow consensus predictions
MCL (Dongen 2000)	Tool to cluster sequences based on all-against-all BLAST searches
CD-HIT (Li and Godzik 2006)	Tool to cluster sequences by counting the number of identical words in a pair of sequences
T-Coffee (Di Tommaso et al. 2011)	Tool for multiple sequence alignment using progressive approach
MUSCLE (Edgar 2004)	Tool for multiple sequence alignment using iterative improvements to the progressive alignments
ProbCons (Do et al. 2005)	Tool for multiple sequence alignment using progressive approach using HMM formalism
MAFFT (Katoh and Standley 2013)	Tool for multiple sequence alignment using fast Fourier transforms (FFT) with residue volume and polarity

many orthologs whose function are not well defined but conserved across multiple organisms. In the eggNOG database (Jensen et al. 2008) which is a resource of orthologs to be used for functional annotation, 74% of the orthologous groups (OGs) are provided with nontrivial descriptions, whereas only 54% of the OGs are assigned to informative functional categories (Powell et al. 2014). In addition, these methods also generate other errors while assigning a function to an unknown protein. Various studies have quantified such intrinsic errors in annotation of genomes and proteomes (Brenner 1999; Devos and Valencia 2001). Since these methods are based on comparison of sequences to predict the function of a protein, there are different possible reasons in transferring incorrect function annotations to novel sequences. These errors might include mistakes in original method that identifies the homologs to be similar enough to transfer the annotation (Liu et al. 2004), errors in the annotations of original database which is being used to transfer the functional information resulting in extrapolation of wrong annotation and errors caused when the homolog (gene/protein) is lost or acquired a function due to evolutionary divergence (Ofran et al. 2005).

In addition to these errors, homology-based methods tend to investigate only the molecular functions of proteins and provide very little information about the context in which proteins operate within the cell. While assigning a function to an unknown protein, it is crucial to understand that proteins never function in an isolated manner within a cell but interact with other biomolecules. Protein networks and protein interactions constitute an important area of study to understand such behaviour of proteins. Therefore, a new class of computational methods has been gradually evolved to understand the cellular function of proteins. These methods are not based on comparison of sequence or structure but draw inferences about relationship between proteins by analysing the context in which they are found and are referred to as non-homology-based methods or context-based methods (Huynen et al. 2000; Snel et al. 2000; Ofran et al. 2005).

Undoubtedly homology-based analysis remains the core methodology of functional annotation, but the context-based methods go beyond sequence or structure comparisons. Context-based methods include all types of associations between genes and proteins of the same or different genomes providing information on functional interactions between them (Aravind 2000). Such use of contextual information in genome analysis provides a simple logic to ensure a systematic and powerful way to assign function to genes or proteins that have no sequence similarity to experimentally characterized homologs (Pellegrini et al. 1999).

Acknowledgements This work was supported by grants from Jawaharlal Nehru University and Open source drug discovery, Council of scientific and industrial research (OSDD-CSIR) project.

References

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C (2011) OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res 39:D289–D294. https://doi.org/10.1093/nar/gkq1238

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389

Berezovsky IN, Grosberg AY, Trifonov EN (2000) Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett 466:283–286

Biegert A, Söding J (2009) Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A 106:3770–3775. https://doi.org/10.1073/pnas.0810767106

Brendel V, Bucher P, Nourbakhsh IR et al (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci U S A 89:2002–2006

Brenner SE (1999) Errors in genome annotation. Trends Genet 15:132–133. https://doi.org/10.1016/S0168-9525(99)01706-0

Claros MG, von Heijne G (1994) TopPred II: an improved software for membrane protein structure predictions. Comput Appl Biosci 10:685–686

Claverie J-M, States DJ (1993) Information enhancement methods for large scale sequence analysis. Comput Chem 17:191–201. https://doi.org/10.1016/0097-8485(93)85010-A

Cserzö M, Eisenhaber F, Eisenhaber B, Simon I (2002) On filtering false positive transmembrane protein predictions. Protein Eng 15:745–752

Cserzo M, Eisenhaber F, Eisenhaber B, Simon I (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics 20:136–137

Desai DK, Nandi S, Srivastava PK, Lynn AM (2011) Mod Enz a: accurate identification of metabolic enzymes using function specific profile HMMs with optimised discrimination threshold and modified emission probabilities. Adv Bioinforma 2011:743782. https://doi.org/10.1155/2011/743782

Devos D, Valencia A (2001) Intrinsic errors in genome annotation. Trends Genet 17:429–431. https://doi.org/10.1016/S0168-9525(01)02348-4

Di Tommaso P, Moretti S, Xenarios I et al (2011) T-coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res 39:W13–W17. https://doi.org/10.1093/nar/gkr245

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S (2005) Prob cons: probabilistic consistency-based multiple sequence alignment. Genome Res 15:330–340. https://doi.org/10.1101/gr.2821705

Dosztányi Z (2018) Prediction of protein disorder based on IUPred. Protein Sci 27:331–340. https://doi.org/10.1002/pro.3334

Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433–3434. https://doi.org/10.1093/bioinformatics/bti541

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 23:205–211

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

Eisenhaber B, Eisenhaber F (2007) Posttranslational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure? Curr Protein Pept Sci 8:197–203

Eisenhaber F, Frömmel C, Argos P (1996) Prediction of secondary structural content of proteins from their amino acid composition alone. II The paradox with secondary structural class. Proteins 25:169–179. https://doi.org/10.1002/(SICI)1097-0134(199606)25:2<169::AID-PROT3>3.0.CO;2-D

Eisenhaber B, Bork P, Eisenhaber F (1999) Prediction of potential GPI-modification sites in proprotein sequences. J Mol Biol 292:741–758. https://doi.org/10.1006/jmbi.1999.3069

Eisenhaber B, Eisenhaber F, Maurer-Stroh S, Neuberger G (2004) Prediction of sequence signals for lipid post-translational modifications: insights from case studies. Proteomics 4:1614–1625. https://doi.org/10.1002/pmic.200300781

Eisenhaber B, Kuchibhatla D, Sherman W et al (2016) The recipe for protein sequence-based function prediction and its implementation in the ANNOTATOR software environment. Methods Mol Biol 1415:477–506. https://doi.org/10.1007/978-1-4939-3572-7_25

Finn RD, Bateman A, Clements J et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:222–230. https://doi.org/10.1093/nar/gkt1223

Frishman D, Argos P (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Eng 9:133–142

Frishman D, Argos P (1997) Seventy-five percent accuracy in protein secondary structure prediction. Proteins 27:329–335

Hannenhalli SS, Russell RB (2000) Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol 303:61–76. https://doi.org/10.1006/jmbi.2000.4036

Hargbo J, Elofsson A (1999) Hidden Markov models that use predicted secondary structures for fold recognition. Proteins 36:68–76

Huynen M, Snel B, Lathe W, Bork P (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 10:1204–1210. https://doi.org/10.1101/gr.10.8.1204

Jaakkola T, Diekhans M, Haussler D (2000) A discriminative framework for detecting remote protein homologies. J Comput Biol 7:95–114. https://doi.org/10.1089/10665270050081405

Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P (2008) egg NOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36 (Database issue):D250–D254 Epub 2007 Oct 16

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637. https://doi.org/10.1002/bip.360221211

Käll L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016

Kamran M, Sinha S, Dubey P et al (2016) Identification of putative Z-ring-associated proteins, involved in cell division in human pathogenic bacteria *Helicobacter pylori*. FEBS Lett 590:2158–2171. https://doi.org/10.1002/1873-3468.12230

Karchin R, Karplus K, Haussler D (2002) Classifying G-protein coupled receptors with support vector machines. Bioinformatics 18:147–159

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010

Kawabata T, Nishikawa K (2000) Protein structure comparison using the markov transition model of evolution. Proteins 41:108–122

Kelley LA, MacCallum RM, Sternberg MJ (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299:499–520. https://doi.org/10.1006/jmbi.2000.3741

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580. https://doi.org/10.1006/jmbi.2000.4315

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. https://doi.org/10.1093/bioinformatics/btl158

Linding R, Jensen LJ, Diella F et al (2003a) Protein disorder prediction: implications for structural proteomics. Structure 11:1453–1459

Linding R, Russell RB, Neduva V, Gibson TJ (2003b) GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res 31:3701–3708

Liu J, Hegyi H, Acton TB et al (2004) Automatic target selection for structural genomics on eukaryotes. Proteins 56:188. https://doi.org/10.1002/prot.20012

Mamitsuka H (1996) A learning method of hidden Markov models for sequence discrimination. J Comput Biol 3:361–373

Marchler-Bauer A, Lu S, Anderson JB et al (2011) CDD: A conserved domain database for the functional annotation of proteins. Nucleic Acids Res 39:D225–D229. https://doi.org/10.1093/nar/gkq1189

Marcotte EM, Pellegrini M, Thompson MJ et al (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402:83–86. https://doi.org/10.1038/47048

Marcotte EM, Xenarios I, van der Bliek AM, Eisenberg D (2000) Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci 97:12115–12120. https://doi.org/10.1073/pnas.220399497

Maurer-Stroh S, Eisenhaber F (2004) Myristoylation of viral and bacterial proteins. Trends Microbiol 12:178–185. https://doi.org/10.1016/j.tim.2004.02.006

Maurer-Stroh S, Washietl S, Eisenhaber F (2003a) Protein Prenyltransferases: Anchor Size, Pseudogenes and Parasites. Biol Chem 384:977–989. https://doi.org/10.1515/BC.2003.110

Maurer-Stroh S, Washietl S, Eisenhaber F (2003b) Protein prenyltransferases. Genome Biol 4:212. https://doi.org/10.1186/GB-2003-4-4-212

Mott R (2000) Accurate formula for P-values of gapped local sequence and profile alignments. J Mol Biol 300:649–659. https://doi.org/10.1006/jmbi.2000.3875

Neuberger G, Maurer-Stroh S, Eisenhaber B et al (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. J Mol Biol 328:581–592

Nielsen H (2017) Predicting secretory proteins with SignalP. In: Methods in molecular biology. Humana Press, Clifton, pp 59–73

Ofran Y, Punta M, Schneider R, Rost B (2005) Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discov Today 10:1475–1482. https://doi.org/10.1016/S1359-6446(05)03621-4

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132:185–219

Pearson WR (2013) An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinformatics 42:3.1.1–3.1.8. https://doi.org/10.1002/0471250953.bi0301s42

Pellegrini M, Marcotte EM, Thompson MJ et al (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96:4285–4288

Powell S, Forslund K, Szklarczyk D et al (2014) EggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res 42:231–239. https://doi.org/10.1093/nar/gkt1253

Promponas VJ, Enright AJ, Tsoka S et al (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. Bioinformatics 16:915–922

Puntervoll P, Linding R, Gemünd C et al (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31:3625–3630

Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9:173–175. https://doi.org/10.1038/nmeth.1818

Schäffer AA, Wolf YI, Ponting CP et al (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15:1000–1011

Schneider G, Wildpaner M, Sirota FL et al (2010) Integrated tools for biomolecular sequence-based function prediction as exemplified by the ANNOTATOR software environment. Methods Mol Biol 609:257–267. https://doi.org/10.1007/978-1-60327-241-4_15

Sigrist CJA, Cerutti L, Hulo N et al (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3:265–274

Sinha S, Lynn AM (2014) HMM-ModE: implementation, benchmarking and validation with HMMER3. BMC Res Notes 7:483. https://doi.org/10.1186/1756-0500-7-483

Sirota FL, Ooi H-S, Gattermayer T et al (2010) Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. BMC Genomics 11:S15. https://doi.org/10.1186/1471-2164-11-S1-S15

Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28:3442–3444. https://doi.org/10.1093/nar/28.18.3442

Söding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960. https://doi.org/10.1093/bioinformatics/bti125

Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244–W248. https://doi.org/10.1093/nar/gki408

Srivastava PK, Desai DK, Nandi S, Lynn AM (2007) HMM-ModE--improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. BMC Bioinformatics 8:104. https://doi.org/10.1186/1471-2105-8-104

Stein L (2001) Genome annotation: from sequence to biology. Nat Rev Genet 2:493–503. https://doi.org/10.1038/35080529

Tusnády GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849–850

van Dongen SM (2000) Graph clustering by flow simulation. PhD thesis, Utrecht University Repository

von Heijne G (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14:4683–4690

Ward JJ, McGuffin LJ, Bryson K et al (2004) The DISOPRED server for the prediction of protein disorder. Bioinformatics 20:2138–2139. https://doi.org/10.1093/bioinformatics/bth195

Wistrand M, Sonnhammer ELL (2004) Improving profile HMM discrimination by adapting transition probabilities. J Mol Biol 338:847–854. https://doi.org/10.1016/j.jmb.2004.03.023

Wong W-C, Maurer-Stroh S, Eisenhaber F (2010) More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. PLoS Comput Biol 6:e1000867. https://doi.org/10.1371/journal.pcbi.1000867

Wong W-C, Maurer-Stroh S, Schneider G, Eisenhaber F (2012) Transmembrane helix: simple or complex. Nucleic Acids Res 40:W370–W375. https://doi.org/10.1093/nar/gks379

Wootton JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 18:269–285. https://doi.org/10.1016/0097-8485(94)85023-2

Yoon B-J (2009) Hidden Markov models and their applications in biological sequence analysis. Curr Genomics 10:402–415

Drug Discovery: An In Silico Approach

14

Sukriti Goyal, Salma Jamal, Abhinav Grover, and Asheesh Shanker

14.1 Introduction

14.1.1 Traditional Drug Discovery and Its Disadvantages

The course of development of one new medicine from an original idea to the time it is available for treating patients is a complex and tedious task that can take up to 15 years. An average amount of expenditure required for research and development (R&D) of a drug with successful outcome lies within the range of $800 to $1000 million, including the cost of failures. For instance, only one potent lead is approved out of a batch of 5000–10,000 that enter the R&D pipeline. The preliminary research, usually taking place in academic world, produces information to generate a hypothesis that either the blocking or activation of the target protein or a particular pathway will produce a remedial effect in the diseased condition. The result of this step is the selection of a drug target that might entail more validation before entering the lead discovery phase so as to rationalize the effort of drug discovery. Lead discovery is an exhaustive search with an aim to identify small molecules or biological therapeutic, usually termed as development candidate, having drug-like properties. It proceeds to preclinical phase followed by clinical phase if successful and eventually be a marketed medicine (Fig. 14.1).

S. Goyal (✉) · S. Jamal
Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan, India

A. Grover
School of Biotechnology, Jawaharlal Nehru University, New Delhi, India

A. Shanker
Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan, India

Department of Bioinformatics, Central University of South Bihar, Gaya, Bihar, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_14

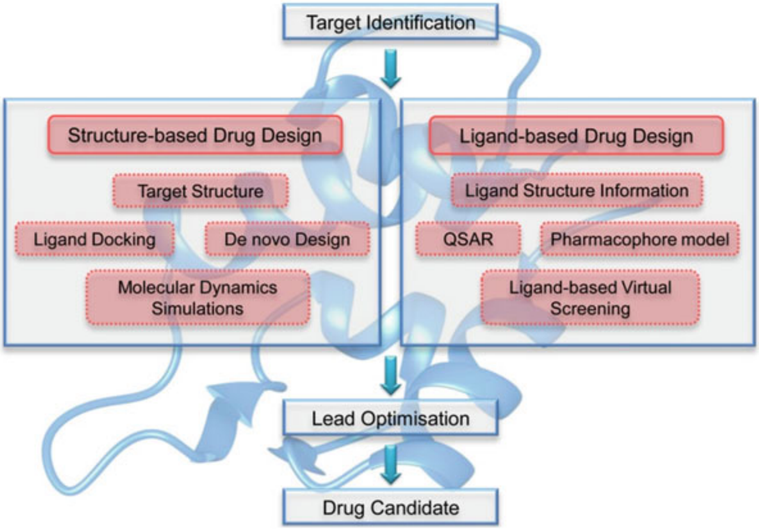


Fig. 14.1 Flowchart depicting the basic steps involved in process of drug discovery

Traditionally, the process of drug discovery is time-consuming involving synthesis of compounds, by a step-by-step protocol against a series of *in vivo* biological screens followed by investigation of the potential lead compounds for their pharmacokinetic properties (Hughes et al. 2011). For the last two decades, researchers employed in the drug discovery domain at various places such as universities, biotech, and pharmaceutical companies have employed a reductionist target-based approach, which focuses on identifying and validating small-molecule compounds possessing specific activity against a specific drug target (usually a protein). This entire procedure can be explained in three main phases: drug target identification and validation, lead identification, and optimization. The repercussion of this complete procedure is high attrition rates with failures accredited to adverse effects in humans (10%), animal toxicity (11%), lack of efficacy (30%), poor pharmacokinetics (39%), and numerous commercial and various factors. A plethora of highly effective technologies including high-throughput screening, *in vitro* and *in silico* ADMET screening, de novo design, combinatorial chemistry, and virtual screening along with bioinformatics, genomics, and proteomics in addition to structure-based drug design have revolutionized the process of drug discovery.

14.1.2 Significance and Advantages of In Silico Methodology

Fortune magazine, in an issue published in 1981 on October 5, featured a cover story titled the "Next Industrial Revolution: Designing Drugs by Computer at Merck" (Van Drie 2007). This article has been accredited as "the commencement of extraordinary interest in the potential for computer-aided drug design (CADD)" which in addition to requiring minimum ligand design or prior information provides an advantage of technologies capable to screen large libraries more efficiently. With preexisting small-molecule libraries representing the standard compound catalogue of pharmaceutical companies, novel drug candidate designing is a significant evolving element of drug discovery. The integration of *in silico* CADD into *in vitro* experimental methodology has led to a new approach known as "smart" drug design (Drews 2000; Kapetanovic 2008; Owen 2002). CADD possesses the ability of escalating the hit rate of new drug molecules owing to a more targeted search unlike traditional drug discovery methodology. Furthermore, while it explains the molecular basis of remedial or inhibitory activity, it also predicts potent compound derivatives that might enhance activity. The expenditure of drug discovery can be reduced significantly by using computational tools (Kotz 2013; Van Drie 2007).

CADD usually serves three main purposes in any drug discovery campaign: the first is filtering huge chemical repositories into smaller sets of potent lead molecules for *in vitro* validation; the second is optimization of those lead compounds by improving its affinity or by optimizing the drug metabolism and pharmacokinetic properties; and the last is designing of new therapeutics, by either building up new compounds one chemical entity at a time or by amalgamating different moieties into new chemotypes (Jorgensen 2004; Terstappen and Reggiani 2001).

14.1.3 Methods of Computer-Aided Drug Discovery

Based on the amount of data present for the chosen protein drug target, drug discovery protocols can be classified into two main groups. The first category, structure-based drug discovery (SBDD) approach, depends on the information of the 3D model of protein to compute interaction energies for lead molecules, whereas ligand-based drug discovery (LBDD), the second approach, utilizes the information of biologically identified active and inactive compounds through common moiety search or by generation of quantitative structure-activity relationship (QSAR) models with predictive abilities (Kalyaanamoorthy and Chen 2011). SBDD is usually favored when high-resolution structural information of the drug target is present, whereas LBDD is usually favored when structural information is not present or is trivial (Jorgensen 2004; Terstappen and Reggiani 2001). Computing plays a very useful role in both methods.

14.1.4 Structure-Based Computer-Aided Drug Discovery

SBDD has emerged as a new approach which involves comprehending the fundamental concept of molecular recognition in ligand-bound protein structures. The information of the experimentally obtained 3D crystal structures or a model prepared based on the crystal structure of the target protein's close homolog using computational modeling approaches, preferably in complex with a ligand, can be used for retrieval and development of potent lead candidates. This complex deciphers the mode of binding along with the conformational state of the compound under investigation and exhibits the imperative features responsible for its binding potential. The observed data is exploited for producing novel ideas for enhancing an already existing ligand or for development of novel alternative bonding chemical moieties. Computational methods along with molecular graphics are implemented to aid this step of hypothesis generation. The essential characteristics of the protein binding pocket can be rendered into queries that can then be utilized either for high-throughput virtual screening of huge chemical repositories or for designing new ligands de novo. Later these ligands are enhanced toward increased affinity and better selectivity. The selectivity aspect is of vital significance in understanding and controlling the pharmacological profile of a ligand.

14.1.5 Ligand-Based Computer-Aided Drug Discovery

In case information regarding experimental structure of the target of interest is unavailable, a different protocol called LBDD, based on the validated inhibitors of a drug target, can be used. It involves methodologies like quantitative structure-activity relationships (QSAR), molecular similarity approaches and pharmacophore modeling (Blake and Laird 2003; Lahana 1999). Information about the molecular fingerprints of validated compounds is exploited to virtually screen large compound databases to find molecules with analogous fingerprints (Gillet et al. 1999). Pharmacophore modeling approach can be employed to obtain common structural characteristics of ligands that can be utilized to virtually screen chemical repositories to find hits with those characteristics (Taylor et al. 2002). Since pharmacophore models only specify the essential characteristics responsible for inhibitory properties of a ligand, QSAR models are better in elucidating the relationship between the various characteristics of drug and the biological activity and thus predicting the activity of the ligand (Doman et al. 2002).

14.2 Structure-Based Drug Discovery

14.2.1 Molecular Modeling of Protein Structures

The first essential step after identification of the drug target is obtaining precise structural information. The four primary methodologies utilized for this purpose include nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, comparative modeling in case of presence of some initial information and ab initio modeling in case of complete absence of any structural information. The first two methods are experimental and offer several advantages over modeling methods. Structures acquired by application of experimental methods are of high resolution and possess ordered water molecules which are helpful when designing a lead compound. The acceleration in the rate of determination of target structures can be owed to structural genomics. However, numerous virtual screening campaigns with successful results have been described using comparative modeling approach of target proteins in absence of experimentally resolved structure (Becker et al. 2006; Budzik et al. 2010; Warner et al. 2006).

Comparative modeling is an approach applied for prediction of target structure that works on the basic principle that proteins having homologous sequences possess similar structures. Homology modeling, a particular form of comparative modeling, can be used when evolutionary origin of the template and target protein is same. The complete process of comparative modeling comprises of three main steps. The beginning step involves identification of homologous proteins with known crystal structures as templates and alignment of the sequences of target protein with template protein. The next step comprises constructing the coordinate file of the target protein by duplicating 3D coordinates for confidently aligned regions and modeling coordinates for missing atoms. The last step is model refinement and evaluation. The comparative modeling process can be automated by various computer programs such as MODELLER (Martí-Renom et al. 2000) and web servers such as PSIPRED (Buchan et al. 2010). The steps involved in homology modeling are as follows:

14.2.2 Template Identification and Alignment

In this step, a Basic Local Alignment Sequence Tool (BLAST; Altschul et al. 1990) search is performed using a Protein Data Bank (PDB) database against the query (protein sequence of the drug target) for recognition of template structures with high sequence similarity (Altschul et al. 1990). In case PDB-BLAST search fails to result in any hits, additional sophisticated fold recognition methods available can be utilized (Kelley and Sternberg 2009; Söding and Remmert 2011). After obtaining structures of template protein, sequence alignment for template and target protein sequences is performed using tools such as ClustalW (Thompson et al. 1994). For a group of structurally related proteins, conserved regions in the sequence alignment are recognized and utilized for building the homology model. Repeated production and assessment of numerous homology models from various high-scoring sequence alignments may upgrade the standard of homology model (Chivian and Baker 2006; Misura et al. 2006). The key to successful homology modeling is template selection. Meticulous attention should be given to the resolution of the template structure along with alignment length and sequence identity of query with hit obtained.

14.2.3 Model Building

The primary cause of occurrence of gaps or insertions in the sequence alignment is that they lie mainly outside the secondary structure elements, thus leading to chain breaks. The anchor amino acids or the N- or C-terminal amino acids of the target protein sequence on both sides of the missing regions need to be connected in order to model the same. Based on the information present for the missing region, two methods, namely, *de novo* and knowledge-based methods, are known. Knowledge-based technique utilizes structured regions from crystal structures having roughly the same anchors as present in target structure which is then applied to the target model. On the other hand, *de novo* methods produce numerous loop conformations, quality of which is then judged using various energy functions (Hillisch et al. 2004). The next action in this step involves prediction of conformations of side chain. Rotamer library, a cluster of 3D coordinates of conformations of side chain, is generally employed during all side-chain prediction protocols (Krivov et al. 2009). Side-chain conformation sampling employs methodology, for instance, dead-end elimination (Desmet et al. 1992) applied in SCRWL (Bower et al. 1997; Dunbrack and Karplus 1993, 1994) and Monte Carlo searches (Rohl et al. 2004).

14.2.4 Model Refinement and Evaluation

Steps, namely, addition of correct bond geometries and removal of unfavorable interactions brought in by the initial modeling process, are performed to refine developed atomic models. This step involves energetically minimizing the generated atomic models by employing techniques such as genetic algorithms (Xiang 2006), minimization using Monte Carlo Metropolis (Misura and Baker 2005), or molecular dynamics simulations (Raval et al. 2012).

Model evaluation is performed by comparing structural characteristics observed in developed atomic models with experimentally determined structures. Many research groups contribute in a worldwide experiment named, Critical Assessment of Structure Prediction (CASP; Cozzetto et al. 2009), for an objective assessment of their respective structure prediction methods. In CASP assessment, a generated atomic model and the corresponding experimental structure are compared to evaluate if they are statistically similar and thus implemented for numerically assessing and ranking the developed models. Alignment accuracy (AL0 score), global distance test-total scores (GDTTS) and full model root mean square deviation are some examples of evaluation methods used in CASP (Cozzetto et al. 2009).

14.3 Molecular Docking

An eminent prerequisite for drug activity is protein-ligand interaction. Ligand-bound experimentally determined crystal models of the protein target or its analogous protein with the natural substrate or non-natural substrate is often the source of possible binding cavities for small molecules. However, if the information regarding the binding cavity or catalytic site of the protein is unknown, new binding sites can be identified using various computational platforms including Q-SiteFinder, SURFNET and POCKET (Henrich et al. 2010; Laurie et al. 2006).

The degree of flexibility considered for compound and protein during the docking process classifies these methods into rigid-body docking and flexible docking (Dias et al. 2008; Halperin et al. 2002). The rigid-body docking method takes into account only static geometric/physiochemical complementarities during complex generation while disregarding flexibility as well as induced-fit binding models (Halperin et al. 2002). In conformational selection paradigm, further advanced algorithms also account various potential conformations of either compound or receptor or both simultaneously (Changeux and Edelstein 2011). This docking methodology is usually chosen when the target is to be screened against a huge chemical repository during an initial screening. After filtering out potential hits using the initial screening, flexible docking methods are applied for optimization and refinement of binding conformations resulting from initial rigid docking process. Flexible docking approach is now being used more frequently owing to the advancement of computational resources and efficiency. Systematic arrangement of conformations, Monte Carlo search algorithms with Metropolis criterion (MCM), molecular dynamics simulations and genetic algorithms are examples of some popular approaches used for flexible docking.

Structure-based high-throughput virtual screening (HTVS), an *in silico* method, is employed for finding potent lead compounds from a large chemical library using docking approach and it thus depends on the match of crystal structure of the compound with the binding cavity of target protein. HTVS provides an advantage over traditional HTS by selecting ligands with potential ability to bind to a particular binding site instead of experimentally asserting its general ability to hinder the target protein's function. HTVS is a time-efficient approach that screens a huge chemical repository in finite time by limiting the conformational space of the target and chemical compound and by rapid computing of binding energy through simplified approximation. Although these approximations introduce some inaccuracies which in turn leads to false-positive hits, refined docking with more specialized protocols including iterative docking and clustering technique of the top resulting molecules is a good and time-efficient approach. The main steps involved in this technique include preparation of the target and chemical repository for docking, identifying favorable binding conformation for each molecule and then ranking the ligand-bound protein structures. The most probable lead compounds are then evaluated with additional refined scoring functions. The goals of applying this extra-precise protocol for refining the initial docking poses include improving the judgment how well the compound will interact with drug target, precise prediction of docked structures poses and precise prediction of binding affinity. HTVS has been employed successfully for identification of novel and potent lead compounds in numerous drug designing studies (Becker et al. 2006; Dhanjal et al. 2014a, b; Grover et al. 2014; Lu et al. 2006; Ruiz et al. 2008; Tyagi et al. 2013, 2015; Zhao et al. 2006).

14.4 ADMET Properties and Its Prediction

A satisfactory profile describing absorption, distribution, metabolism, excretion and toxicity (ADMET) properties plays an important role in deciding the success of the drug. Even though a large range of *in vitro* ADMET screens are present, the capability to foretell few of these characteristics *in silico* is helpful. An acceptable balance of safety, toxicity, potency and appropriate pharmacokinetics recognizes a successful medicine. Numerous studies (Kennedy 1997) carried out during the late 1990s reported that the main causes of late-stage failures of potent lead molecules in the process of drug discovery were poor pharmacokinetics and toxicity. *In silico* ADMET tools are an effort toward addressing this problem. *In silico* tools provide many advantages including speeding the process of drug discovery by screening out more potent molecules in small amount of time, selecting an appropriate balance of numerous properties from huge virtual repositories for chemical synthesis by screening these models on virtual molecules and gaining a correlation between the structural profile and physiochemical properties of a drug and its corresponding ADMET properties. These *in silico* tools differ in their prediction accuracy and throughput and also in the array of statistical methodology and descriptors. For instance, descriptors can be either derived from properties of simple whole molecule (e.g., logP/D, size, hydrogen-bond donors and acceptors, etc.) or from quantum theory-based semi-empirical methods (van de Waterbeemd 2002; Van De Waterbeemd and Gifford 2003).

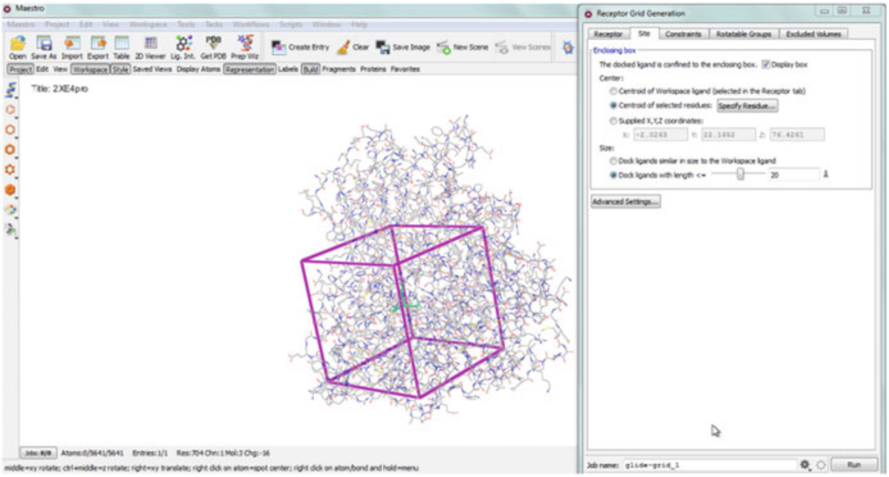
14.5 A Case Study Using Schrödinger Glide Module

Schrödinger is an integrated software platform comprising of many modules for drug designing including those for docking, QSAR, and pharmacophore development. Glide module of this platform provides tools for HTVS of huge chemical repository and extremely accurate binding conformation predictions.

The purpose of this study is to search for small natural compounds with potent antileishmanial properties targeting Oligopeptidase B (OPB) and to comprehend the inhibitory action of the natural compounds against OPB. The protocol employs screening a large repository of compounds with natural origin against OPB.

14.5.1 Preparation of Protein and Ligand

The experimental structure of target protein was extracted from RCSB Protein Data Bank [PDB: 2XE4] (Rose et al. 2011). Accelrys ViewerLite Version 5.0 (Lite 1998) was utilized for preprocessing the protein structure. The protein structure was again processed followed by optimization by employing Maestro's protein preparation wizard (Schrödinger 2008b, 2009) before docking analysis. A natural compound



Atoms:0/5641/5641 Entries:1/1 Res:704 Chn:1 Mol:3 Chg:-16

middle=xy rotate; ctrl+middle=z rotate; right=xy translate; right click on atom=spot center; right click on atom/bond and hold=menu

Receptor Grid Generation

Receptor | Site | Constraints | Rotatable Groups | Excluded Volumes

Enclosing box

The docked ligand is confined to the enclosing box. ☑ Display box

Center:

○ Centroid of Workspace ligand (selected in the Receptor tab)

● Centroid of selected residues: Specify Residue...

○ Supplied X,Y,Z coordinates:

X: -2.0243 Y: 22.1652 Z: 76.4241

Size:

○ Dock ligands similar in size to the Workspace ligand

● Dock ligands with length <= 20 Å

Advanced Settings...

Job name: glide-grid_1

Fig. 14.2a Generation of grid around receptor binding site

library of about 0.2 million natural compounds (Irwin and Shoichet 2005) was obtained and prepared using Schrödinger's LigPrep module (Ligprep 2009).

14.5.2 Receptor Grid Generation

A receptor grid of size 20 × 20 × 20 Å was generated around the binding site, comprising of the catalytic triad of OPB (Ser 577, Asp 662, and His 697) along with Tyr 499 and Glu 621 which are known to be conserved throughout the OPB family, by employing receptor grid generation utility of Glide module (Fig. 14.2a) (Friesner et al. 2004; Schrödinger 2008a, b).

14.5.3 Docking and Scoring

Glide high-throughput virtual screening and Glide extra precision (XP) (Friesner et al. 2004; Halgren et al. 2004) were applied for screening of the natural compound repository against OPB. Compounds possessing HTVS docking scores more than (in magnitude) −7 kcal/mol (in this case 423) were screened further using a refined XP docking protocol (Fig. 14.2b). The top two scoring ligands with a score of −13.183 kcal/mol and − 10.3 kcal/mol resulting from XP docking were chosen for evaluation of their ADMET properties.

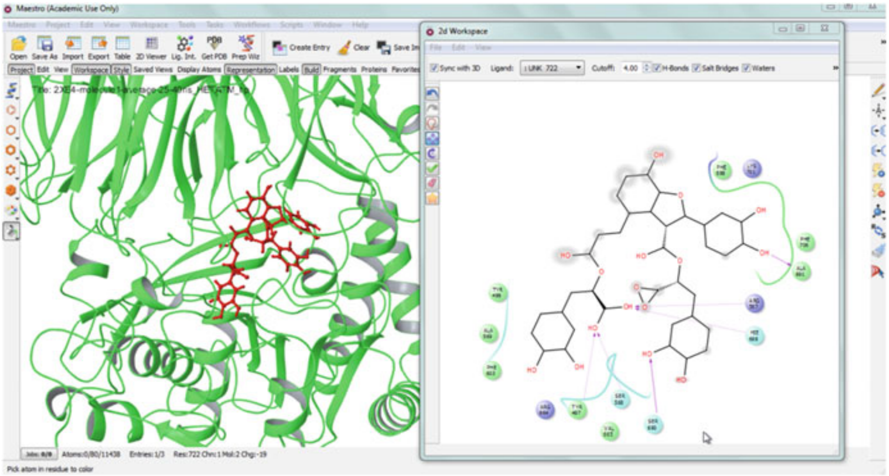

Maestro (Academic Use Only)
2d Workspace
Title: 2XE4-molecule1-average-25-40ns_HETATM_hp
☑ Sync with 3D Ligand: UNK 722 Cutoff: 4.00 ☑ H-Bonds ☑ Salt Bridges ☑ Waters
Atoms:0/80/11438 Entries:1/3 Res:722 Chn:1 Mol:2 Chg:-19
Pick atom in residue to color

Fig. 14.2b Hydrogen-bond formation and hydrophobic interactions between target enzyme and ligand

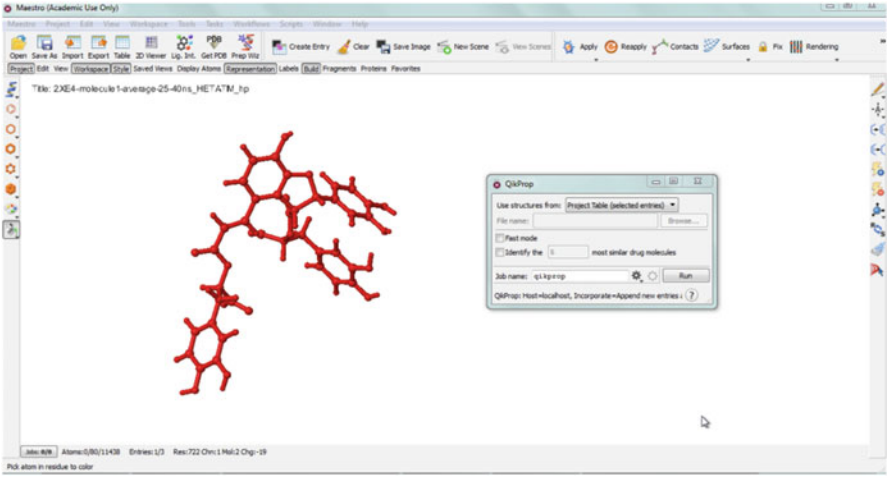

Maestro (Academic Use Only)
Title: 2XE4-molecule1-average-25-40ns_HETATM_hp
QikProp
Use structures from: Project Table (selected entries)
File name:
Browse...
☐ Fast mode
☐ Identify the 5 most similar drug molecules
Job name: qikprop
Run
QikProp: Host=localhost, Incorporate=Append new entries
Atoms:0/80/11438 Entries:1/3 Res:722 Chn:1 Mol:2 Chg:-19
Pick atom in residue to color

Fig. 14.2c Calculation of ADMET properties of ligand

14.5.4 Calculation of ADMET Properties

ADMET properties of the resulting compounds were predicted using Schrödinger's QikProp module (Fig. 14.2c) (Ioakimidis et al. 2008; Schrödinger 2008b). An online web server admetSAR was used for prediction of toxicity (Cheng et al. 2012). Forty-nine descriptors were calculated and the results were compared with 95% known drugs. The two resulting compounds exhibited satisfactory ADMET properties with good absorption power. They were observed to be non-carcinogenic with toxic properties toward *Tetrahymena pyriformis*, fish and honeybee. For detailed results please refer to the original paper (Goyal et al. 2014).

14.6 Ligand-Based Drug Discovery

14.6.1 Quantitative Structure-Activity Relationship

In case the crystal structure of the drug target is unresolved, the approach for drug development can instead be based on the knowledge available for known inhibitors of the corresponding drug target. Information about compounds that are known to bind and affect their target receptor in a three-dimensional manner forms the basis of 3D techniques in ligand-based virtual screening. One such ligand-based 3D procedure is quantitative structure-activity relationship (QSAR) that quantifies or generates a relationship between properties of chemical profile of the compound with its biological or chemical activity by application of mathematical models (Put et al. 2003; Tropsha 2003). If the biological activity of a group of compounds against the corresponding target protein can be determined, a mathematical model describing this relationship can be developed. In addition to encoding only the crucial characteristics of a biologically active ligand as in a pharmacophore model, the QSAR model also delineates the effect of a particular property of the molecule on its inhibitory activity. The set of active molecules chosen for QSAR is required to cover an extensive activity range (three orders of magnitude is the minimum range) against that particular target for the generation of a robust and statistically sound QSAR model. Quality of the dataset, its activity range, and the specificity of its activity decide the quality of the developed model. Since the aim of QSAR is quantifying the structure and activity of a compound, quantifying the structure of compound poses a crucial problem as representation of the structure by a numerical value is not viable. To address this problem, a collection of features, also called descriptors, are assessed using the structure, and these are utilized to enumerate the same. A QSAR model is built to illustrate the relationship between independent variable (evaluated descriptors) and dependent variable (inhibitory activity) (Esposito et al. 2004; Svetnik et al. 2003). Post QSAR model building and its validation, the biological or inhibitory activity of new molecules can be predicted using their structural descriptors. This QSAR model can also be utilized to virtually screen a large chemical database to identify potent lead compounds. Owing to the wide range of biological, chemical, or physical descriptors, QSAR model can also be employed in industries other than drug design (Du et al. 2008; Put et al. 2004), such as toxicology (Bradbury 1995), food chemistry (Martinez-Mayorga and Medina-Franco 2009), and other fields.

The first step to develop a QSAR model comprises of collection of compounds and their inhibitory activities. It is then followed by calculation and selection of descriptors prior to choosing a mathematical modeling method which together with activity values is used for the generation of QSAR models. Protocols of internal as well as external validation are performed to test the developed models. Validated

QSAR models are employed for the activity prediction of novel compounds. Inhibitory or biological activity is generally quantified in terms of half maximal inhibitory concentration (IC50s) (de Melo 2010), half maximal effective concentration (EC50s; Zhou et al. 2010), and Ki values (inhibition constant; Karolidis et al. 2010). However, depending on the property of the ligand to be predicted, more activity indexes can be employed for the quantification of inhibitory activity during QSAR model development. The descriptors quantifying ligand structures need to be calculated and verified prior to building a QSAR model.

After the dependent variable (inhibitory activity) and independent variables (descriptors) are specified for the dataset of compounds, a variable selection method and a model building method are chosen. To eliminate redundancy, removing all invariable descriptors and building the QSAR model using the remaining descriptors with unique values is a rational approach (Karelson 2000). For instance, in case two descriptors correspond to an analogous biological or chemical factor, either of them should be eliminated. Stepwise, principle component analysis (Xue et al. 2000), simulated annealing, genetic algorithms (Chen et al. 1998; Rogers and Hopfinger 1994), and artificial neural networks (Wikel and Dow 1993) are some selection methods employed for descriptor selection. Conventional statistical protocols such as principle least square, multiple linear regression, and k-nearest neighbor (Itskowitz and Tropsha 2005) are employed for generation of a linear QSAR model. The key difference between frequently applied QSAR algorithms exists in their methods of descriptor generation. CoMFA (Cramer et al. 1988b), CoMSIA (Klebe et al. 1994), CoMMA (Silverman and Platt 1996), and HypoGen (Kurogi and Guner 2001) are some examples of QSAR algorithms that utilize comparable linear statistical models for exploring the structure-activity relationship. In the first two methods (CoMFA and CoMSIA), the aligned compounds are positioned into a grid. Calculation of descriptors is carried out by the contact of the compound and the probe, positioned at every intersecting point of the grid. The difference between CoMFA and CoMSIA lies in the application of distinct probes along with functions used to calculate interactions. Probes signifying only steric and electrostatic contacts are utilized in CoMFA, while probes signifying hydrophobic and hydrogen bonds are also selected in CoMSIA. CoMSIA employs a Gaussian-type function for calculation of contacts between probe and molecule. Using this smooth function provides an advantage as the result value is more rational in comparison with the function employed in CoMFA, and thus specifying a threshold value for removal of invariable descriptors is not required. In CoMMA, the descriptors are produced by computation of the spatial moments of the compounds. The single descriptor applied in HypoGen model development procedure is fit value which explains robustness of the alignment of a compound with a pharmacophore model (explained in the next section).

Validation of the developed QSAR model is imperative before it can be employed for prediction of inhibitory activity. Protocols like internal validation (“leave-one-out- LOO method” or “leave-n-out” methods) (Cramer et al. 1988a) and external validation are examples of some popular protocols utilized for validation of QSAR model (Verma et al. 2010). In internal validation, either one (in case of LOO) or more (in case of leave-n-out) training set compounds are eliminated, and the model is reconstructed with reduced training set that is utilized for prediction of inhibitory activity of the excluded compounds. This process is continued till all compounds belonging to the training set have been eliminated and their activity is predicted. The accuracy of predicted activity determines the robustness of the QSAR model (Cramer et al. 1988a). Unlike internal validation, which uses training set compounds for validation of the model, this second validation protocol, known as external validation, is an extensively employed protocol which tests the potential of the QSAR model with compounds not present in the training set (Consonni et al. 2009). In order to ensure reliability of the developed QSAR model, both internal and external validation protocols are executed. After these validation procedures, if the model satisfies these tests, it can be employed for the prediction of inhibitory activity of novel compounds.

14.7 Pharmacophore Modeling

Ehrlich was the first to define the term “pharmacophore” as “a molecular framework that carries the essential features responsible for a drug’s biological activity” (Ehrlich 1909). It can be understood from the description that a pharmacophore model describes the imperative characteristics that any compound needs to possess in order to be active. A pharmacophore model usually encodes characteristic type, its position, and direction of an active compound in addition to its probable steric constraints (van Drie 2003). A three-dimensional pharmacophore depicts the positioning of important active site amino acids in cavity of the drug target (Wolber et al. 2008). For instance, a residue acting as an acceptor of hydrogen bond must be located near a hydrogen-bond donor characteristic in the pharmacophore model, responsible for ligand interaction with the protein. Based on the process of protein interaction with the ligand, the target protein might change its conformation or lock itself upon interaction (Drews 2000). A set of known inhibitors can be utilized for generation of a pharmacophore model. However, data related to crystal structure or structure in complex with ligand in combination with data on active site residues can also be utilized for pharmacophore modeling (Yang 2010). The building of pharmacophore models is based on the information of target protein structure by analyzing the binding cavity, probable molecular contacts between the ligand (active) and target protein. Pharmacophore models have been widely used for generation of disease-related protein-specific inhibitors, for instance, inhibitors against enzymes, ion channels, and G-protein-coupled receptors (Kubinyi 2006). It can also be implemented in combination with other drug designing methods.

Although a comprehensive work plan of pharmacophore model development depends on the software used, the common protocol followed is as mentioned by van Drie (2003). Generally, a software package is used for the generation of pharmacophore models as they conduct the whole workflow and comprise of all tools required for this purpose. The first step involves assembling a set of active ligands, usually with the help of literature reviews and querying molecular databases. All accessible databases including commercial or public can be employed for this purpose. However, a steady threshold should be applied to classify a compound as active while identifying active compounds from several sources. For construction of a 3D pharmacophore model, generation of conformers of the ligand is a prerequisite. International Union of Pure and Applied Chemistry (IUPAC) defines the conformations of any molecule as "the spatial arrangement of the atoms affording distinction between stereoisomers" (McNaught and Wilkinson 1997). After generation of the conformers, the set of ligands are superimposed or aligned (Yang 2010) to determine the common characteristics of the selected set of ligands. Pharmacophore elucidation algorithm is employed for construction of pharmacophore models after completion of the ligand alignment. Usually, multiple pharmacophore models are constructed from the set of selected ligands. The next step comprises ranking of all the models generated using the scoring function. In general, the pharmacophore model possessing the highest score is chosen. The final step includes validation of the pharmacophore model generated. If information pertaining to binding mechanism of ligand is clearly known, it can be applied for model validation, since pharmacophore models might also reflect the 3D structure of the binding cavity (Wolber et al. 2008). A rational approach in case of any discrepancies observed between the experimentally reported binding data and the chosen pharmacophore model would be rejecting the model. Any selected pharmacophore model must always be validated using compounds not belonging to training set irrespective of whether information regarding binding mechanisms is present or not.

14.8 A Case Study Using VLife Software

14.8.1 Selection of Dataset and Its Presentation

An experimentally reported set of 38 thiazolyl-pyrazoline derivatives along with its template was prepared using ChemSketch (Spessard 1998) and was minimized using VLife molecular design suite (Fig. 14.3a) (Akamatsu 2002; Vlife 2008). These molecules were then aligned using VLife Engine (Vlife 2008).

14.8.2 Force Field Computation

Dataset of derivatives of thiazolyl-pyrazoline along with their pIC50 values (negative logarithm of IC50) were provided for the calculation of force field. The grid dimensions were kept as default (21.6 × 6.9 × 21.4) and all three classes of descriptors, namely, steric, electrostatic, and hydrophobic, were calculated (Fig. 14.3b).

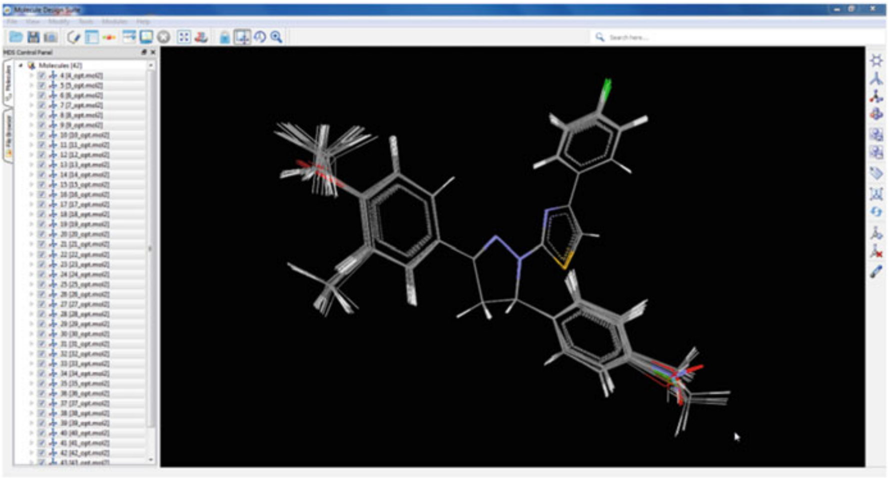
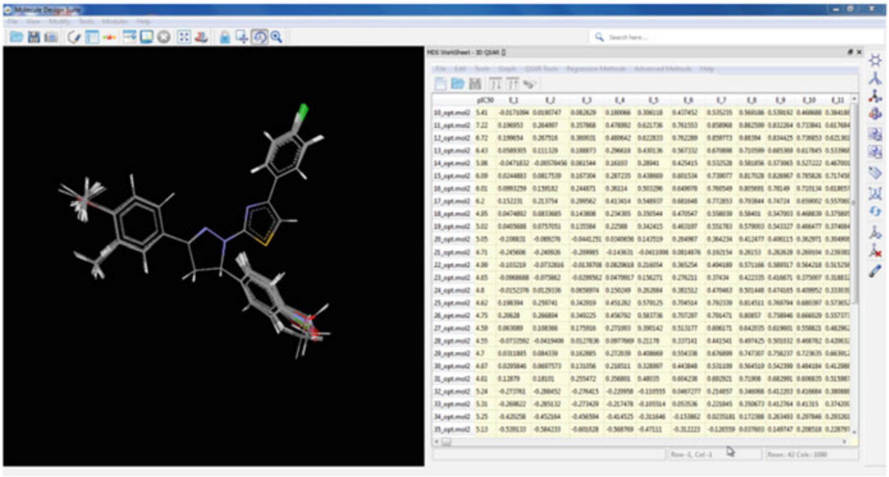


Fig. 14.3a An alignment of thiazolyl-pyrazoline derivatives


Molecule Design Suite
File View Modify Tools Modules Help
Search here...
MDS WorkSheet - 3D QSAR
File Edit Tools Graph QSAR Tools Regression Methods Advanced Methods Help

	pIC50	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_10	E_11
10_opt.mol2	5.41	-0.0171094	0.0190747	0.082629	0.180066	0.306118	0.437452	0.535235	0.560186	0.539192	0.466688	0.38418
11_opt.mol2	7.22	0.198953	0.264907	0.357868	0.478992	0.621736	0.761553	0.858968	0.882509	0.832264	0.733841	0.61768
12_opt.mol2	6.72	0.199654	0.267516	0.360031	0.480642	0.622833	0.762289	0.859773	0.88394	0.834425	0.736853	0.62138
13_opt.mol2	6.43	0.0589305	0.111329	0.188873	0.296618	0.430136	0.567332	0.670898	0.710509	0.685369	0.617845	0.53396
14_opt.mol2	5.08	-0.0471832	-0.00578456	0.061544	0.16103	0.28041	0.425415	0.532528	0.581856	0.573065	0.527222	0.46700
15_opt.mol2	6.09	0.0244883	0.0817539	0.167304	0.287235	0.438869	0.601534	0.739077	0.817028	0.826967	0.785826	0.71745
16_opt.mol2	6.01	0.0991259	0.159182	0.244871	0.36114	0.501296	0.649078	0.760549	0.805691	0.78149	0.710134	0.61865
17_opt.mol2	6.2	0.152231	0.213754	0.299562	0.413414	0.548937	0.681648	0.772853	0.793844	0.74724	0.659002	0.55708
18_opt.mol2	4.05	0.0474892	0.0833685	0.143808	0.234305	0.350544	0.470547	0.558039	0.58401	0.547003	0.468839	0.37589
19_opt.mol2	5.02	0.0409688	0.0757051	0.135584	0.22588	0.342415	0.463197	0.551783	0.579083	0.543327	0.466477	0.37468
20_opt.mol2	5.05	-0.108831	-0.089276	-0.0441251	0.0340656	0.143519	0.264987	0.364234	0.412477	0.408115	0.362971	0.30490
21_opt.mol2	4.71	-0.245606	-0.240926	-0.209985	-0.143631	-0.0411006	0.0814876	0.192154	0.26153	0.282628	0.269104	0.23938
22_opt.mol2	4.00	-0.103219	-0.0732816	-0.0139708	0.0829618	0.216054	0.365254	0.494189	0.571166	0.589517	0.564218	0.51525
23_opt.mol2	4.65	-0.0968688	-0.075862	-0.0299562	0.0479917	0.156271	0.276211	0.37434	0.422335	0.418671	0.375007	0.31881
24_opt.mol2	4.8	-0.0152376	0.0129136	0.0658974	0.150249	0.262984	0.381512	0.470463	0.501448	0.474165	0.409952	0.33303
25_opt.mol2	4.62	0.198394	0.259741	0.342919	0.451282	0.570125	0.704514	0.792139	0.814511	0.768794	0.680387	0.57365
26_opt.mol2	4.75	0.20628	0.266894	0.349225	0.456792	0.583736	0.707287	0.796471	0.80857	0.758946	0.666029	0.55737
27_opt.mol2	4.59	0.063089	0.108366	0.175916	0.271003	0.390142	0.513177	0.606171	0.642035	0.619601	0.558821	0.48296
28_opt.mol2	4.55	-0.0732592	-0.0419406	0.0127836	0.0977669	0.21178	0.337141	0.441541	0.497425	0.501032	0.468782	0.42963
29_opt.mol2	4.7	0.0311885	0.084139	0.162885	0.272039	0.408669	0.554338	0.676899	0.747307	0.758237	0.723635	0.66391
30_opt.mol2	4.67	0.0295846	0.0697573	0.131056	0.218511	0.328097	0.443848	0.531109	0.564519	0.542399	0.484184	0.41298
31_opt.mol2	4.61	0.12879	0.18101	0.255472	0.356801	0.48035	0.604238	0.692921	0.71908	0.682991	0.606835	0.51398
32_opt.mol2	5.24	-0.273781	-0.288452	-0.276415	-0.229958	-0.110555	0.0467277	0.234857	0.346068	0.412203	0.416684	0.38088
33_opt.mol2	5.31	-0.260622	-0.285132	-0.273429	-0.217478	-0.105514	0.052536	0.221845	0.350673	0.412764	0.41315	0.37420
34_opt.mol2	5.25	-0.420258	-0.452164	-0.456594	-0.414525	-0.311646	-0.153862	0.0235181	0.172388	0.263493	0.297846	0.29326
35_opt.mol2	5.13	-0.539133	-0.584233	-0.601028	-0.568769	-0.47111	-0.312223	-0.128559	0.037683	0.149747	0.208518	0.22879

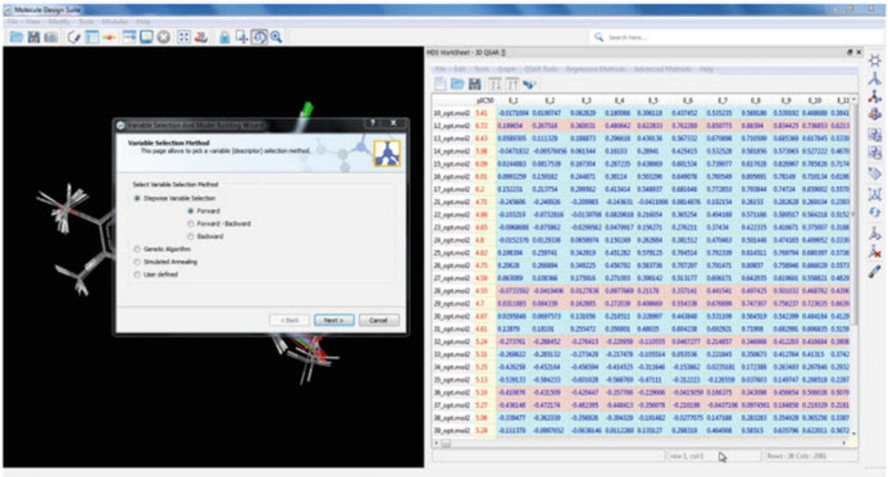
Row -1, Col -1

Rows : 42 Cells : 1080

Fig. 14.3b Calculation of steric, electrostatic, and hydrophobic descriptors

14.8.3 Building the 3D-QSAR Model of Thiazolyl-Pyrazoline-Derived Compounds

The experimentally reported compounds were classified into two sets, training and test, using sphere exclusion method. This step resulted in test set with 11 compounds, while the remaining 27 molecules in training dataset. Wizard for variable selection and building of QSAR model was employed using stepwise forward multiple regression method with default values for generation of 3D-QSAR model (Fig. 14.3c).


Molecule Design Suite
File View Modify Tools Modules Help
Search here...
MDS Worksheet - 3D QSAR
File Edit Tools Graph QSAR Tools Regression Methods Advanced Methods Help
Variable Selection And Model Building Wizard
< Back
Next >
Cancel

Variable Selection Method
This page allows to pick a variable (descriptor) selection method.

Select Variable Selection Method

● Stepwise Variable Selection
  - ● Forward
  - ○ Forward - Backward
  - ○ Backward

○ Genetic Algorithm
○ Simulated Annealing
○ User defined

	pIC50	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_10	E_11
10_opt.mol2	5.41	-0.0171094	0.0190747	0.062829	0.180066	0.306118	0.437452	0.535235	0.569186	0.539192	0.468688	0.3841
12_opt.mol2	6.72	0.199654	0.267516	0.360031	0.480642	0.622833	0.762289	0.850773	0.88394	0.834425	0.736853	0.6213
13_opt.mol2	6.43	0.0589305	0.111329	0.188873	0.296618	0.430136	0.567332	0.670698	0.710599	0.685369	0.617845	0.5330
14_opt.mol2	5.98	-0.0471832	-0.00578456	0.061544	0.16103	0.28941	0.425415	0.532528	0.581856	0.573065	0.527222	0.4670
15_opt.mol2	6.09	0.0244083	0.0817539	0.167304	0.287235	0.438869	0.601534	0.739077	0.817928	0.826967	0.785826	0.7174
16_opt.mol2	6.01	0.0991259	0.159182	0.244871	0.38114	0.503296	0.649078	0.760549	0.805691	0.78149	0.710134	0.6186
17_opt.mol2	6.2	0.152231	0.213754	0.299562	0.413414	0.548837	0.681648	0.772853	0.793844	0.74724	0.659002	0.5570
21_opt.mol2	4.71	-0.245606	-0.240026	-0.209085	-0.143631	-0.0411006	0.0814876	0.192154	0.26153	0.282628	0.260104	0.2303
22_opt.mol2	4.98	-0.103219	-0.0732816	-0.0130708	0.0829618	0.216054	0.365254	0.494189	0.571166	0.589517	0.564218	0.5152
23_opt.mol2	4.65	-0.0968688	-0.075862	-0.0299562	0.0479917	0.156271	0.276211	0.37434	0.422335	0.416671	0.375907	0.3188
24_opt.mol2	4.8	-0.0152376	0.0129336	0.0658974	0.150249	0.262684	0.381512	0.470463	0.501448	0.474165	0.409952	0.3330
25_opt.mol2	4.62	0.198394	0.259741	0.342919	0.451282	0.578125	0.704514	0.792339	0.814511	0.769794	0.680397	0.5736
26_opt.mol2	4.75	0.20628	0.266894	0.349225	0.456792	0.583736	0.707297	0.791471	0.80857	0.758946	0.669029	0.5573
27_opt.mol2	4.58	0.063089	0.108366	0.175916	0.271003	0.390142	0.513177	0.606171	0.642035	0.610601	0.558821	0.4820
28_opt.mol2	4.55	-0.0733592	-0.0419406	0.0127836	0.0977669	0.21178	0.337141	0.441541	0.497425	0.501032	0.468782	0.4206
29_opt.mol2	4.7	0.0311885	0.084339	0.162885	0.272039	0.408669	0.554338	0.676899	0.747307	0.758237	0.723635	0.6639
30_opt.mol2	4.67	0.0295846	0.0697573	0.131056	0.218511	0.328997	0.443848	0.531109	0.564519	0.542399	0.484184	0.4129
31_opt.mol2	4.61	0.12879	0.18101	0.255472	0.356801	0.48035	0.604238	0.692921	0.71908	0.682981	0.606835	0.5158
32_opt.mol2	5.24	-0.273761	-0.288452	-0.276415	-0.229958	-0.110555	0.0467277	0.214857	0.346068	0.412203	0.416684	0.3808
33_opt.mol2	5.31	-0.268822	-0.285132	-0.273429	-0.217478	-0.105514	0.053536	0.221845	0.350673	0.412764	0.41315	0.3742
34_opt.mol2	5.25	-0.426258	-0.452164	-0.456594	-0.414525	-0.311646	-0.153862	0.0275181	0.172388	0.263493	0.297846	0.2932
35_opt.mol2	5.13	-0.539133	-0.584233	-0.601028	-0.568769	-0.47111	-0.312223	-0.128559	0.037603	0.149747	0.208518	0.2287
36_opt.mol2	5.16	-0.410876	-0.431509	-0.420447	-0.357766	-0.229066	-0.0415059	0.166375	0.343098	0.456654	0.506038	0.5079
37_opt.mol2	5.27	-0.436148	-0.472174	-0.482395	-0.448413	-0.356078	-0.210198	-0.0437106	0.0974561	0.184858	0.218129	0.2181
38_opt.mol2	5.06	-0.339477	-0.362339	-0.356926	-0.304329	-0.191482	-0.0277075	0.147188	0.283283	0.354029	0.365256	0.3387
39_opt.mol2	5.26	-0.111379	-0.0097652	-0.0638146	0.0112289	0.135127	0.298319	0.464008	0.58515	0.635796	0.622011	0.5672

row 1, col 1
Rows : 38 Cols : 2081

Fig. 14.3c Generation of QSAR model

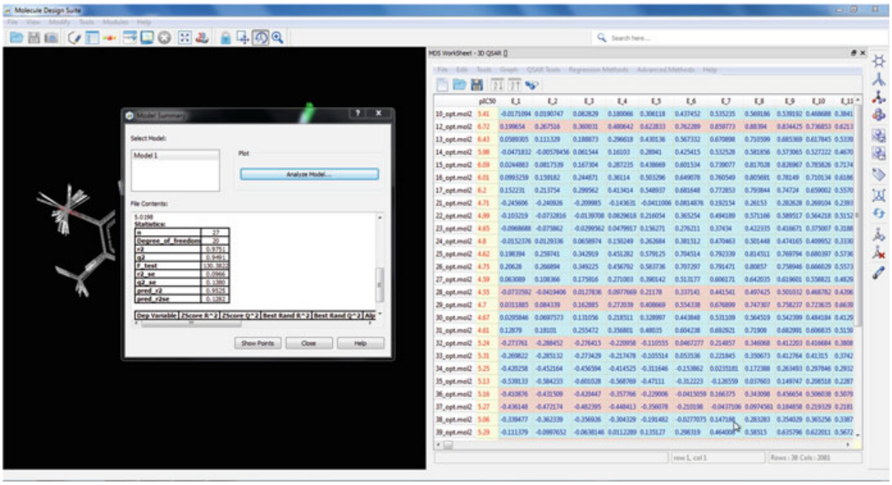
14.8.4 Validation of Developed 3D-QSAR Model

Reliability of generated 3D-QSAR model is established using internal and external validation protocols. Statistical criteria setup for testing reliability of generated model is (correlation coefficient) r^2 and (cross-validated correlation coefficient) q^2 > 0.6 and (predicted correlation coefficient) pred_r^2 > 0.5. The descriptors selected in QSAR model were E_337, E_832, E_424, S_151, S_335, and E_721. Alphabets preceding the chosen descriptors refer to the steric and electrostatic class, while the associated numbers represent their corresponding spatial grid points (Fig. 14.3d). The 3D-QSAR model obtained was:

$$
\begin{aligned}
pIC50 = {} & [0.2989(\pm 0.0020) \times E\_337] + [3.2763(\pm 0.5560) \times S\_335] \\
& + [0.1785(\pm 0.0003) \times E\_832] + [0.4938(\pm 0.0033) \times E\_424] \\
& - [11.7460(\pm 0.3402) \times S\_151] - [0.6486(\pm 0.0019) \times E\_721] \\
& + 5.0198
\end{aligned}
$$

14.8.5 Model Cross-Validation

As described previously, validation of QSAR model was performed using both internal and external protocols. The statistical parameters for generated QSAR model comprised of r^2 (0.9751), q^2 (0.9491), pred_r^2 (0.9525), standard error value, r^2_se (0.0966), q^2_se (0.1380), and pred_r^2_se (0.1282) which validates model as reliable (Fig. 14.3e). The detailed study can be seen in our paper (Goyal et al. 2015).


Molecule Design Suite
File View Modify Tools Modules Help
Search here...
Model Summary
Analyze Model...
Show Points
Close
Help
MDS WorkSheet - 3D QSAR
File Edit Tools Graph QSAR Tools Regression Methods Advanced Methods Help

Select Model:

Model 1

Plot

File Contents:

5-0.198

Statistics:

n	27
Degree_of_freedom	20
r2	0.9751
q2	0.9491
F_test	130.3822
r2_se	0.0966
q2_se	0.1380
pred_r2	0.9525
pred_r2se	0.1282

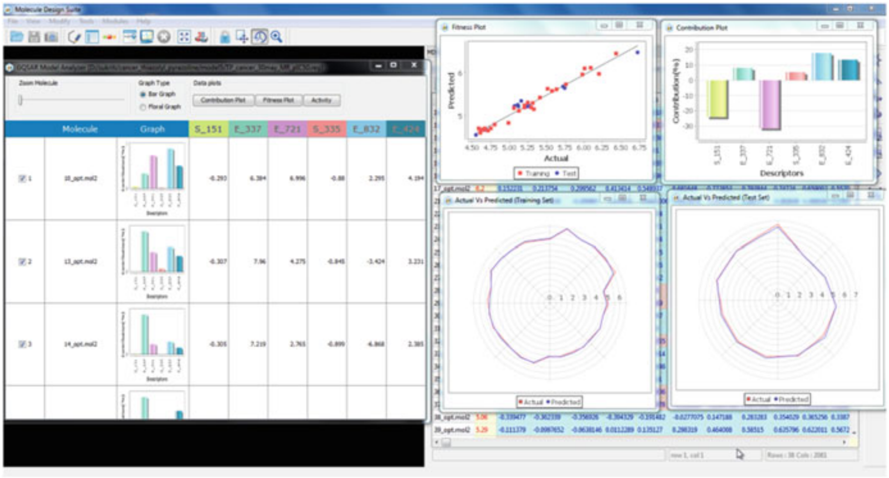
Dep Variable	ZScore R^2	ZScore Q^2	Best Rand R^2	Best Rand Q^2	Alp

	pIC50	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_10	E_11
10_opt.mol2	5.41	-0.0171094	0.0190747	0.082829	0.180066	0.306118	0.437452	0.535235	0.569186	0.539192	0.468688	0.3841
12_opt.mol2	6.72	0.199654	0.267516	0.360031	0.480642	0.622833	0.762289	0.858773	0.88394	0.834425	0.736853	0.6213
13_opt.mol2	6.43	0.0589305	0.111329	0.188873	0.296618	0.430136	0.567332	0.670898	0.710599	0.685369	0.617845	0.5339
14_opt.mol2	5.98	-0.0471832	-0.00578456	0.061544	0.16103	0.28941	0.425415	0.532528	0.581856	0.573065	0.527222	0.4670
15_opt.mol2	6.09	0.0244883	0.0817539	0.167304	0.287235	0.438669	0.601534	0.739077	0.817028	0.828967	0.785826	0.7174
16_opt.mol2	6.01	0.0993259	0.159182	0.244871	0.36114	0.503296	0.649078	0.760549	0.805691	0.78149	0.710134	0.6186
17_opt.mol2	6.2	0.152231	0.213754	0.299562	0.413414	0.548937	0.681648	0.772853	0.793844	0.74724	0.659002	0.5570
21_opt.mol2	4.71	-0.245606	-0.240926	-0.209985	-0.143631	-0.0411006	0.0814876	0.192154	0.26153	0.282628	0.269104	0.2393
22_opt.mol2	4.99	-0.103219	-0.0732816	-0.0139708	0.0829618	0.216054	0.365254	0.494189	0.571166	0.589517	0.564218	0.5152
23_opt.mol2	4.65	-0.0968688	-0.075862	-0.0299562	0.0479817	0.156271	0.276211	0.37434	0.422335	0.416671	0.375007	0.3188
24_opt.mol2	4.8	-0.0152376	0.0129336	0.0658974	0.150249	0.262684	0.381512	0.470463	0.501448	0.474165	0.409952	0.3330
25_opt.mol2	4.62	0.198394	0.259741	0.342919	0.451282	0.579125	0.704514	0.792339	0.814511	0.769794	0.680397	0.5736
26_opt.mol2	4.75	0.20628	0.266894	0.349225	0.456792	0.583736	0.707297	0.791471	0.80857	0.758946	0.666029	0.5573
27_opt.mol2	4.59	0.063089	0.108366	0.175916	0.271003	0.390142	0.513177	0.606171	0.642035	0.619601	0.558821	0.4829
28_opt.mol2	4.55	-0.0733592	-0.0419406	0.0127836	0.0977669	0.21178	0.337141	0.441541	0.497425	0.501032	0.468782	0.4206
29_opt.mol2	4.7	0.0311885	0.084339	0.162885	0.272039	0.408669	0.554338	0.676899	0.747307	0.758237	0.723635	0.6639
30_opt.mol2	4.67	0.0295846	0.0697573	0.131056	0.218511	0.328997	0.443848	0.531109	0.564519	0.542399	0.484184	0.4129
31_opt.mol2	4.81	0.12879	0.18101	0.255472	0.356801	0.48035	0.604238	0.692921	0.71908	0.682991	0.606835	0.5159
32_opt.mol2	5.24	-0.273761	-0.288452	-0.276415	-0.220958	-0.110555	0.0467277	0.254857	0.346068	0.412203	0.416684	0.3808
33_opt.mol2	5.31	-0.269822	-0.285132	-0.273429	-0.217478	-0.105514	0.053536	0.221845	0.350673	0.412764	0.41315	0.3742
34_opt.mol2	5.25	-0.420258	-0.452164	-0.456594	-0.414525	-0.311646	-0.153862	0.0235181	0.172388	0.263493	0.297846	0.2932
35_opt.mol2	5.13	-0.539133	-0.584233	-0.601028	-0.568769	-0.47111	-0.312223	-0.126559	0.037603	0.149747	0.208518	0.2287
36_opt.mol2	5.16	-0.410876	-0.431509	-0.420447	-0.357766	-0.229006	-0.0415059	0.166375	0.343098	0.456654	0.506038	0.5079
37_opt.mol2	5.27	-0.436148	-0.472174	-0.482395	-0.448413	-0.356078	-0.210198	-0.0437106	0.0974561	0.184858	0.219329	0.2181
38_opt.mol2	5.06	-0.339477	-0.362339	-0.356926	-0.304329	-0.191482	-0.0277075	0.147188	0.283283	0.354029	0.365256	0.3387
39_opt.mol2	5.29	-0.111379	-0.0997652	-0.0638146	0.0112289	0.135127	0.298319	0.464008	0.58515	0.635796	0.622011	0.5672

row 1, col 1

Rows : 38 Cols : 2081

Fig. 14.3d The QSAR model generated



	Molecule	Graph	S_151	E_337	E_721	S_335	E_832	E_424
☑ 1	10_opt.mol2		-0.293	6.384	6.996	-0.88	2.295	4.194
☑ 2	13_opt.mol2		-0.307	7.96	4.275	-0.845	-3.424	3.231
☑ 3	14_opt.mol2		-0.305	7.219	2.765	-0.899	-4.868	2.385

Fig. 14.3e Parameters calculated for cross-validation of generated QSAR model

Among the various computational tools available for drug discovery, the performance of each method differs with the target protein, other information, and resources used. Even though computer-aided drug discovery has been employed extensively, some advantageous targets such as interactions between protein-protein or protein-DNA are still difficult issues due to the massive size of interaction sites. Interfaces which are not user-friendly, a large number of variables, and the expertise required to derive successful results are a few problems faced in this field.

Nevertheless, ongoing efforts to make user-friendly softwares and protocols have improved the issues and promise better tools in the future.

References

Akamatsu M (2002) Current state and perspectives of 3D-QSAR. Curr Top Med Chem 2:1381–1394

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

Becker OM et al (2006) An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. J Med Chem 49:3116–3135

Blake JF, Laird ER (2003) Chapter 30: recent advances in virtual ligand screening. Annu Rep Med Chem 38:305–314

Bower MJ, Cohen FE, Dunbrack RL (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 267:1268–1282

Bradbury SP (1995) Quantitative structure-activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research. Toxicol Lett 79:229–237

Buchan DW, Ward S, Lobley AE, Nugent T, Bryson K, Jones DT (2010) Protein annotation and modelling servers at University College London. Nucleic Acids Res 38:W563–W568

Budzik B et al (2010) Novel N-substituted benzimidazolones as potent, selective, CNS-penetrant, and orally active M1 mAChR agonists. ACS Med Chem Lett 1:244–248

Changeux J-P, Edelstein S (2011) Conformational selection or induced fit? 50 years of debate resolved. F1000 biology reports 3

Chen H, Zhou J, Xie G (1998) PARM: a genetic evolved algorithm to predict bioactivity. J Chem Inf Comput Sci 38:243–250

Cheng F et al (2012) admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. J Chem Inf Model 52:3099–3105

Chivian D, Baker D (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res 34:e112–e112

Consonni V, Ballabio D, Todeschini R (2009) Comments on the definition of the Q 2 parameter for QSAR validation. J Chem Inf Model 49:1669–1678

Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A (2009) Evaluation of template-based models in CASP8 with standard measures. Proteins: Struct Funct Bioinf 77:18–28

Cramer RD, Bunce JD, Patterson DE, Frank IE (1988a) Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. Quant Struct Act Relat 7:18–25

Cramer RD, Patterson DE, Bunce JD (1988b) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967

de Melo EB (2010) Multivariate SAR/QSAR of 3-aryl-4-hydroxyquinolin-2 (1H)-one derivatives as type I fatty acid synthase (FAS) inhibitors. Eur J Med Chem 45:5817–5826

Desmet J, De Maeyer M, Hazes B, Lasters I (1992) The dead-end elimination theorem and its use in protein side-chain positioning. Nature 356:539–542

Dhanjal JK, Goyal S, Sharma S, Hamid R, Grover A (2014a) Mechanistic insights into mode of action of potent natural antagonists of BACE-1 for checking Alzheimer's plaque pathology. Biochem Biophys Res Commun 443:1054–1059

Dhanjal JK, Grover S, Paruthi P, Sharma S, Grover A (2014b) Mechanistic insights into mode of action of a potent natural antagonist of orexin Receptor-1 by means of high throughput screening and molecular dynamics simulations. Comb Chem High Throughput Screen 17:124–131

Dias R, de Azevedo J, Walter F (2008) Molecular docking algorithms. Curr Drug Targets 9:1040–1047

Doman TN et al (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 45:2213–2221

Drews J (2000) Drug discovery: a historical perspective. Science 287:1960–1964

Du Q-S, Huang R-B, Chou K-C (2008) Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Curr Protein Pept Sci 9:248–259

Dunbrack RL, Karplus M (1993) Backbone-dependent rotamer library for proteins application to side-chain prediction. J Mol Biol 230:543–574

Dunbrack RL, Karplus M (1994) Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Mol Biol 1:334–340

Ehrlich P (1909) Über den jetzigen Stand der Chemotherapie. Ber Dtsch Chem Ges 42:17–47

Esposito EX, Hopfinger AJ, Madura JD (2004) Methods for applying the quantitative structure-activity relationship paradigm. Methods Mol Biol 275:131–213

Friesner RA et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47:1739–1749. https://doi.org/10.1021/jm0306430

Gillet VJ, Willett P, Bradshaw J, Green DV (1999) Selecting combinatorial libraries to optimize diversity and physical properties. J Chem Inf Comput Sci 39:169–177

Goyal S et al (2014) Mechanistic insights into mode of actions of novel oligopeptidase B inhibitors for combating leishmaniasis. J Mol Model 20:1–9

Goyal S, Jamal S, Shanker A, Grover A (2015) Structural investigations of T854A mutation in EGFR and identification of novel inhibitors using structure activity relationships. BMC Genomics 16:S8

Grover S, Dhanjal JK, Goyal S, Grover A, Sundar D (2014) Computational identification of novel natural inhibitors of glucagon receptor for checking type II diabetes mellitus. BMC Bioinformatics 15:S13

Halgren TA et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47:1750–1759

Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins: Struct Funct Bioinf 47:409–443

Henrich S, Salo-Ahen OM, Huang B, Rippmann FF, Cruciani G, Wade RC (2010) Computational approaches to identifying and characterizing protein binding sites for ligand design. J Mol Recognit 23:209–219

Hillisch A, Pineda LF, Hilgenfeld R (2004) Utility of homology models in the drug discovery process. Drug Discov Today 9:659–669

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

Ioakimidis L, Thoukydidis L, Mirza A, Naeem S, Reynisson J (2008) Benchmarking the reliability of QikProp. Correlation between experimental and predicted values. QSAR Comb Sci 27:445–456

Irwin JJ, Shoichet BK (2005) ZINC – a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182. https://doi.org/10.1021/ci049714+

Itskowitz P, Tropsha A (2005) K nearest neighbors QSAR modeling as a variational problem: theory and applications. J Chem Inf Model 45:777–785

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

Kalyaanamoorthy S, Chen Y-PP (2011) Structure-based drug design to augment hit discovery. Drug Discov Today 16:831–839

Kapetanovic I (2008) Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. Chem Biol Interact 171:165–176

Karelson M (2000) Molecular descriptors in QSAR/QSPR. Wiley-Interscience, New York

Karolidis DA, Agatonovic-Kustrin S, Morton DW (2010) Artificial neural network (ANN) based modelling for D1 like and D2 like dopamine receptor affinity and selectivity. Med Chem 6:259–270

Kelley LA, Sternberg MJ (2009) Protein structure prediction on the web: a case study using the Phyre server. Nat Protoc 4:363–371

Kennedy T (1997) Managing the drug discovery/development interface. Drug Discov Today 2:436–444

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

Kotz J (2013) In silico drug design. SciBX: Science-Business eXchange 6

Krivov GG, Shapovalov MV, Dunbrack RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins Struct Funct Bioinf 77:778–795

Kubinyi H (2006) Success stories of computer-aided design. In: Ekins S, Wang B (eds) Computer applications in pharmaceutical research and development. Wiley-Interscience, pp 377–424

Kurogi Y, Guner OF (2001) Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. Curr Med Chem 8:1035–1055

Lahana R (1999) How many leads from HTS? Drug Discov Today 4:447–448

Laurie R, Alasdair T, Jackson RM (2006) Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. Curr Protein Pept Sci 7:395–406

Ligprep V. 2.3 (2009) Schrodinger. LLC, New York

Lite V (1998) Version 5.0. Accelrys Inc., 9685

Lu I-L et al (2006) Structure-based drug design of a novel family of PPARγ partial agonists: virtual screening, X-ray crystallography, and in vitro/in vivo biological activities. J Med Chem 49:2703–2712

Martinez-Mayorga K, Medina-Franco JL (2009) Chemoinformatics—applications in food chemistry. Adv Food Nutr Res 58:33–56

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

McNaught AD, Wilkinson A (1997) Compendium of chemical terminology, IUPAC Recommendations, The Gold Book, 2nd edn. Blackwell Science, Oxford

Misura K, Baker D (2005) Progress and challenges in high-resolution refinement of protein structure models. Proteins: Struct Funct Bioinf 59:15–29

Misura KM, Chivian D, Rohl CA, Kim DE, Baker D (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci 103:5361–5366

Owen D (2002) Channelling drug discovery. Drug Discov World 3:48–61

Put R, Perrin C, Questier F, Coomans D, Massart D, Vander Heyden Y (2003) Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure–retention relationship studies. J Chromatogr A 988:261–276

Put R, Xu Q, Massart D, Vander Heyden Y (2004) Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies. J Chromatogr A 1055:11–19

Raval A, Piana S, Eastwood MP, Dror RO, Shaw DE (2012) Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. Proteins: Struct Funct Bioinf 80:2071–2079

Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf Comput Sci 34:854–866

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

Rose PW et al (2011) The RCSB protein data Bank: redesigned web site and web services. Nucleic Acids Res 39:D392–D401

Ruiz FM, Gil-Redondo R, Morreale A, Ortiz AR, Fábrega C, Bravo J (2008) Structure-based discovery of novel non-nucleosidic DNA alkyltransferase inhibitors: virtual screening and in vitro and in vivo activities. J Chem Inf Model 48:844–854

Schrödinger L (2008a) Glide, version 5.0. Schrödinger. LLC, New York

Schrödinger L (2008b) SCHRODINGER SUITE 2009. Maestro Version 8

Schrödinger M (2009) Version 9.2. LLC, New York

Silverman B, Platt DE (1996) Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. J Med Chem 39:2129–2140

Söding J, Remmert M (2011) Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr Opin Struct Biol 21:404–411

Spessard GO (1998) ACD Labs/LogP dB 3.5 and ChemSketch 3.5. J Chem Inf Comput Sci 38:1250–1253

Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:1947–1958

Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. J Comput Aided Mol Des 16:151–166

Terstappen GC, Reggiani A (2001) In silico research in drug discovery. Trends Pharmacol Sci 22:23–26

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 22:4673–4680

Tropsha A (2003) Recent trends in quantitative structure activity relationships. In: Abraham D (ed) Burger's medicinal chemistry and drug discovery, vol 1. Wiley, New York, pp 49–75

Tyagi C, Grover S, Dhanjal JK, Goyal S, Goyal M, Grover A (2013) Mechanistic insights into mode of action of novel natural cathepsin L inhibitors. BMC Genomics 14:S10

Tyagi C et al (2015) Targeting the intersubunit cavity of Plasmodium falciparum glutathione reductase by a novel natural inhibitor: computational and experimental evidence. Int J Biochem Cell Biol 61:72–80

van de Waterbeemd H (2002) High-throughput and in silico techniques in drug metabolism and pharmacokinetics. Curr Opin Drug Discov Devel 5:33–43

Van De Waterbeemd H, Gifford E (2003) ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2:192–204

van Drie JH (2003) Pharmacophore discovery-lessons learned. Curr Pharm Des 9:1649–1664

Van Drie JH (2007) Computer-aided drug design: the next 20 years. J Comput Aided Mol Des 21:591–601

Verma J, Khedkar VM, Coutinho EC (2010) 3D-QSAR in drug design-a review. Curr Top Med Chem 10:95–115

Vlife M (2008) Software package, version 3.0, supplied by Vlifescience Technologies Pvt. Ltd., Pune

Warner SL et al (2006) Identification of a lead small-molecule inhibitor of the aurora kinases using a structure-assisted, fragment-based approach. Mol Cancer Ther 5:1764–1773

Wikel JH, Dow ER (1993) The use of neural networks for variable selection in QSAR. Bioorg Med Chem Lett 3:645–651

Wolber G, Seidel T, Bendix F, Langer T (2008) Molecule-pharmacophore superpositioning and pattern matching in computational drug design. Drug Discov Today 13:23–29

Xiang Z (2006) Advances in homology protein structure modeling. Curr Protein Pept Sci 7:217

Xue L, Godden JW, Bajorath J (2000) Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. J Chem Inf Comput Sci 40:1227–1234

Yang S-Y (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug Discov Today 15:444–450

Zhao L et al (2006) FK506-binding protein ligands: structure-based design, synthesis, and neurotrophic/neuroprotective properties of substituted 5, 5-dimethyl-2-(4-thiazolidine) carboxylates. J Med Chem 49:4059–4071

Zhou T et al (2010) Anti-AIDS agents 79. Design, synthesis, molecular modeling and structure–activity relationships of novel dicamphanoyl-2′, 2′-dimethyldihydropyranochromone (DCP) analogs as potent anti-HIV agents. Bioorg Med Chem 18:6678–6689

Advanced In Silico Tools for Designing of Antigenic Epitope as Potential Vaccine Candidates Against *Corona*virus

15

Mehak Dangi, Rinku Kumari, Bharat Singh, and Anil Kumar Chhillar

15.1 Introduction

Coronaviruses are remarkably large, positive-stranded RNA viruses that are enveloped with the nucleocapsid having helical symmetry. The *corona in coronavirus* is a Latin word that means a “crown”, and it indicates to the typical presentation of virions underneath electron microscopy with a periphery of hefty, globular surface projections similar to that of a crown. Coronavirus is a pathogen associated with severe respiratory symptoms and was first identified from the nasal cavities of sufferers with the common cold in the early 1960s (de Groot et al. 2013; Brown et al. 2012). These were named human coronavirus OC43 and human coronavirus 229E. A total of 40 sequenced genomes of different strains of coronavirus are accessible from National Center for Biotechnology Information (NCBI), out of which 7 are pathogenic to humans. A coronavirus, i.e. SARS-CoV, was responsible for outbreak of severe acute respiratory syndrome (SARS) in the year 2003, whereas Middle East respiratory syndrome coronavirus (MERS-CoV) caused the most recent outbreak in 2012 causing acute respiratory disease in affected people with signs of fever, cough and difficulty in breathing. After first reported from Saudi Arabia in 2012, this novel virus has also dispersed to other countries like the United States and was known to have high death rate. MERS-CoV infections are highly communicable, and no explicit antiviral cure has been designed for it till date (Azhar et al. 2017).

M. Dangi
Centre for Bioinformatics, Maharshi Dayanand University, Rohtak, Haryana, India

Centre for Biotechnology, Maharshi Dayanand University, Rohtak, Haryana, India

R. Kumari
Centre for Bioinformatics, Maharshi Dayanand University, Rohtak, Haryana, India

B. Singh · A. K. Chhillar (✉)
Centre for Biotechnology, Maharshi Dayanand University, Rohtak, Haryana, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_15

It compelled us to apply the well-known reverse vaccinology (RV) approach on available proteome of coronavirus. RV approach has been successfully applied on many prokaryotes, but there are very few known applications on eukaryotes and viruses. So, it is worthwhile to explore the potential of this approach to identify potential vaccine candidates for coronavirus. RV basically does the in silico examination of the viral proteome to hunt antigenic and surface-exposed proteins. This approach was initially applied successfully to *Neisseria meningitidis* serogroup B (Kelly and Rappuoli 2005) against which none of the prevailing techniques could develop a vaccine. The present book chapter is intended to explore the potential of RV approach to select the probable vaccine candidates against coronavirus and validate the results using docking studies.

15.2 The Elementary Concept of Reverse Vaccinology

Undoubtedly, the traditional approaches for vaccine development are fortunate enough to efficiently resist the alarming pathogenic diseases of its time. However, the traditional approach suffers from certain limitations like it is very time-consuming, the pathogens which can't be cultivated in the lab conditions are out of reach, and certain non-abundant proteins are not accessible using this approach (Rappuoli 2000). Consequently, a number of pathogenic diseases are left without any vaccine against them. All these limitations are conquered by reverse vaccinology approach utilizing genome sequence information which ultimately is translated into proteins. Hence all the proteins expressed by the genome are accessible irrespective of their abundance, conditions in which they expressed. The credit of fame of reverse vaccinology should go to the advancements in the sequencing strategies worldwide. Accordingly, improvement in the sequencing technologies has flooded the genome databases with huge amount of data which can be computationally undertaken to reveal the various crucial aspects of the virulence factors of the concerned pathogen. Reverse vaccinology is based on same approach of computationally analysing the genome of pathogen and proceeds step by step to ultimately identify the highly antigenic, secreted proteins with high epitope densities. The best epitopes are selected as potential vaccine candidates (Pizza et al. 2000). This approach has brought the unapproachable pathogens of interest in spotlight and is evolving as the most reassuring tool for precise selection of vaccine candidates and brought the use of peptide vaccines in trend (Sette and Rappuoli 2010; Kanampalliwar et al. 2013).

15.3 Successful Applications of Reverse Vaccinology

Bexsero is the first universal serogroup B meningococcal vaccine developed using RV, and it has currently earned positive judgement from the European Medicines Agency (Gabutti 2014). Whether it is discovery of pili in gram-positive pathogens which were thought to not have any pili or the sighting of factor G-binding protein in meningococcus (Alessandro and Rino 2010), the reverse vaccinology steals all the credits from other conventional approaches. Most of the applications of RV are against prokaryotes and very few against eukaryotes and viruses because of complexity of their genome. *Corynebacterium urealyticum* (Guimarães et al. 2015), *Mycobacterium tuberculosis* (Monterrubio-López et al. 2015), *H. pylori* (Naz et al. 2015), *Acinetobacter baumannii* (Chiang et al. 2015), *Rickettsia prowazekii* (Caro-Gomez et al. 2014), *Neospora caninum* (Goodswen et al. 2014) and *Brucella melitensis* (Vishnu et al. 2017) are the examples of some pathogens that are recently approached using this in silico technique in order to spot some epitopes having potential of being a vaccine candidate. *Herpesviridae* (Bruno et al. 2015) and hepatitis C virus (HCV) (Kolesanova et al. 2015) are the examples of the viruses that are addressed using this approach.

15.4 Workflow of Reverse Vaccinology (with Example of Coronavirus)

15.4.1 Retrieval of Proteome of Different Strains of Coronavirus from NCBI

The proteome of different strains of the coronavirus of interest was downloaded from NCBI's ftp site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/; NCBI Resource Coordinators 2017). The proteome information is available for download in many formats including FASTA format for different sequenced viruses. Strains pathogenic to humans were selected for further analysis. Among them a single strain was selected as the seed genome on the basis of literature. Sequence similarity searches using Blastp (http://blast.ncbi.nlm.nih.gov/blast, http://ugene.unipro.ru/) were performed to reveal the orthologs in different strains (Altschul et al. 1990; Okonechnikov et al. 2012; Golosova et al. 2014). Multiple sequence alignment (MSA) was done via ClustalW, and the phylogenetic tree was constructed using NJ method from Unipro UGENE 1.16.1 bioinformatics toolkit (Okonechnikov et al. 2012).

15.4.2 Analysis of Secondary Structure of Proteins from Seed Genome

Analysis of secondary structure of the proteins of seed genome was done by means of ExPASy portal. The aim is to forecast the solvent accessibility, instability index, theoretical pI, molecular weight, grand average of hydropathicity (GRAVY), aliphatic index, number of charged residues, extinction coefficient etc. (http://web.expasy.org/protparam/; Gasteiger et al. 2005).

15.4.3 Subcellular Localization Predictions and Count of Transmembrane Helices

Virus-mPLoc was used to identify the localization of proteins of virus in the infected cells of host (http://www.csbio.sjtu.edu.cn/bioinf/virus-multi/; Hong-Bin Shen and Kuo-Chin Chou 2010). This information is important to understand the destructive role and mechanism of the viral proteins in causing the disease. In total six different subcellular locations, namely, host cytoplasm, viral capsid, host plasma membrane, host nucleus, host endoplasmic reticulum and secreted proteins, were covered. These predictions could help in formulation of better therapeutic options against the virus. As per the protocol of RV, secreted and membrane proteins are of special interest, therefore, filtered for further analysis. To predict the number of transmembrane helices TMHMM Server *v. 2.0* (http://www.cbs.dtu.dk/services/TMHMM/; Krogh et al. 2001) was used.

15.4.4 Signal Peptides

Signal peptides are known to impact the immune responses and possess high epitope densities. Moreover, most of the known vaccine candidates also possess signal peptides. Hence, it is worthwhile to predict signal peptides in proteins prior to epitope predictions. Signal-BLAST web server is used to predict the signal peptides without any false predictions (http://sigpep.services.came.sbg.ac.at/signalblast.html; Frank and Sippl 2008). The prediction options include best sensitivity, balanced prediction, best specificity and detect cleavage site only. We choose to make the predictions using each option, and the proteins predicted as signal peptide by all the four options were preferred for further investigation.

15.4.5 Adhesion Probability

The most appropriate targets as vaccine candidates are those which possess the adhesion-like properties because they not only mediate the adhesion of pathogen's proteins with cells of host but also facilitate transmission of virus. Adhesions are known to be crucial for virulence and are located on surface which makes them promptly approachable to antibodies. The stand-alone SPAAN with a sensitivity of 89% and specificity of 100% was used to carry out the adhesion probability predictions, and the proteins with having adhesion probabilities higher than or equal to 0.4 were selected (Sachdeva et al. 2004).

15.4.6 BetaWrap Motifs

BetaWrap motifs are dominant in virulence factors of the pathogens. If the proteins are predicted to possess such motifs, then they are appropriate to be taken under reverse vaccinology studies. BetaWrap server is the only online web server to make such predictions. The proteins having P-value lower than 0.1 were anticipated to contain BetaWraps (http://groups.csail.mit.edu/cb/betawrap/betawrap.html; Bradley et al. 2001).

15.4.7 Antigenicity Predictions

For added identification of the antigenic likely of the proteins, they were subjected to VaxiJen server *version 2.0.* It is basically an empirical method to hunt antigenic proteins. So, if the proteins are not found antigenic using other sequence-based methods, then they can be identified using this method. This step confirms the antigenicity of proteins selected using above-mentioned steps (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html; Doytchinova and Flower 2007).

15.4.8 Allergenicity Predictions

For being a probable vaccine candidate, the protein should not exhibit the characteristics of an allergen as they trigger the type-1 hypersensitivity reactions causing allergy. Therefore, to escape out such possibilities, the proteins were also subjected to allergenicity predictions using Allertop (http://www.pharmfac.net/allertop; Dimitrov et al. 2014) and AlgPred tools (http://www.imtech.res.in/raghava/algpred/submission.html; Saha and Raghava 2006a, b).

15.4.9 Similarity with Host Proteins

To check whether the filtered proteins possess any similarity to host proteins or not, the standard Blastp (http://blast.ncbi.nlm.nih.gov/blast) searches were performed. In case of sequence similarity, there is a feasibility of generation of immune responses against own cells.

15.4.10 Epitope Mapping

Predicting the epitopes binding to MHC class I is the main decisive phase of the RV to carry out valid vaccine predictions. The epitopes showing their affinity for T-cells were first selected via IEDB (http://tools.immuneepitope.org/mhci/), ProPred-I (http://www.imtech.res.in/raghava/propred1/; Singh and Raghava 2003), BIMAS (http://www-bimas.cit.nih.gov/molbio/hla_bind/; Parker et al. 1994) and NetCTL tools (http://www.cbs.dtu.dk/services/NetCTL/; Larsen et al. 2005). For the epitope to be included in the hit list, it must be predicted by any three of these four mentioned tools. For making the predictions of B-cell epitopes, BepiPred (http://www.cbs.dtu.dk/services/BepiPred/; Larsen et al. 2006) and ABCPred tools (http://www.imtech.res.in/raghava/abcpred/ABC_submission.html; Saha and Raghava 2006a, b) were used. The overlapping B-cell and T-cell epitopes were identified.

15.4.11 Docking of the Predicted Epitopes with HLA-A*0201

The predicted epitopes were docked with receptor that is HLA-A*0201 using ClusPro (http://cluspro.bu.edu/login.php; Kozakov et al. 2017) that is an automated protein-protein docking web server. The literature searches provided the information of conserved residues of the receptor site. The default parameters were used for docking (Comeau et al. 2004a, b; Kozakov et al. 2006).

15.5 Results and Discussion

15.5.1 Retrieval of Proteome from NCBI

A total of 40 different sequenced strains of coronavirus are available at NCBI. Among them 7 strains are pathogenic to humans. Various information regarding source, host and collection of these strains are presented in Table 15.1 and 15.2. This information can be obtained from NCBI's genome database, the Virus Pathogen Database and Analysis Resource and Genomes OnLine Database (Liolios et al. 2006; Pickett et al. 2012). The MERS strain is taken as seed genome as it is the most prevalent and disastrous strain among others. Its proteome consists of total 11 proteins as shown in Table 15.3. The results of sequence similarity to reveal orthologs using Blastp are shown in Table 15.4. The sequences with greater than 30% identity score are considered as homologs. The phylogenetic tree is depicted in Fig. 15.1 and the MERS-CoV, taken as seed genome, found clustered with different Bat coronaviruses.

15.5.2 Analysis of Secondary Structure

The results of analysis of secondary structure of the proteome using ExPASy tools are shown in the Table 15.5. From the analysis of charge on the residues and pH values, it is concluded that six of the proteins are basic and positively charged unlike allergens which are acidic in nature. However, five proteins are acidic and show negative charge. The negative GRAVY score of five proteins justify them to be of hydrophilic nature with majority of the residues positioned towards the surface. For the rest of six proteins, the GRAVY score is positive; it means that these are

Table 15.1 Information of coronavirus strains available at NCBI

S. no.	Infected host	Coronavirus strains	Genome#	No of proteins	S. no.	Infected host	Coronavirus strains	Genome#	No of proteins
1	Human	Human coronavirus 229E	NC_002645	8	21	Thrush	Thrush coronavirus HKU12–600	NC_011549	9
2	Human	SARS coronavirus	NC_004718	14	22	Munia	Munia coronavirus HKU13–3514	NC_011550	9
3	Human	Human coronavirus OC43	NC_005147	9	23	Rat	Rat coronavirus Parker	NC_012936	10
4	Human	Human coronavirus NL63	NC_005831	6	24	Bovine	Bovine respiratory coronavirus AH187	NC_012948	12
5	Human	Human coronavirus HKU1	NC_006577	8	25	Bovine	Bovine respiratory coronavirus bovine/US/OH-440-TC/1996	NC_012949	11
6	Human	Human enteric coronavirus strain 4408	NC_012950	12	26	Bat	Bat coronavirus BM48–31/BGR/2008	NC_014470	9
7	Human	Middle East respiratory syndrome coronavirus	NC_019843	11	27	Porcine	Porcine coronavirus HKU15	NC_016990	7
8	Bovine	Bovine coronavirus	NC_003045	12	28	White-eye bird	White-eye coronavirus HKU16	NC_016991	8
9	Bat	Bat coronavirus (BtCoV/133/2005)	NC_008315	8	29	Sparrow	Sparrow coronavirus HKU17	NC_016992	8
10	Bat	Tylonycteris bat coronavirus HKU4	NC_009019	9	30	Magpie-robin	Magpie-robin coronavirus HKU18	NC_016993	9
11	Bat	Pipistrellus bat coronavirus HKU5	NC_009020	9	31	Night-heron	Night-heron coronavirus HKU19	NC_016994	8
12	Bat	Rousettus bat coronavirus HKU9	NC_009021	8	32	Wigeon	Wigeon coronavirus HKU20	NC_016995	10

(continued)

Table 15.1 (continued)

S. no.	Infected host	Coronavirus strains	Genome#	No of proteins	S. no.	Infected host	Coronavirus strains	Genome#	No of proteins
13	Bat	Scotophilus bat coronavirus 512	NC_009657	6	33	Moorhen	Common moorhen coronavirus HKU21	NC_016996	9
14	Bat	Rhinolophus bat coronavirus HKU2	NC_009988	8	34	Rabbit	Rabbit coronavirus HKU14	NC_017083	11
15	Equine	Equine coronavirus	NC_010327	11	35	Bat	Rousettus bat coronavirus HKU10	NC_018871	9
16	Bat	Bat coronavirus 1B	NC_010436	7	36	Bat	Bat coronavirus CDPHE15/USA/2006	NC_022103	7
17	Bat	Bat coronavirus 1A	NC_010437	7	37	*Erinaceus*	Betacoronavirus Erinaceus/VMC/DEU/2012	NC_022643	12
18	Bat	Miniopterus bat coronavirus HKU8	NC_010438	8	38	Mink	Mink coronavirus strain WD1127	NC_023760	10
19	Whale	Beluga whale coronavirus SW1	NC_010646	14	39	Bat	Bat Hp-betacoronavirus/Zhejiang2013	NC_025217	9
20	Turkey	Turkey coronavirus	NC_010800	11	40	Birds	Betacoronavirus HKU2	NC_026011	10

Table 15.2 Detail information about seven strains of coronavirus which are pathogenic to humans

S. no.	Strain name	Acc. No.	Source information	Length	Proteins	Host	Collection date	Sequencing centre	Sequencing strategy	NCBI BioProject ID	Culture type	Host taxonomy ID	Completion date	GC %
1.	*Human coronavirus 229E*	NC_002645	Strain:229E	27,317 nt	8	Vertebrates, human	1/11/2001	University of Wurzburg (Germany)	Whole genome sequencing	14913	Isolate	9606	10/2/2006	38.3
2.	*SARS coronavirus*	NC_004718	Isolate:Tor2	29,751 nt	14	Vertebrates, human	4/14/2003	National Microbiology Laboratory, Public Health Agency of Canada	Whole genome sequencing	15500	Isolate	196,296	10/2/2006	40.8
3.	*Human coronavirus OC43*	NC_005147	Strain: ATCC VR-759; isolate: OC43	30,738 nt	9	Vertebrates, human	10/29/2003	University of Leuven – KU Leuven (Belgium)	Whole genome sequencing	15438	Isolate	9606	10/2/2006	36.8
4.	*Human coronavirus NL63*	NC_005831	Strain: Amsterdam I	27,553 nt	6	Vertebrates, human	3/23/2004	University of Amsterdam (Netherlands)	Whole genome sequencing	14960	Isolate	9606	10/2/2006	34.5
5.	*Human coronavirus HKU1*	NC_006577	Isolate: HKU1	29,926 nt	8	Vertebrates, human	12/28/2004	University of Hong Kong (Hong Kong)	Whole genome sequencing	15139	Isolate	9606	10/2/2006	32.1

(continued)

Table 15.2 (continued)

S. no.	Strain name	Acc. No.	Source information	Length	Proteins	Host	Collection date	Sequencing centre	Sequencing strategy	NCBI BioProject ID	Culture type	Host taxonomy ID	Completion date	GC %
6.	*Human enteric coronavirus strain 4408*	NC_012950	Strain:4408	30,953 nt	12	Vertebrates, human	7/13/2009	–	Whole genome sequencing	39335	Isolate	9606	10/13/2009	37.1
7.	Middle East respiratory syndrome coronavirus	NC_019843	Strain: HCoV-EMC	30,119 nt	11	Vertebrates	12/13/2012	Health Protection Agency (HPA), UK (United Kingdom)	Whole genome sequencing	183710	Isolate	9606	12/8/2014	41.2

Table 15.3 Information of MERS-CoV (NC_019843.3) proteins

S. no.	GeneID	Start	Stop	Locus	Locus tag	Protein Accession	Length	Protein name
1.	14254602	279	13454	orf1ab	G128_gp01	YP_009047203.1	4391	1A polyprotein
2.	14254602	279	21514	orf1ab	G128_gp01	YP_009047202.1	7078	1ab polyprotein
3.	14254594	21456	25517	S	G128_gp02	YP_009047204.1	1353	Spike glycoprotein
4.	14254595	25532	25843	orf3	G128_gp03	YP_009047205.1	103	NS3 protein
5.	14254596	25852	26181	orf4a	G128_gp04	YP_009047206.1	109	NS4A protein
6.	14254597	26093	26833	orf4b	G128_gp05	YP_009047207.1	246	NS4B protein
7.	14254598	26840	27514	orf5	G128_gp06	YP_009047208.1	224	NS5 protein
8.	14254599	27590	27838	E	G128_gp07	YP_009047209.1	82	Envelope protein
9.	14254600	27853	28512	M	G128_gp08	YP_009047210.1	219	Membrane protein
10.	14254601	28566	29807	N	G128_gp09	YP_009047211.1	413	Nucleoprotein
11.	19910005	28762	29100	orf8b	G128_gp10	YP_009047212.1	112	ORF8b protein

Table 15.4 Results of Homology searches of the proteins of seed genome using Blastp

			SARS coronavirus		*Human coronavirus OC43*		Human coronavirus NL63		Human coronavirus HKU1		Human enteric coronavirus strain 4408		Human coronavirus 229E	
S. no.	Accession no.	Protein name	Accession no.	Identity	Accession no.	Identity	Accession no.	Identity	Accession no.	Identity	Accession no.	Identity	Accession no.	Identity
1	YP_009047202.1	1ab polyprotein	NP_828849.2| orf1ab polyprotein	42.89%	NP_937947.2| replicase polyprotein	48.12%	YP_003766.2| replicase polyprotein	45.49%	YP_173236.1| orf1ab polyprotein	48.23%	YP_003038518.1| orf1ab polyprotein	35.83%	NP_073549.1| replicase polyprotein	33.76%
			NP_828850.1| orf1a polyprotein	45.94%	–	–	–	–	YP_173236.1| orf1ab polyprotein	55.34%	–	–	–	–
2	YP_009047203.1	1A polyprotein	NP_828850.1| orf1a polyprotein	42.62%	NP_937947.2| replicase polyprotein	35.85%	YP_003766.2| replicase polyprotein 1ab	33.87%	YP_173236.1| orf1ab polyprotein	36.10%	YP_003038519.1| orf1a polyprotein	35.82%	NP_073550.1| replicase polyprotein	33.76%
3	YP_009047204.1|	Spike glycoprotein	–	–	–	–	–	–	YP_173238.1| spike glycoprotein	41.28%	YP_003038522.1| spike protein	37.72%	NP_073551.1| surface glycoprotein	35.79%
4	YP_009047205.1	NS3 protein	–	–	–	–	–	–	–	–	–	–	–	–
5	YP_009047206.1	NS4A protein	–	–	–	–	–	–	–	–	–	–	–	–
6	YP_009047207.1	NS4B protein	–	–	–	–	–	–	–	–	–	–	–	–
7	YP_009047208.1	NS5 protein	–	–	–	–	–	–	–	–	–	–	–	–
8	YP_009047209.1	Envelope protein	–	–	–	–	–	–	–	–	–	–	–	–
9	YP_009047210.1	Membrane protein	NP_828855.1| matrix protein	45.45%	NP_937953.1| M protein	33.81%	–	–	–	–	YP_003038527.1| membrane protein	44.39%	NP_073555.1| membrane protein	34.52%
10	YP_009047211.1	Nucleoprotein	NP_828858.1| nucleocapsid protein	38.69%	–	–	–	–	YP_173242.1| nucleocapsid	40.83%	YP_003038528.1| nucleocapsid	40.83%	–	–
11	YP_009047212.1	ORF8b protein	–	–	–	–	–	–	–	–	–	–	–	–

The accession number and identity of orthologs obtained in different strains is shown in the table

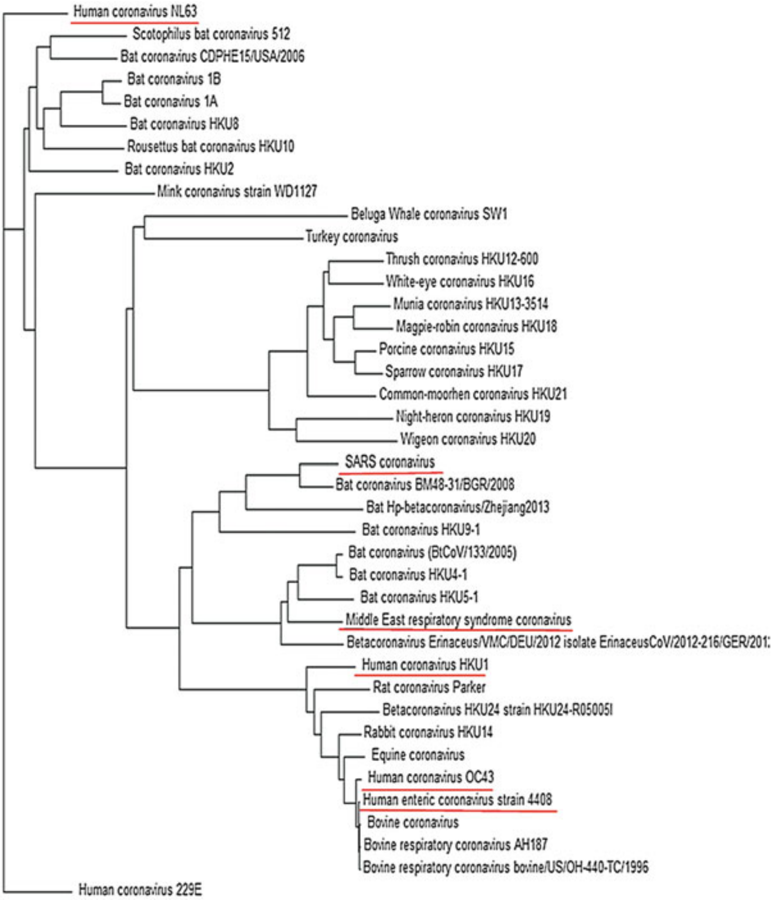

Human coronavirus NL63
Scotophilus bat coronavirus 512
Bat coronavirus CDPHE15/USA/2006
Bat coronavirus 1B
Bat coronavirus 1A
Bat coronavirus HKU8
Rousettus bat coronavirus HKU10
Bat coronavirus HKU2
Mink coronavirus strain WD1127
Beluga Whale coronavirus SW1
Turkey coronavirus
Thrush coronavirus HKU12-600
White-eye coronavirus HKU16
Munia coronavirus HKU13-3514
Magpie-robin coronavirus HKU18
Porcine coronavirus HKU15
Sparrow coronavirus HKU17
Common-moorhen coronavirus HKU21
Night-heron coronavirus HKU19
Wigeon coronavirus HKU20
SARS coronavirus
Bat coronavirus BM48-31/BGR/2008
Bat Hp-betacoronavirus/Zhejiang2013
Bat coronavirus HKU9-1
Bat coronavirus (BtCoV/133/2005)
Bat coronavirus HKU4-1
Bat coronavirus HKU5-1
Middle East respiratory syndrome coronavirus
Betacoronavirus Erinaceus/VMC/DEU/2012 isolate ErinaceusCoV/2012-216/GER/201:
Human coronavirus HKU1
Rat coronavirus Parker
Betacoronavirus HKU24 strain HKU24-R05005I
Rabbit coronavirus HKU14
Equine coronavirus
Human coronavirus OC43
Human enteric coronavirus strain 4408
Bovine coronavirus
Bovine respiratory coronavirus AH187
Bovine respiratory coronavirus bovine/US/OH-440-TC/1996
Human coronavirus 229E

Fig. 15.1 Phylogenetic tree of 40 different strains of coronavirus using whole genome sequences (Alignment of genome sequences is done using ClustalW, and tree is created using NJ method from Unipro UGENE 1.15.1 bioinformatics toolkit)

hydrophobic proteins. The proteins with less than 40 value of instability index are quite stable than those with higher values. All the proteins are having the molecular weight less than 110 kDa except 3 (YP_009047202.1, YP_009047203.1 and YP_009047204.1). This exhibits the effectiveness of lightweight proteins as targets as they can be easily purified because of their low molecular weights. The protein YP_009047204.1 is reported as a spike glycoprotein. It is acidic with prominent negative charge, with negative GRAVY score which suggests its hydrophilicity and

Table 15.5 Secondary structure analysis of MERS-CoV proteins

S. no	Accession no.	Molecular weight	Theoretical PI	GRAVY	Aliphatic Index	Instability index	Estimated half life	Extinction coefficient	Positive charged	Negative charged
1	YP_009047202.1	789461.2	6.47	0.013	88.03	34.24	30	978,520	647	687
2	YP_009047203.1	485956.4	6.28	0.081	91.54	34.07	30	575,415	385	416
3	YP_009047204.1	149,368	5.7	−0.074	82.71	36.6	30	170,865	95	112
4	YP_009047205.1	11137.5	5.39	−0.022	86.02	25.54	30	6085	6	9
5	YP_009047206.1	12242.8	6.25	−0.263	83.3	47.02	30	27,055	8	9
6	YP_009047207.1	28,603	9.74	−0.228	94.23	51.09	30	34,505	29	18
7	YP_009047208.1	25280.1	9.26	0.807	124.33	51.4	30	19,410	17	9
8	YP_009047209.1	9354.2	7.64	0.795	111.59	33	30	10,220	5	4
9	YP_009047210.1	24536.8	9.27	0.436	104.61	43.67	30	53,525	16	11
10	YP_009047211.1	45062.3	10.05	−0.866	57	48.62	30	47,900	55	33
11	YP_009047212.1	12279.6	10.15	0.131	113.21	49.65	30	5500	11	7

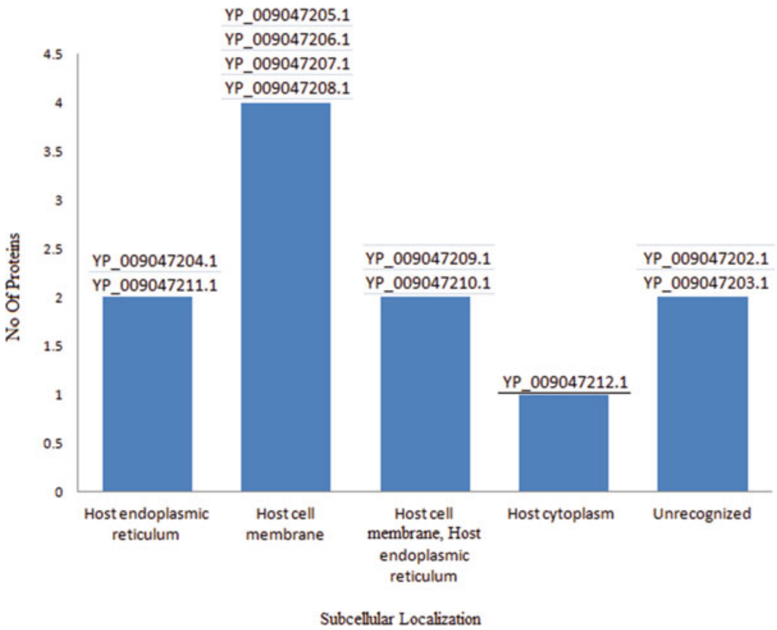


Fig. 15.2 Subcellular localization of seed genome proteins predicted using Virus-mPLoc

presence on surface. However the envelope protein YP_009047209.1 and membrane protein YP_009047210.1 are basic and hydrophobic.

15.5.3 Subcellular Localization Predictions

Figure 15.2 depicts the subcellular localization of proteins of the seed genome, i.e. MERS-CoV. Only one protein was predicted to be localized in host cytoplasm, four in host membrane, two in both host cell membrane and endoplasmic reticulum (ER) while two in only ER, and two are left unrecognized. The known spike protein is predicted to be localized in host ER. From these results we decided to pick the proteins which are located in host membrane or were predicted to be localized in both host membrane and ER. The two are known envelop protein and membrane protein from bibliographic studies, and along with that, the known spike protein was also included in the filtered results. Out of the filtered proteins, only two (YP_009047210.1 and YP_009047208.1) contain more than two transmembrane helices, therefore filtered out. The results of transmembrane helices prediction are tabulated in Table 15.6. Figure 15.3 depicts the subcellular localization of proteins of all the four selected genomes using Virus-mPLoc prediction tool.

Table 15.6 Subcellular Localization prediction results using Virus-mPloc

Subcellular localization	Human coronavirus NL63	Human coronavirus HKU1	Human enteric coronavirus strain 4408	Human coronavirus 229E	Human coronavirus OC43	SARS coronavirus	MERS coronavirus
Host cell membrane. Host endoplasmic reticulum	YP_003769.1	–	YP_003038527.1	–	–	–	YP_009047209.1
	YP_003770.1						YP_009047210.1
Host cell membrane. Host cytoplasm	YP_003768.1						
Host endoplasmic reticulum	YP_003767.1	YP_173238.1	YP_003038522.1	NP_073551.1	NP_937950.1	NP_828851.1	YP_009047204.1
	–	YP_173241.1	–	NP_073554.1	NP_937953.1	NP_828856.1	YP_009047211.1
		–		–	–	NP_828857.1	–
						NP_828858.1	
Host cell membrane	–	YP_173240.1	YP_003038526.1	NP_073552.1	NP_937952.1	NP_828852.2	YP_009047205.1
		–	–	NP_073555.1	–	NP_828854.1	YP_009047206.1
				–		NP_828855.1	YP_009047207.1
						–	YP_009047208.1
Host cytoplasm	-	YP_173239.1	YP_003038520.1	NP_073553.1	NP_937948.1	NP_828853.1	YP_009047212.1
		YP_173243.1	YP_003038525.1	–	NP_937951.1	NP_849177.1	–
		–	YP_003038529.1		NP_937955.1	NP_828859.1	
Host endoplasmic reticulum. Host cytoplasm	YP_003771.1	YP_173237.1	YP_003038521.1	NP_073556.1	NP_937949.1	–	–
	–	YP_173242.1	YP_003038528.1	–	NP_937954.1		
		–	–		–		
Unrecognized	YP_003766.2	YP_173236.1	YP_003038518.1	NP_073549.1	NP_937947.2	NP_828849.2	YP_009047202.1
	–	–	YP_003038519.1	NP_073550.1	–	NP_828850.1	YP_009047203.1
			YP_003038523.1	–		NP_849175.1	–
			YP_003038524.1			NP_849176.1	

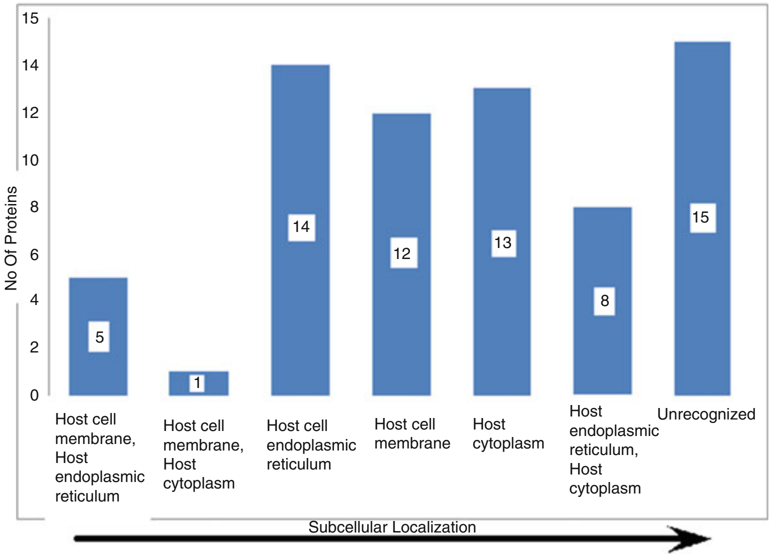


Fig. 15.3 Subcellular localization of proteins of all four selected genomes predicted using Virus-mPLoc

15.5.4 Signal Peptides

The proteins that are predicted to possess the signal peptides by Signal-BLAST web server are YP_009047204.1 and YP_009047205.1. The results of Signal-BLAST web server are tabulated in the Table 15.7.

15.5.5 Adhesion Probability

This step takes into account the concept of adhesion-based virulence. Adhesions cause pathogen recognition and initiation of inflammatory responses by the host. SPAAN predicted 2 (YP_009047204.1 and YP_009047205.1) out of 11 proteins of MERS strain as adhesive (Table 15.8).

15.5.6 BetaWrap

Only one protein (YP_009047204.1) was predicted to contain BetaWrap motifs within it (Table 15.8). Hence, it is considered virulent and might be responsible for initializing the infection in the host.

Table 15.7 The signal peptide prediction results for proteins of MERS coronavirus strain

S. no.	Accession no.	Signal blast (Sensitivity)	Specificity	Balanced prediction	Cleavage site
1	YP_009047202.1	No	No	No	Yes
2	YP_009047203.1	No	No	No	Yes
3	YP_009047204.1	Yes	Yes	Yes	Yes
4	YP_009047205.1	Yes	Yes	Yes	Yes
5	YP_009047206.1	No	No	No	Yes
6	YP_009047207.1	No	No	No	Yes
7	YP_009047208.1	No	No	No	Yes
8	YP_009047209.1	No	No	No	No alignment found, unable to predict
9	YP_009047210.1	No	No	No	No alignment found, unable to predict
10	YP_009047211.1	No	No	No	Yes
11	YP_009047212.1	No	No	No	No alignment found, unable to predict

Table 15.8 Table illustrating the prediction results made for selecting adhesion proteins using SPAAN, BetaWrap predictions and antigenicity predictions using Vaxijen *version* 2.0

S. no	Accession no.	Adhesion probability	P-value	Vaxijen value	TMHMM
1	YP_009047202.1	0.439813	No	0.4908	14
2	YP_009047203.1	0.442577	No	0.4884	14
3	YP_009047204.1	0.634711	0.0046	0.4849	1
4	YP_009047205.1	0.635586	No	0.4226	0
5	YP_009047206.1	0.44212	No	0.3288	0
6	YP_009047207.1	0.269269	No	0.4978	0
7	YP_009047208.1	0.237608	No	0.3369	3
8	YP_009047209.1	0.389879	No	0.5119	1
9	YP_009047210.1	0.461965	No	0.5503	3
10	YP_009047211.1	0.690125	No	0.6036	0
11	YP_009047212.1	0.342692	No	0.6078	0

The transmembrane prediction results using TMHMM are also tabulated

15.5.7 VaxiJen *2.0*

A total of 9 out of 11 proteins of MERS strain were predicted antigenic (prediction values greater than 0.4). The protein with accession number YP_009047206.1 and YP_009047208.1 were among the filtered proteins, however, not predicted antigenic, therefore filtered out. As a result, only four proteins (YP_009047204.1, YP_009047205.1, YP_009047207.1 and YP_009047209.1) were kept for further analyses.

15.5.8 AlgPred and Allertop

None of the 11 proteins of MERS-CoV possessed any clue of allergenicity as per prediction results from AlgPred and Allertop tools; it means that no vigorous immune responses will be mounted if the epitopes from these proteins will be adopted as vaccine candidates.

15.5.9 Similarity with Host Proteome

None of the protein of MERS strain shows similarity with the proteins of host that demonstrates that the epitopes from these proteins can safely elicit the required immune response without the hazard of autoimmunity.

15.5.10 Epitope Mapping

In total 12 different 9-mer epitopes with potential to bind to receptors of both B-cell and T-cell were predicted. The list of the predicted epitopes can be found in the Table 15.9 and are specific for MERS-CoV strain. All these epitopes displayed no conservancy with proteins of other human and non-human pathogenic strains.

15.5.11 Docking Analysis

Docking permits to reveal the binding energy or potency of connection among epitopes and the receptor in appropriate orientation. The ClusPro docking server was used to dock the predicted 90 epitopes against HLA-A*0201. The structure of the receptor was available from PDB and was optimized before docking to free it from the complexed self-peptide (4U6Y, Resolution 1.47 Å, Bouvier et al. 1998). PEPstr (Peptide Tertiary Structure Prediction Server; Kaur et al. 2007) was used to derive the tertiary structure of the predicted peptides.

Table 15.9 The results of overlapping T-cell and B-cell epitope predictions for four filtered proteins

S. no	Accession no	Starting position	9-mer epitope	Binding energy	H-bonding	Receptor residues involved in H-bonding
1	YP_009047204.1	18	YVDVGPDSV	−854.8	13	Lys-66, 146, His-70, Thr-73, Arg-97, Trp-147, Tyr-159,59,99, Glu-63, Asp-3
		110	KQFANGFVV	−764.6	9	Lys-66, His-70, Glu-63,166, Thr-163, Tyr-99,159
		160	KMGRFFNHT	−849.7	11	Lys-66, Arg-97, Tyr-159,171, Glu-166,63, Gln-155, His-70
		388	LLSGTPPQV	−713.1	10	Lys-66,146, Arg-97, Trp-147, Tyr-159,99, Gln-155, Asp-77
		553	WLVASGSTV	−843.6	13	Lys-66,Thr-73,163, Arg-97, Trp-147,167, Asp-77, Tyr-99, Glu-63
		630	FVYDAYQNL	−785.6	11	Thr-73, Arg-97, Lys-146, Trp-147, Gln-155, Asp-77, Gln-155
		716	GLVNSSLFV	−887.1	14	Tyr-59,159,99, Lys-66,146, Arg-97, Gln-155, Glu-63
		1275	ALNESYIDL	−650.1	11	Lys-66,146, Thr-73, Arg-75,97, Trp-147, Gln-155, Tyr-159
2	YP_009047205.1	18	LVTASSKPL	−813.9	9	Lys-66, Arg-97, Trp-147, Gln-155, Glu-63, Tyr-159, Thr-73, Asp-77
3	YP_009047207.1	187	KLHALDDVT	−649.3	12	Lys-66,146, Tyr-116, Trp-147, Glu-63
4	YP_009047209.1	21	VVCAITLLV	−951.7	12	Lys-66,146, Arg-97, Tyr-116,99, Trp-147,167, Thr-163, Glu-166,63
		27	TLLVCMAFL	−916	7	Lys-66, Trp-167, Glu-63, Tyr-159,99,Gln-155, Thr-163

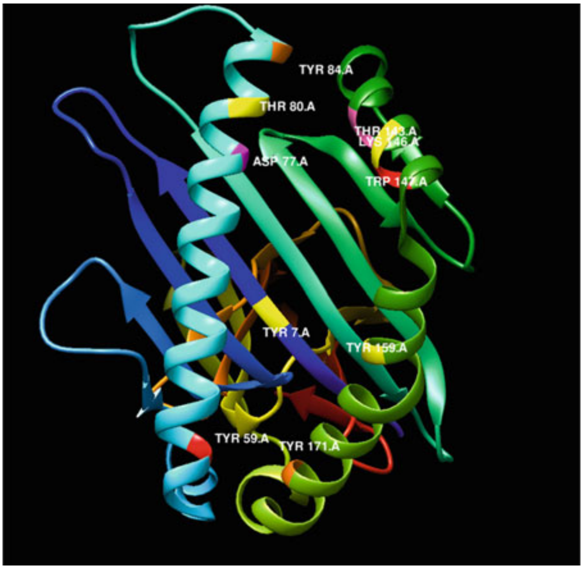


Fig. 15.4 3D structure of receptor site of HLA-A*0201 visualized using Swiss PDB viewer 4.10. The residues shown in globular structure are known to be conserved and form hydrogen bonds with the binding peptides

Figure 15.4 depicts the quaternary structure of the receptor HLA-A*0201 with its conserved active site known to form complex with the peptides (Bouvier et al. 1998). The binding energy results obtained after performing docking analysis are listed in Table 15.9.

The 9-mer epitope VVCAITLLV at site 21 of protein YP_009047209.1 docked to the receptor with smallest amount of binding energy (−951.7) and 12 hydrogen bonds. The next epitope in the list was also from the same protein YP_009047209.1 at site 27, i.e. TLLVCMAFL. The predicted structure of the top 5 potent epitopes on the basis of docking energy and the snapshots of docking results are displayed in Figs. 15.5, 15.6, 15.7, 15.8 and 15.9.

The most chief restriction for developing a safe and sound vaccine against any of the virus is to identify the protective antigens. The present study is an effort of application of reverse vaccinology approach to investigate a choice of coronavirus proteomes to identify possible vaccine targets. This technique has demonstrated to be a competent way to forecast 12 different epitopes from the selected seed genome. These epitopes are from spike glycoprotein, NS3 protein, NS4B protein and envelope protein. Unfortunately none of the epitope is found conserved in other strains, and all are specific to MERS-CoV. The docking analysis studies revealed perfect binding between HLA-A*0201 receptor and epitopes. The conserved residues of the receptor site are also involved in H-bonding with epitope residues. Further, the selected antigenic epitopes must be validated using in vitro and in vivo studies to confirm their potential as vaccine candidates.

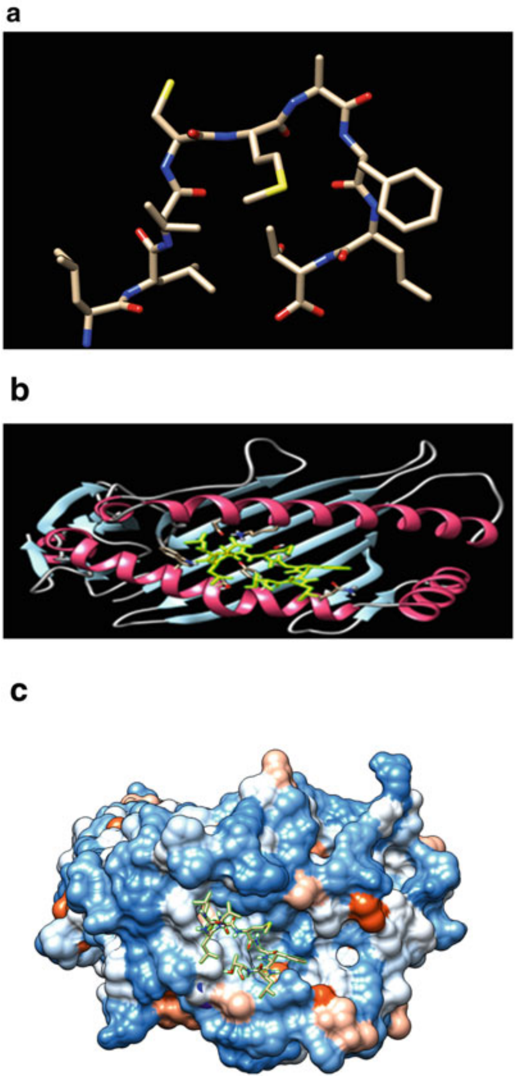


Fig. 15.5 (**a**) 3D Structure of the 9-mer epitope starting from 21(VVCAITLLV) position of protein YP_009047209.1 (**b**) Docking results of epitope “VVCAITLLV” with A chain of HLA-A*0201 using ClusPro. (**c**) The snapshot representing the epitope docked in the pocket of molecular surface of the receptor (all the structures are visualized using Chimera 1.10.1)

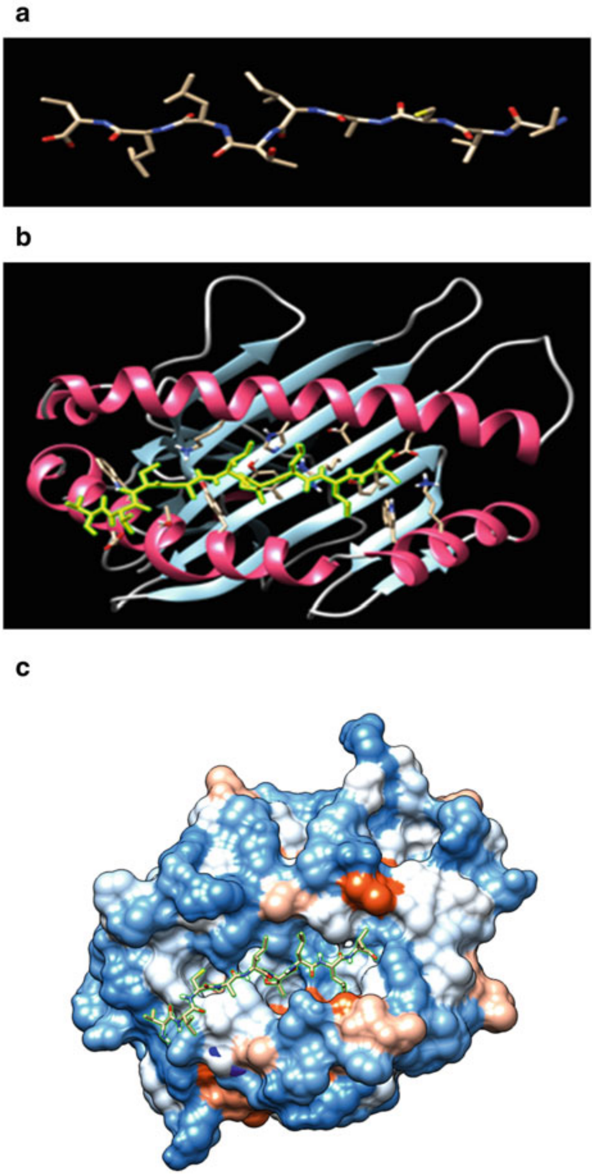


Fig. 15.6 (**a**) 3D Structure of the 9-mer epitope starting from 27(TLLVCMAFL) position of protein YP_009047209.1. (**b**) Docking results of epitope “TLLVCMAFL” with A chain of HLA-A*0201 using ClusPro. (**c**) The snapshot representing the epitope docked in the pocket of molecular surface of the receptor (all the structures are visualized using Chimera 1.10.1)

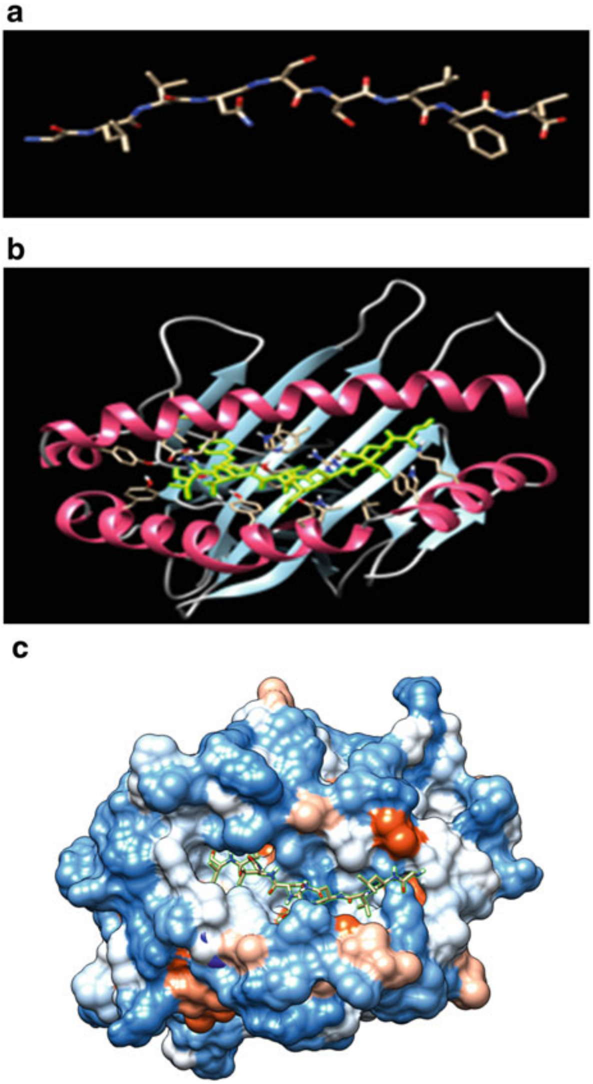


Fig. 15.7 (**a**) 3D Structure of the 9-mer epitope starting from 716(GLVNSSLFV) position of protein YP_009047204.1. (**b**) Docking results of epitope "GLVNSSLFV" with A chain of HLA-A*0201 using ClusPro. (**c**) The snapshot representing the epitope docked in the pocket of molecular surface of the receptor (all the structures are visualized using Chimera 1.10.1)

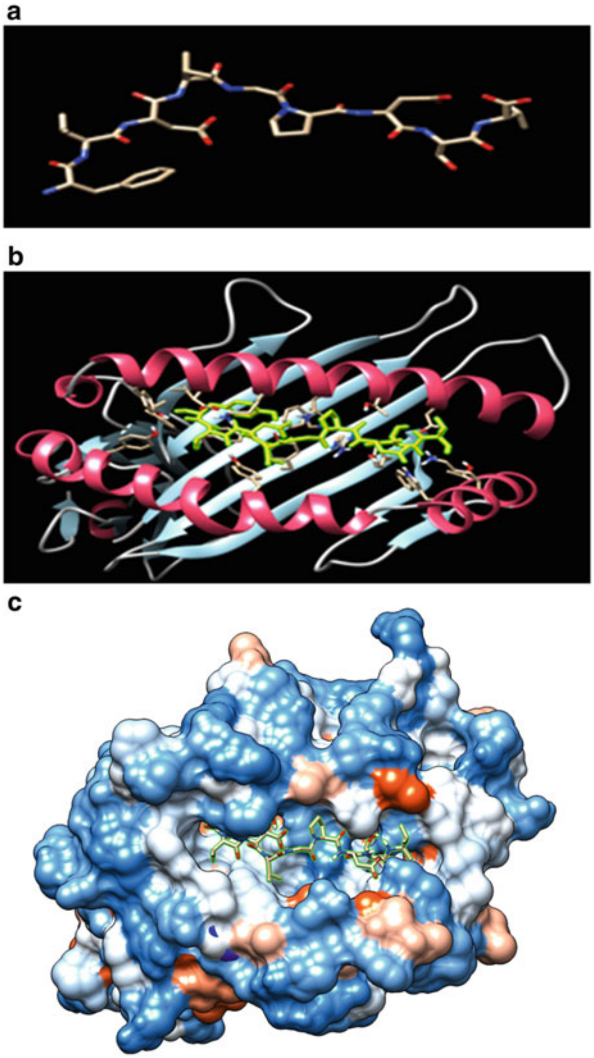


Fig. 15.8 (**a**) 3D Structure of the 9-mer epitope starting from 18(YVDVGPDSV) position of protein YP_009047204.1. (**b**) Docking results of epitope “YVDVGPDSV” with A chain of HLA-A*0201 using ClusPro. (**c**) The snapshot representing the epitope docked in the pocket of molecular surface of the receptor (all the structures are visualized using Chimera 1.10.1)

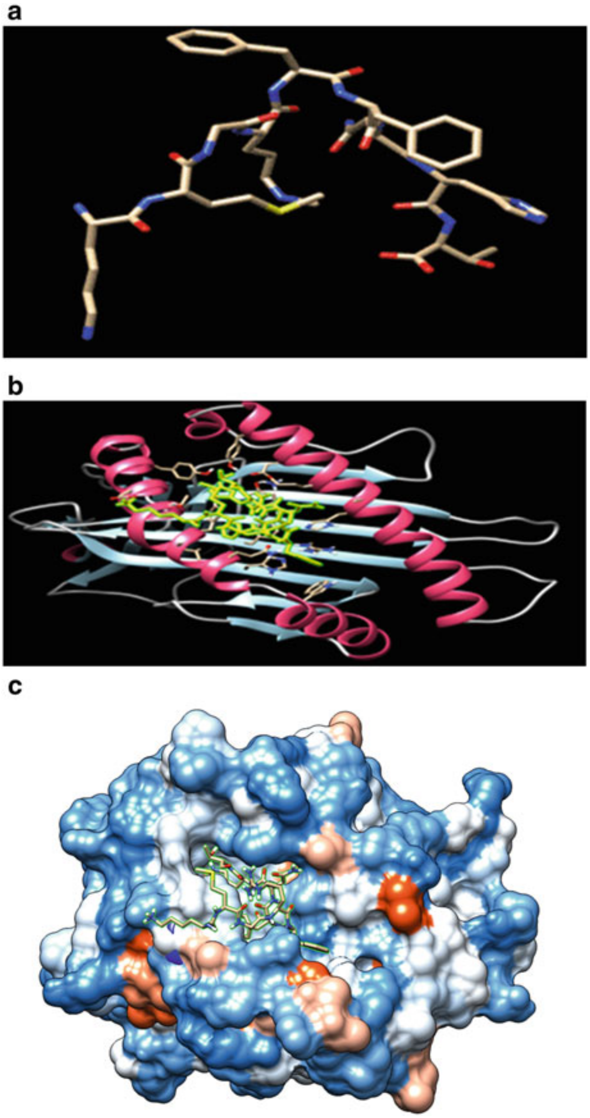


Fig. 15.9 (**a**) 3D Structure of the 9-mer epitope starting from 160(KMGRFFNHT) position of protein YP_009047204.1. (**b**) Docking results of epitope "KMGRFFNHT" with A chain of HLA-A*0201 using ClusPro. (**c**) The snapshot representing the epitope docked in the pocket of molecular surface of the receptor (all the structures are visualized using Chimera 1.10.1)

References

Alessandro S, Rino R (2010) Review: reverse vaccinology: developing vaccines in the era of genomics. Immunity 33:530–541. https://doi.org/10.1016/j.immuni.2010.09.017

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Azhar EI, Lanini S, Ippolito G, Zumla A (2017) The Middle East respiratory syndrome coronavirus – a continuing risk to global health security. Adv Exp Med Biol 972:49–60. https://doi.org/10.1007/5584_2016_133.PMID:27966107

Bouvier M, Guo HC, Smith KJ, Wiley DC (1998) Crystal structures of HLA-A*0201 complexed with antigenic peptides with either the amino- or carboxyl-terminal group substituted by a methyl group. Proteins Struct Funct Genet 33(1):97–106. PMID: 9741848

Bradley P, Cowen L, Menke M, King J, Berger B (2001) BETAWRAP: successful prediction of parallel beta -helices from primary sequence reveals an association with many microbial pathogens. Proc Natl Acad Sci U S A 98(26):14819–14824. https://doi.org/10.1073/pnas.251267298

Brown JK, Fauquet CM, Briddon RW, Zerbini M, Moriones E, Navas-Castillo J, King AM, Adams MJ, Carstens EB, Lefkowitz EJ (2012) Geminiviridae. In: Virus taxonomy—ninth report of the International Committee on Taxonomy of Viruses. Elsevier Science, Burlington, pp 351–373

Bruno L, Cortese M, Rappuoli R, Merola M (2015) Lessons from Reverse Vaccinology for viral vaccine design. Curr Opin Virol 11:89–97. https://doi.org/10.1016/j.coviro.2015.03.001

Caro-Gomez E, Gazi M, Goez Y, Valbuena G (2014) Discovery of novel cross-protective Rickettsia prowazekii T-cell antigens using a combined reverse vaccinology and in vivo screening approach. Vaccine 32(39):4968–4976. https://doi.org/10.1016/j.vaccine.2014.06.089

Chiang MH, Sung WC, Lien SP, Chen YZ, Lo AF, Huang JH, Kuo SC, Chong P (2015) Identification of novel vaccine candidates against Acinetobacter baumannii using reverse vaccinology. Hum Vaccin Immunother 11(4):1065–1073. https://doi.org/10.1080/21645515.2015.1010910

Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004a) ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res 32(suppl_2):W96–W99. https://doi.org/10.1093/nar/gkh354

Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004b) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 20(1):45–50. https://doi.org/10.1093/bioinformatics/btg371

De Groot RJ, Baker SC, Baric RS et al (2013) Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the coronavirus study group. J Virol 87(14):7790–7792. https://doi.org/10.1128/JVI.01244-13

Dimitrov I, Bangov I, Flower DR, Doytchinova I (2014) AllerTOP v.2-a server for in silico prediction of allergens. J Mol Model BioMed Central Ltd 20(6):227. DOI: https://doi.org/10.1007/s00894-014-2278-5

Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinf 8(1):4. https://doi.org/10.1186/1471-2105-8-4

Frank K, Sippl MJ (2008) High performance signal peptide prediction based on sequence alignment techniques. Bioinformatics 24:2172–2176. https://doi.org/10.1093/bioinformatics/btn422

Gabutti G (2014) Meningococcus B: control of two outbreaks by vaccination. J Prev Med Hyg 55 (2):35–41. PMCID: PMC4718325

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD et al (2005) Protein identification and analysis tools on the ExPASy server. Proteomics Protoc Handb:571–607. https://doi.org/10.1385/1-59259-890-0:571

Golosova O, Henderson R, Vaskin Y, Gabrielian A, Grekhov G, Nagarajan V, Oler AJ, Quiñones M, Hurt D, Fursov M, Huyen Y (2014) Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. Peer J 2:e644. https://doi.org/10.7717/peerj.644

Goodswen SJ, Kennedy PJ, Ellis JT (2014) Discovering a vaccine against neosporosis using computers: is it feasible? Trends Parasitol 30(8):401–411. https://doi.org/10.1016/j.pt.2014.06.004

Guimarães L, Soares S, Trost E, Blom J, Ramos R, Silva A, Barh D, Azevedo V (2015) Genome informatics and vaccine targets in Corynebacterium urealyticum using two whole genomes, comparative genomics, and reverse vaccinology. BMC Genomics 16(Suppl 5):S7. https://doi.org/10.1186/1471-2164-16-S5-S7

Kanampalliwar AM, Soni R, Tiwari A, Giridhar A (2013) Reverse vaccinology: basics and applications. J Vaccin Vaccin 4(6):194–198. https://doi.org/10.4172/2157-7560.1000194

Kaur H, Garg A, Raghava GPS (2007) PEPstr: a de novo method for tertiary structure prediction of small bioactive peptides. Protein Pept Lett 14:626–630. https://doi.org/10.2174/092986607781483859

Kelly DF, Rappuoli R (2005) Reverse vaccinology and vaccines for serogroup B Neisseria meningitidis. Adv Exp Med Biol 568:217–223

Kolesanova EF, Sobolev BN, Moysa AA, Egorova EA, Archakov AI (2015) Way to the peptide vaccine against hepatitis C. Biomed Khim 61(2):254–264. https://doi.org/10.18097/PBMC20156102254

Kozakov D, Brenke R, Comeau SR, Vajda S (2006) PIPER: an FFT-based protein docking program with pair wise potentials. Proteins 65(2):392–406. https://doi.org/10.1002/prot.21117

Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, Beglov D, Vajda S (2017) The ClusPro web server for protein-protein docking. Nat Protoc 12(2):255–278

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305 (3):567–580. https://doi.org/10.1006/jmbi.2000.4315

Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, Nielsen M (2005) An integrative approach to CTL epitope prediction, a combined algorithm integrating MHC-I binding, TAP transport efficiency, and proteasomal cleavage prediction. J Immunol 35:2295–2303. https://doi.org/10.1002/eji.200425811

Larsen JEP, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes. Immunome Res 2(1):2. https://doi.org/10.1186/1745-7580-2-2

Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The genomes on line database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 34:D332–D334. https://doi.org/10.1093/nar/gkj145

Monterrubio-López GP, González-Y-Merchand JA, Ribas-Aparicio RM (2015) Identification of novel potential vaccine candidates against tuberculosis based on reverse vaccinology. Biomed Res Int 2015:483150. https://doi.org/10.1155/2015/483150

Naz A, Awan FM, Obaid A, Muhammad SA, Paracha RZ, Ahmad J, Ali A (2015) Identification of putative vaccine candidates against Helicobacter pylori exploiting exoproteome and secretome: a reverse vaccinology based approach. Infect Genet Evol 32:280–291. https://doi.org/10.1016/j.meegid.2015.03.027

NCBI Resource Coordinators (2017) Database resources of the national center for biotechnology information. Nucleic Acids Res 45(D1):D12–D17. https://doi.org/10.1093/nar/gkw1071

Okonechnikov K, Golosova O, Fursov M, UGENE team (2012) Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28:1166–1167. https://doi.org/10.1093/bioinformatics/bts091

Parker KC, Bednarek MA, Coligan JE (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 152 (1):163–75. PMID: 8254189

Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G et al (2012) Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses 11:3209–3226. https://doi.org/10.3390/v4113209

Pizza M, Scarlato V, Masignani V et al (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 287:1816–1820. https://doi.org/10.1126/science.287.5459.1816

Rappuoli R (2000) Reverse vaccinology. Curr Opin Microbiol 3:445–450. https://doi.org/10.1016/S1369-5274(00)00119-3

Sachdeva G, Kumar K, Preti J, Ramachandran S (2004) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. Bioinformatics 21:483–491. https://doi.org/10.1093/bioinformatics/bti028

Saha S, Raghava GPS (2006a) AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res 34:W202–W209. https://doi.org/10.1093/nar/gkl343

Saha S, Raghava GPS (2006b) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 65(1):40–48. https://doi.org/10.1002/prot.21078

Sette A, Rappuoli R (2010) Reverse vaccinology: developing vaccines in the era of genomics. Immunity 33(4):530–541. https://doi.org/10.1016/j.immuni.2010.09.017

Shen H-B, Chou K-C (2010) Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. J Biomol Struct Dyn 28:175–186. https://doi.org/10.1080/07391102.2010.10507351

Singh H, Raghava GPS (2003) ProPred1: prediction of promiscuous MHC class-I binding sites. Bioinformatics 19(8):1009–1014. https://doi.org/10.1093/bioinformatics/btg108

Vishnu US, Sankarasubramanian J, Gunasekaran P, Rajendhran J (2017) Identification of potential antigens from non-classically secreted proteins and designing novel multitope peptide vaccine candidate against Brucella melitensis through reverse vaccinology and immunoinformatics approach. Infect Genet Evol 55:151–158

Machine Learning: What, Why, and How? 16

Salma Jamal, Sukriti Goyal, Abhinav Grover,
and Asheesh Shanker

16.1 Introduction

Machine learning involves a set of algorithms which deal with the automatic recognition of hidden patterns in data and making predictions about the future unseen data (Kohavi and Provost 1998). It has been defined by Arthur Samuel (1959) as "Field of study that gives computers the ability to learn without being explicitly programmed" (Simon 2013). As quoted from Tom M. Mitchell's definition of machine learning which is "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1997). Due to its importance, machine learning has become the integral part of analysis pipeline in this era of ever-increasing amounts of data.

The fundamental part of machine learning is to learn from the known properties of the data and from past experiences and then give accurate predictions on new cases based on learning from the trained sets. A specific set of methods/algorithms including decision tree-based learning, support vector machines, Bayesian networks, instance-based learning, and artificial neural networks has been used for training the model system. Various parameters are needed to be tuned to optimize the performance of the learned model systems (Bishop 2006). Machine learning helps in

S. Jamal (✉) · S. Goyal
Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan, India

A. Grover
School of Biotechnology, Jawaharlal Nehru University, New Delhi, India

A. Shanker
Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan, India

Department of Bioinformatics, Central University of South Bihar, Gaya, Bihar, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_16

finding solutions to a wide range of problems, and its applications include search engines, information retrieval, bioinformatics, cheminformatics, disease diagnosis, speech and handwriting recognition, image processing, and many more to mention.

16.1.1 Types of Machine Learning

Machine learning is usually classified into two types based on the availability of the data and the input given to the learning system; these include supervised learning and unsupervised learning approach (Stuart and Peter 2003). A third type of learning approach, rather less frequently used, is the reinforcement learning approach.

16.1.1.1 Supervised Learning

In the supervised learning approach which is also known as the predictive approach, the task is to predict for the unknown from a labeled set of training data (Mohri et al. 2012). The training data comprises a set of input objects say *i*, which are basically represented by vectors describing the properties of the objects and the corresponding output classes, forming input-output pairs. Consider the input space as X and output space as Y in the training data and an example which lies in the form, $\{(x_{i,}\, y_i)_{i\,=\,1\text{ to }N}\}n$; then x_i is the feature vector of the i-th object of the training data and y_i where y is the output label of the *i*-th object. These properties, which may be anything, say the age and height of a person contain information about the input objects and are known as features or attributes. It is advised that the number of features must not be too large as it would result in many dimensions which may confuse the learning algorithm. A supervised learning approach examines the training data and results in a function using which it further attempts to determine the output class for unobserved cases (Murphy 2012). Various supervised learning algorithms use a subset of training data, known as a validation set, to determine the accuracy of the learning algorithm by means of cross-validation.

16.1.1.2 Unsupervised Learning

In unsupervised learning, also known as descriptive learning, the data is unlabeled and the task is to find some hidden interesting patterns in the data (Murphy 2012). This differentiates the unsupervised learning approach from the supervised learning and the reinforcement learning. The data consists of only input space which lies in the form $\{(x_{i,})_{i\,=\,1\text{ to }N}\}$, and there are no input-output pairs. This algorithm does not use any explicit output labels; therefore, it uses various other approaches such as clustering of the input data based on the similarities in the features and further placing the unseen instances into one or the other cluster (Daumé 2012). Other approaches used by unsupervised learning include discovering the most contributing attributes employing dimensionality reduction techniques like principal component analysis (PCA) and also by determining the correlated variables using graph theory.

16.1.1.3 Reinforcement Learning

Reinforcement learning is a less commonly used area of machine learning in which learning is associated with a reward or a punishment and the behavior of the learning algorithm or agent is based on a set of environment states. The task of the agent is to determine the ideal behavior and maximize the rewards (Sutton and Barto 1998).

16.1.2 Applications of Machine Learning

Machine learning is used widely in the plethora of tasks, mostly for classification purposes; some examples include email filtering, web page ranking, disease diagnosis, face detection, and many more (Alex and Vishwanathan 2008).

Through machine learning, a system can be generated which can sort the emails by redirecting received emails containing significant information into the inbox, sent mails in the outbox, and other emails about discounted products and offers into the spam section. The learning algorithm is trained to classify the mails based on the information they carry and process them automatically as spam or not spam (Tretyakov 2004).

One of very interesting application of machine learning is information retrieval where when a user enters a query in a search engine, it displays a list of web pages sorted according to the significance of the information they contain matching the user's query term. The training data consists of the content of various web pages, the link structure, etc., and the learned system classifies the information as relevant or irrelevant (Yong et al. 2008).

Another application of machine learning is face detection in security systems where the computer system or the software identifies a person or tells if the face is unknown. The system takes into account various features like the complexions, person wearing glasses or not, hairstyle, expressions, shape and size of eyes, nose, etc. and learns to recognize a person based on these features (Brunelli and Poggio 1993).

Machine learning can also be used in the disease diagnosis. The learned model predicts if a person suffers from a particular disease or not. The system uses all the data related to the disease including the associated symptoms, the histological data, the time period, and regional information as attributes and deduces if a person is a sufferer or non-sufferer (Sajda 2006).

Translation between two documents is a tedious task as one needs to fully understand a text prior to its translation plus it involves huge chances of loss of accuracy of information and grammatical errors. Machine learning has proved to be quite successful in automatic translation of the documents making it fast and accurate.

Optical character recognition (OCR) involves electronic translation of images of the handwritten documents, say scanned documents, into machine-readable language like ASCII code. The technique is used for entering data into the system for a wide range of documents that include passports, bank statements, printed receipts, and other different documents. The role of machine learning in OCR is to classify a character into a character region; the features of the character regions are already stored in the classifier. Whenever a new character comes across, the classifier tries to match the properties of the character to the character region and assigns a region best matching the properties (Dervisevic 2006).

There are other areas where machine learning is applied efficiently like in bioinformatics, computational and systems biology, and cheminformatics which include bioactivity data analysis, gene expression data classification and gene prediction, protein structure prediction, identification of biomarkers, and many more. Various learning algorithms have been increasingly used for analyzing gene expression data from microarrays for more accurate phenotypic classification of diseases and diagnosis in addition to the prediction of novel disease-associated genes (Jamal et al. 2017; Moore et al. 2007). Protein structure prediction, which is one of the most complex problems in structural biology and bioinformatics, has also been addressed by machine learning methods (Cheng et al. 2008). Machine learning algorithms have been widely used for generating predictive classification models which could identify probable active compounds from large unscreened chemical compounds libraries (Singh et al. 2016).

16.2 Steps to Build a Machine Learning Model

16.2.1 Inputs in the Form of Instances and Features

Machine learning is basically training a model using some objects and then performing predictions on some other objects. An instance can be any example, object, case, or item to be classified by the learned model system. The instance is an object used by the learning algorithm for training a model and on which the model carries out the predictions. These objects or instances are represented by feature vectors (Christopher 2006). Features, also known as descriptors or attributes, are the set of predetermined quantifiable properties of an object; say in flower classification, an object is the flower, so the features might be color of the flower, number of sepals, number of petals, sepal length, petal length, etc., the objects are encoded as features, and then these features are used to decide the class for the object (Murphy 2012). If the instance is a molecule, the chemical information encoded within the molecule is transformed into a mathematical representation of that molecule which is known as the molecular descriptors or features. A number of commercial and free molecular descriptor generation software are available which include ADAPT (Valla et al. 1993), ADMET Predictor [Simulations Plus Inc., Lancaster, CA], Dragon [Talete, Milano, Italy], JOELib (JOELib/JOELib2 cheminformatics library), Marvin Beans [ChemAxon], Molecular Operating Environment (MOE; Chemical Computing Group Inc. 2015), PaDEL (Yap 2011), PowerMV (Liu et al. 2005), and many more.

Choosing a subset of features which contain relevant information toward the classification to overcome the dimensionality curse and simplify the classification process is a primary step in machine learning.

16.2.2 Feature Selection

Feature selection is a method that involves discovering an optimal subset of features from the original set of features. The accuracy and robustness of some machine learning algorithms depend on the number of features chosen to represent the objects/instances (Mitchell 2014). Feature selection techniques, one of the most significant steps in machine learning, are used to simplify and fasten the learned system generation process and increase the accuracy of the classification by reducing the dimensionality and noise from the data. The irrelevant features, those which do not give any information in making predictions by the classifier, add noise and increase the complexity of the data. In feature selection, descriptors which contribute most toward the prediction task have been searched. A subset of relevant features, though may be a few in number, prove to be extremely important for the prediction task (Daumé 2012). The remaining irrelevant features are not considered during the training process.

Another issue to be taken care of is the redundancy in the descriptors. If two features have very similar values for the objects, then they are highly correlated and thus can be discarded without much information loss (Ethem 2009).

The basic principle behind feature selection techniques is testing each subset of features and finding the subset which decreases the error the most. Two methods are employed in the subset selection process, backward selection and forward selection. The backward selection method starts with the complete set of features and removes the features by deleting one feature at a time. The process continues until the removal of a feature increases the error. In forward selection algorithm, the process starts with an empty set of features, and then the features are added one by one till the error is decreased (Guyon and Elisseeff 2003).

16.2.3 Methods to Search Features

16.2.3.1 Best First

The best-first search approach employs greedy hill climbing algorithm and derives a subset of features. Once this subset is obtained, its features are examined for the information gain. A new feature is defined on the basis of the information available from the features of this subset, and then previously chosen features are removed. The procedure is repeated until all the features have been taken into account (Dang and Croft 2010).

16.2.3.2 Exhaustive Search

Exhaustive search is a simple approach which starts from a random point, selects an empty set of features, and then performs a comprehensive search over all probable subsets of features (Karuppasamy et al. 2008).

16.2.3.3 Genetic Search

The genetic search approach uses the genetic algorithm and finds an optimal feature subset. The data is in the binary form, i.e., a feature is either present or absent from the subset. Further the fitness function values, the larger the better, for these features are calculated. This process is continual till better solutions are obtained (Tiwari and Singh 2010).

16.2.3.4 Greedy Stepwise

The algorithm performs a greedy forward and backward search throughout the feature space. The algorithm either starts with no attributes or does a random selection of attributes considering the most descriptive attributes and discards the remaining ones. The process stops when addition or deletion of attributes effect the accuracy of prediction (Farahat et al. 2011).

16.2.3.5 Scatter Search

Unlike other feature selection algorithms, the scatter search is a directed search which includes a predefined reference subset of diverse attributes. This subset acts as a reference point and an attempt is made to increase its diversity. Further, the search is applied and the reference set is updated, and the procedure terminates when a predecided threshold is achieved or the search no longer produces improved results (López et al. 2006).

16.3 Machine Learning Algorithms

16.3.1 Naïve Bayes

The Naïve Bayes (NB) algorithm is a simple classifier that employs Bayes formula and estimates the probability of an object belonging to a particular class. The classifier assumes that the occurrence of one feature does not relate to the presence or absence of any other feature and considers all attributes as statistically independent of each other. For example, an animal is an elephant if it has large ears and has trunk and tusks; all these features are dependent on each other, but the Naïve Bayes classifier considers all these features as independently contributing toward the probability of the animal of being an elephant. The algorithm computes the posterior probability of each class, and the object is placed in the class which is the most probable (Friedman et al. 1997). The Bayesian classifier provides a flexible approach to machine learning where the probability for each hypothesis can be increased or decreased and the test instances are assigned the class based on the observed data, i.e., it calculates the prior probability and then the posterior probabilities. The Bayesian learning-based NB classifier finds its application in a wide range of classification problems (Mitchell 2014).

16.3.2 Random Forest

Random Forest (RF) classifier is an ensemble classifier developed by Leo Breiman. The algorithm uses decision trees which are generated by randomly selecting the features from the training data. The nodes of the tree are the features, the branches are the values, and the edges correspond to the classes. Each node in the tree links to an attribute, and each branch from this node represents a value of that attribute. The classifiers consist of a forest of trees which are then used to categorize a new instance. Initially the tree, at each node, uses the subset of features chosen randomly, and the best subset is used to split the node. The attribute which has the maximum information gain provides the best prediction and thus is selected as the decision-making attribute (Ali et al. 2012). The classifier does not involve pruning of the trees, and each tree is grown as long as possible, and the process is terminated when each attribute has been incorporated at least once or if all the training instances associated with that attribute have the same value. When a test instance is encountered, each tree is examined for the features at the nodes, and the instance is assigned the class which is the output of the larger number of trees (Mitchell 2014).

16.3.3 Support Vector Machines

Support vector machines (SVM) are non-probabilistic classifiers that use a kernel function and attempts to find a hyperplane in a high-dimensional space. The algorithm tries to find a linearly separating hyperplane amid the two classes, and then the margins of the hyperplane are maximized. The support vectors lie on either side of the margins of the hyperplane. In case of high-dimensional data, the algorithm makes use of kernel functions which convert the original input space into nonlinear input space. For SVM to perform multiclass classification task, the algorithm will reduce it to several binary classification problems. The various kernel functions include linear, radial basis function (RBF), polynomial, and the sigmoid kernel. The efficiency of the SVM classifier depends on the choice of the kernel function and kernel parameters and one more parameter, which is the trade-off between training error and the margin (Platt 1998). The use of the type of the kernel function depends on the type of the classification problem; however, RBF is the kernel of choice in most cases. The SVM training generates learned model system which classifies the test instances into any of the two categories which are on either side of the separating hyperplane (Hsu et al. 2003).

16.3.4 Artificial Neural Network

Artificial neural network (ANN) is a widely used algorithm inspired by the central nervous system and works on the same principle as the human brain works. ANNs are generally complex interconnected neurons which transfer messages to and from each other. The algorithm consists of three layers, an input layer where the input is given, a hidden layer where the processing takes place, and an output layer which records the output. A number of features are fed into the input unit which is then forwarded to the hidden unit, and the hidden unit further feeds these features to the single output layer. The edges that connect these layers are the weighted neurons, and during the training phase, the algorithm tries to fluctuate these weights for the system to learn to connect between the input and output layers (Mitchell 1997). Initially, the weights are varied in the hidden layer based on the features in the input layer following which the output units are computed based on the hidden layer features and weights.

16.3.5 k-Nearest Neighbors

The k-nearest neighbor algorithm (kNN), also known as the lazy learning algorithm, is one of the simplest nonparametric machine learning algorithms which is based on instance-based learning. The algorithm takes as input the training instances and assigns a test instance the class voted by the majority of its closest neighbors, i.e., where k is a positive integer. Mostly, the value of k is kept small if $k = 1$ the algorithm will assign the instance same class as of its nearest neighbor (Mitchell 2014). The training instances lie as position vectors in the feature space and the distance between the training instances and the query is calculated. Euclidean distance matrix is generally used to calculate the distances. To increase the accuracy of the classification, weights can be added to the closest neighbors so that they contribute more toward the classification. The effectiveness of the classifier depends on the value of k; it is preferred to choose an odd value for k in the binary classification problems (Altman 1992).

16.4 Model Validation

16.4.1 Testing Set

A learned model system generated is only effective if it can make the prediction for the previously uncharacterized data which is known as the testing set. The model systems are generated using the training data in which the class to which a particular instance belongs is already known. However, the model systems generated are validated to assess the performance of the classification algorithms using the testing data in which the outcome is not already known to the learned system. The test set is a set of instances that did not have any role during the learning of the model system.

16.4.2 Cross-Validation

To gain insights into the performance of the learned system on previously unknown data and to use the best parameter values for generating the model, cross-validation technique is used. An internal assessment of the learned system is performed by breaking the training data into subsets which are known as validation sets. The validation sets are used for tuning of the parameters during the formation of the classifiers.

16.4.2.1 N-Fold Cross-Validation

In n-fold cross-validation, the training data is divided into N equally sized folds, and each time during the learning process, N-1 folds are employed for training, and the onefold left is used as test set. This procedure is repeated N times until every fold has been used as the test set at least once following which the average performance over all the folds is taken to produce a single output. Usually, five- or tenfold cross-validation is used depending upon the dimensions of the training set.

16.4.2.2 Leave-One-Out (LOO) Cross-Validation

In LOO cross-validation, if the training data is comprised of K instances, K-1 instances are used to generate the model, and the remaining one instance is used as the testing set. K-learned systems are obtained at the end among which either any of the K model is used as final learned system or a newly learned system can be generated on the whole data using the best parameter values selected by cross-validation. The LOO form of cross-validation makes comprehensive use of the data and thus is computationally very expensive.

16.4.3 Evaluating Classifier Performance

A variety of statistical figures have been suggested to test the predictive ability of the learned model system. A learned system is not considered as an accurate system if it produces an error on training data predictions. In case of binary classification problems, the instances are divided into true positives, TP (positive prediction); true negatives, TN (negative prediction); false positives, FP (negative predicted as positive); and false negatives, FN (positive predicted as negative). There are various metrics used which include true positive rate or sensitivity or recall, (TP/(TP + FN)), which is the proportion of the positive predictions. True negative rate or specificity, (TN/(TN + FP)), is the percentage of negative predictions identified as negative. Another very popularly used metric is precision (TP/(TP + FP)) also referred to as positive prediction value, which is the fraction of positive predictions which are actual positives. Accuracy (TP + TN/(TP + TN + FP + FN)) is the percentage of the correct positive and negative predictions. A good model system is one which gives highly accurate prediction on training data; however, accuracy alone cannot be used as a measure for classification tasks as it will always predict the majority class for all the instances. A balanced measure is required to overcome the accuracy paradox. F-measure or F-balanced score, (2 × PrecisionxRecall/Precision + Recall), is the harmonic mean between precision and recall which is used to evaluate the accuracy of the model systems.

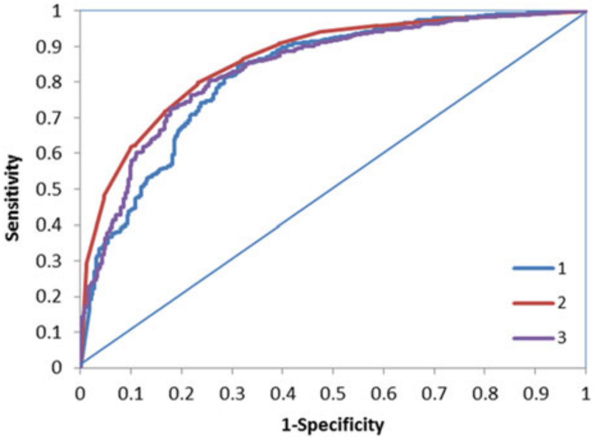


Fig. 16.1 A ROC plot generated using Weka

The performance of the learned systems can also be visualized by computing curve between sensitivity and 1-specificity; the plot is known as receiver operating characteristic (ROC; Fig. 16.1). The area under curve (AUC) value can be computed from the ROC plot which gives information about the performance of the learned systems. The value of AUC lies between 0 and 1 which is the best possible value. A range of other useful evaluation metrics have also been proposed which can be used depending on the requirement of the classification task (Demsar 2006).

16.5 Machine Learning Software

The following software suites are the implementations of various machine learning algorithms:

16.5.1 Open Source Software

- dlib, a C++ based machine learning library
- OpenNN, a C++ implementation of neural networks
- Torch, LuaJIT-based computing framework for machine learning algorithms
- ELKI (for Environment for Developing KDD-Applications Supported by Index-Structures), a Java-based platform for knowledge discovery in databases
- Orange, C++ and Python-based machine learning suite
- Scikit-learn, largely Python-based machine learning library
- R, a programming language that implements a range of techniques, one among which is m machine learning
- Weka (Waikato Environment for Knowledge Analysis), a very popular Java-based suite of machine learning techniques

There is a wide range of other open source software suites including Apache's Spark, Intel's OpenCV (Open Source Computer Vision), Encog, and Shogun.

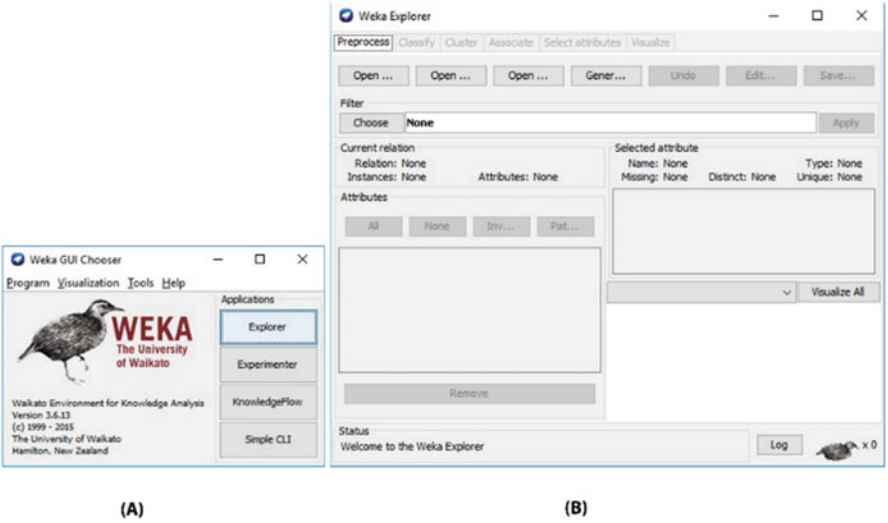
16.5.2 Commercial Software

- Amazon machine learning, machine learning platform offered by Amazon.
- KXEN modeler.
- Neural designer developed by Intelnics.
- Mathematica, written in Wolfram language.
- STATISTICA Data Miner developed by StatSoft.
- MATLAB (matrix laboratory) is a programming language developed by MathWorks that allows implementation of machine learning algorithms.

Other commercial software for machine learning includes Microsoft Azure, RCASE, SAS Enterprise Miner, IBM SPSS Modeler, and NeuroSolutions.

16.6 A Case Study Using Weka Machine Learning Platform

Weka is one of the most popularly used free accessible machine learning suite developed by University of Waikato, New Zealand (Fig. 16.2). The suite consists of tools for preprocessing of data, classification, clustering, regression, and visualization.



Waikato Environment for Knowledge Analysis
Version 3.6.13
(c) 1999 - 2015
The University of Waikato
Hamilton, New Zealand

None

Relation: None
Instances: None
Attributes: None

Name: None
Missing: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer

Fig. 16.2 Weka: (**a**) GUI chooser indicating the explorer interface, (**b**) explorer interface

16.6.1 Predicting Activity Outcome for Chemical Compounds

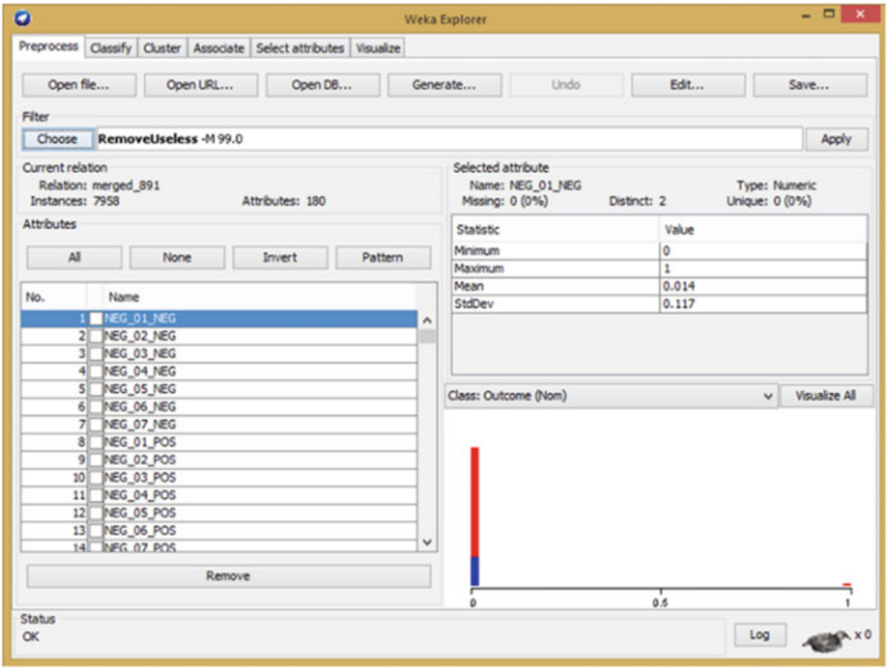
The goal of the present study was to generate machine learning-based predictive model which can find bioactive compounds from a high-throughput bioassay dataset consisting of the active and inactive compounds.

16.6.1.1 Dataset Description

The high-throughput bioassay dataset was downloaded from the PubChem database maintained by National Center of Biotechnology Information (NCBI). The bioassay was conducted to identify inhibitors and substrates of cytochrome P450 2D6. The dataset consisted of 1623 active and 6338 inactive compounds.

16.6.1.2 Data Preparation

The attributes or features for the compounds were generated using the descriptor generation software, PowerMV. A total of 179 attributes were generated, and the problematic attributes were removed. The dimensionality of the dataset was reduced by removing the attributes having identical values throughout the dataset, using the RemoveUseless filter of Weka. Figure 16.3 shows reading in the compounds data and choosing RemoveUseless filter.



Filter: RemoveUseless -M 99.0

Current relation
Relation: merged_891
Instances: 7958 Attributes: 180

Selected attribute
Name: NEG_01_NEG Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.014
StdDev	0.117

Attributes

No.		Name
1	☐	NEG_01_NEG
2	☐	NEG_02_NEG
3	☐	NEG_03_NEG
4	☐	NEG_04_NEG
5	☐	NEG_05_NEG
6	☐	NEG_06_NEG
7	☐	NEG_07_NEG
8	☐	NEG_01_POS
9	☐	NEG_02_POS
10	☐	NEG_03_POS
11	☐	NEG_04_POS
12	☐	NEG_05_POS
13	☐	NEG_06_POS
14	☐	NEG_07_POS

Class: Outcome (Nom)

Status
OK

Fig. 16.3 Reading in the compounds data and choosing RemoveUseless filter

16.6.1.3 Training and Testing Set

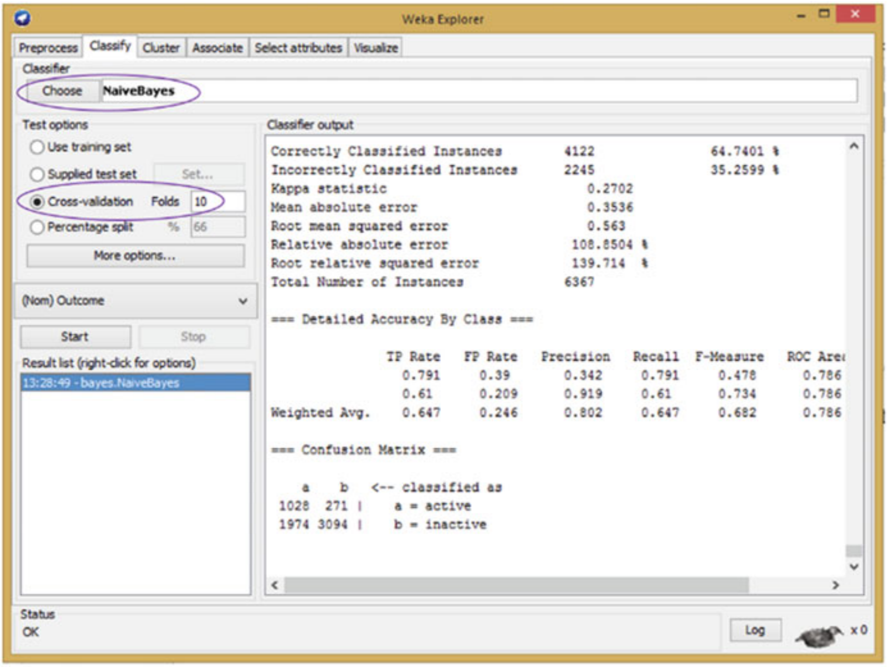
The resultant significant attributes were saved in CSV (comma space value) format, and the data was divided into 80% training set which was to train the model and 20% test set which was used to assess the performance of the generated model.

16.6.1.4 Model Generation Using Different Learning Techniques

The train and test files were converted to ARFF (Attribute-Relation File Format) using Weka, and the different machine learning algorithms were used to generate the predictive models using the training set. The machine learning algorithms can be used by going to "Classify" tab in Weka, and the different algorithms can be chosen under the Classifier category (Fig. 16.4). The number of folds for cross-validation can also be specified in the box placed under "Cross-validation."

Once the model is generated, its performance can be evaluated using the testing set which can be supplied using "Supplied test set" option available in "Classify" tab of Weka (Fig. 16.5).

The performance of the model can be improved by changing the machine learning algorithm used and altering the parameters of the algorithm.



Choose NaiveBayes

● Cross-validation Folds 10

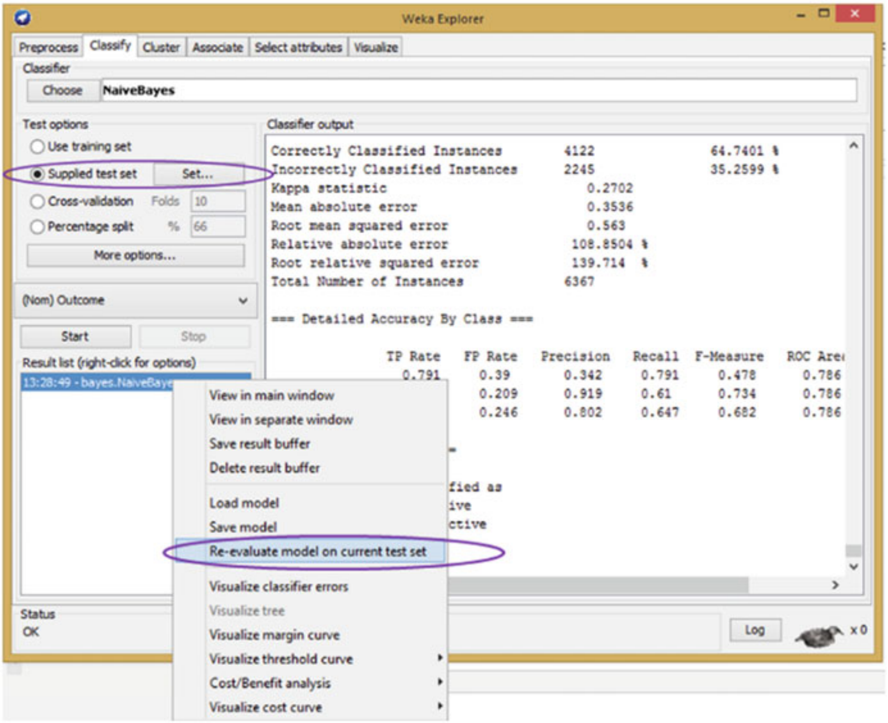
(Nom) Outcome

13:28:49 - bayes.NaiveBayes

```
Correctly Classified Instances        4122               64.7401 %
Incorrectly Classified Instances      2245               35.2599 %
Kappa statistic                          0.2702
Mean absolute error                      0.3536
Root mean squared error                  0.563
Relative absolute error                108.8504 %
Root relative squared error            139.714  %
Total Number of Instances             6367

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
                 0.791     0.39      0.342     0.791     0.478      0.786
                 0.61      0.209     0.919     0.61      0.734      0.786
Weighted Avg.    0.647     0.246     0.802     0.647     0.682      0.786

=== Confusion Matrix ===

    a    b   <-- classified as
 1028  271 |    a = active
 1974 3094 |    b = inactive
```

Status OK

Fig. 16.4 Weka classify tab, model generation using Naïve Bayes classifier



Classifier: NaiveBayes

○ Use training set
● Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Outcome

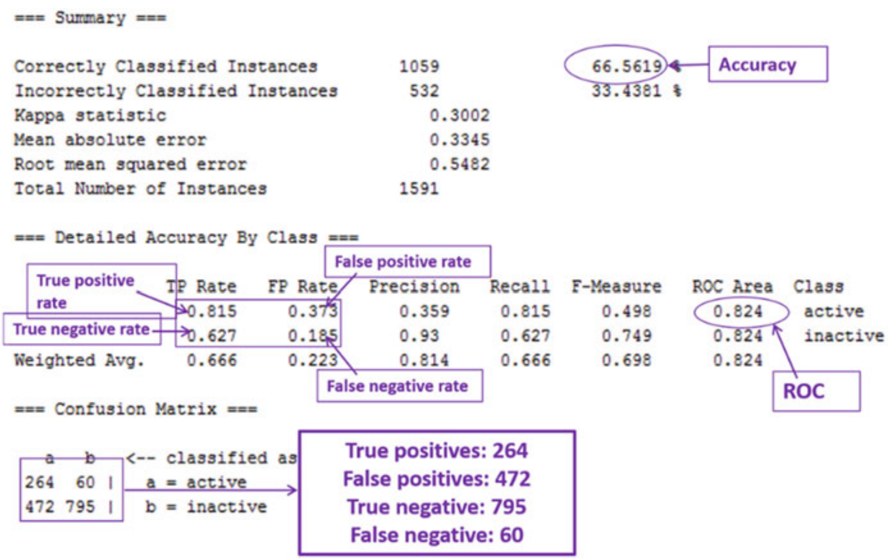
```
Correctly Classified Instances        4122               64.7401 %
Incorrectly Classified Instances      2245               35.2599 %
Kappa statistic                          0.2702
Mean absolute error                      0.3536
Root mean squared error                  0.563
Relative absolute error                108.8504 %
Root relative squared error            139.714  %
Total Number of Instances             6367

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
                   0.791     0.39      0.342     0.791     0.478     0.786
                             0.209     0.919     0.61      0.734     0.786
                             0.246     0.802     0.647     0.682     0.786

                         -

                         fied as
                         ive
                         ctive
```

Fig. 16.5 Supply test set to Weka for evaluation of the generated learning model

16.6.1.5 Statistical Assessment of the Generated Models

The suitable metrics can be computed for the data, and the values can be recorded for the components of the confusion matrix which takes account of TP, FP, TN, and FN (Fig. 16.6). The various statistical figures of merit which can be employed have already been discussed in Sect. 16.4.3.

The increasing amount of data generated in recent years and the growing curiosity in using this data to discover new facts and make better and improved decisions for the future has led to the development of various robust and effective machine learning algorithms discussed in this chapter. The types of learning method to be used depend on the nature of the data and can be employed to various applications of machine learning to generate learned model systems for prediction.



```
=== Summary ===

Correctly Classified Instances        1059               66.5619
Incorrectly Classified Instances       532               33.4381 %
Kappa statistic                          0.3002
Mean absolute error                      0.3345
Root mean squared error                  0.5482
Total Number of Instances             1591

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.815     0.373     0.359      0.815     0.498      0.824    active
                 0.627     0.185     0.93       0.627     0.749      0.824    inactive
Weighted Avg.    0.666     0.223     0.814      0.666     0.698      0.824

=== Confusion Matrix ===

   a   b   <-- classified as
 264  60 |   a = active
 472 795 |   b = inactive
```

Fig. 16.6 Output obtained from the generated Naïve Bayes classifier

Acknowledgments Salma Jamal acknowledges a Senior Research Fellowship from the Indian Council of Medical Research (ICMR).

References

Alex S, Vishwanathan SVN (2008) Introduction to machine learning. Cambridge University Press, Cambridge

Ali J, Khan R, Ahmad N, Maqsood I (2012) Random forests and decision trees. Int J Comput Sci Issues 9(5). JOELib/JOELib2 cheminformatics library

Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46(3):175–185

Bishop CM (2006) Pattern recognition and machine learning. In: Information science and statistics. Springer, New York

Brunelli R, Poggio T (1993) Face recognition: features versus templates. IEEE Trans Pattern Anal Mach Intell 15(10):1042–1052

Chemical Computing Group Inc (2015) Molecular operating environment (MOE). 2013.08 edn., Sherbooke St. West, Suite #910, Montreal, QC, Canada

Cheng J, Tegge AN, Baldi P (2008) Machine learning methods for protein structure prediction. IEEE Rev Biomed Eng 1:41–49. https://doi.org/10.1109/RBME.2008.2008239

Christopher B (2006) Pattern recognition and machine learning. In: Information science and statistics. Springer, New York

Dang V, Croft WB (2010) Feature selection for document ranking using best first search and coordinate ascent. In: Proceedings of SIGIR workshop on feature generation and selection for information retrieval

Daumé H (2012) A course in machine learning. ciml.info

Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Dervisevic I (2006) Machine learning methods for optical character recognition. pp 1–25

Ethem A (2009) Introduction to machine learning. The MIT Press, Cambridge

Farahat AK, Ghodsi A, Kamel MS (2011) An efficient greedy method for unsupervised feature selection. In: 11th IEEE international conference on data mining

Friedman N, Geiger D, GoldSzmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hsu C-W, Chang C-C, Lin C-J (2003) A practical guide to support vector classification. National Taiwan University

Jamal S, Goyal S, Shanker A, Grover A (2017) Computational screening and exploration of disease-associated genes in Alzheimer's disease. J Cell Biochem 118(6):1471–1479. https://doi.org/10.1002/jcb.25806

Karuppasamy S, Indradevi MR, Rajaram R (2008) Combined feature selection and classification – a novel approach for the categorization of web pages. J Inf Comput Sci 3(2):083–089

Kohavi R, Provost F (1998) Glossary of terms. Mach Learn 30:271–274

Liu K, Feng J, Young SS (2005) PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. J Chem Inf Model 45(2):515–522. https://doi.org/10.1021/ci049847v

López FG, Torres MG, Batista BM, JAM P, Moreno-Vega JM (2006) Solving feature subset selection problem by a parallel scatter search. Eur J Oper Res 169(2):477–489

Mitchell TM (1997) Machine learning. McGraw-Hill Science/Engineering/Math, Maidenhead

Mitchell JB (2014) Machine learning methods in chemoinformatics. Wiley Interdiscip Rev Comput Mol Sci 4(5):468–481. https://doi.org/10.1002/wcms.1183

Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of machine learning. The MIT Press, Cambridge (MA)/London

Moore CL, Smagala JA, Smith CB, Dawson ED, Cox NJ, Kuchta RD Rowlen KL (2007) Evaluation of MChip with historic subtype H1N1 influenza A viruses, including the 1918 "Spanish Flu" strain. J Clin Microbiol 45 (11):3807-3810. JCM.01089-07 [pii]https://doi.org/10.1128/JCM.01089-07

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge

Platt JC (1998) Sequential minimal optimization: a fast algorithm for training support vector machines. Microsoft Research

Sajda P (2006) Machine learning for detection and diagnosis of disease. Annu Rev Biomed Eng 8:537–565. https://doi.org/10.1146/annurev.bioeng.8.061505.095802

Simon P (2013) Too big to ignore: the business case for big data. Wiley, Hoboken

Singh H, Kumar R, Singh S, Chaudhary K, Gautam A Raghava GP (2016) Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. BMC Cancer 16:77. https://doi.org/10.1186/s12885-016-2082-y10.1186/s12885-016-2082-y [pii]

Stuart R, Peter N (2003) Artificial intelligence: a modern approach, 2nd edn. Prentice Hall, Upper Saddle River

Sutton R, Barto A (1998) Reinforcement learning: an introduction. MIT Press, Cambridge, MA

Tiwari R, Singh MP (2010) Correlation-based attribute selection using genetic algorithm. Int J Comput Appl 4(8):0975–8887

Tretyakov K (2004) Machine learning techniques in spam filtering. Institute of Computer Science, University of Tartu

Valla A, Giraud M, Dore JC (1993) Descriptive modeling of the chemical structure-biological activity relations of a group of malonic polyethylenic acids as shown by different pharmacotoxicologic tests. Pharmazie 48(4):295–301

Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32(7):1466–1474. https://doi.org/10.1002/jcc.21707

Yong SL, Hagenbuchner M, Tsoi AC (2008) Ranking web pages using machine learning approaches. Web Intelligence and Intelligent Agent Technology, 2008 WI-IAT '08 IEEE/WIC/ACM International Conference 3:677–680

Command-Line Tools in Linux for Handling Large Data Files 17

Deepti Mishra and Garima Khandelwal

17.1 Introduction

Before the 1990s Unix was well known in scientific community and used extensively by the experts. However, in the beginning of the 1990s, Linus Torvalds started working on creating a freely available and academic version of UNIX popularized as Linux which is a stable and reliable operating system (OS). These days it has become very well known and acceptable in scientific community. Bioinformaticians and computational chemists are one of the prime users of Linux system because it handles large data files generated as an output while working in the field of genomics and proteomics. Handling and mining these files to extract useful information with basic and some advanced commands of Linux proved as a blessing for people working in this area.

Linux works on some basic commands which help in the management of files and system resources. The file system is arranged in a hierarchical structure, as shown in Fig. 17.1. The top of the hierarchy is traditionally called **root** (written as forward slash /). Root user's home directory, i.e. /root, differs from primary hierarchy root (/).

To gain access to a Linux-based machine, one should first ask the system administrator to provide an account (username) and password. Once the username and password is done, one should open the terminal and use basic commands to handle file and its processing. Some of the basic Linux commands are shown in Table 17.1.

The one-line description of the UNIX system is also applicable to Linux system, i.e. "On a UNIX system, everything is a file; if something is not a file, it is a process".

D. Mishra
Institute of Plant Molecular Biology, Biology Centre of the Academy of Sciences,
České Budějovice, Czech Republic

G. Khandelwal (✉)
Cancer Research UK Manchester Institute, The University of Manchester, Manchester, UK

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_17

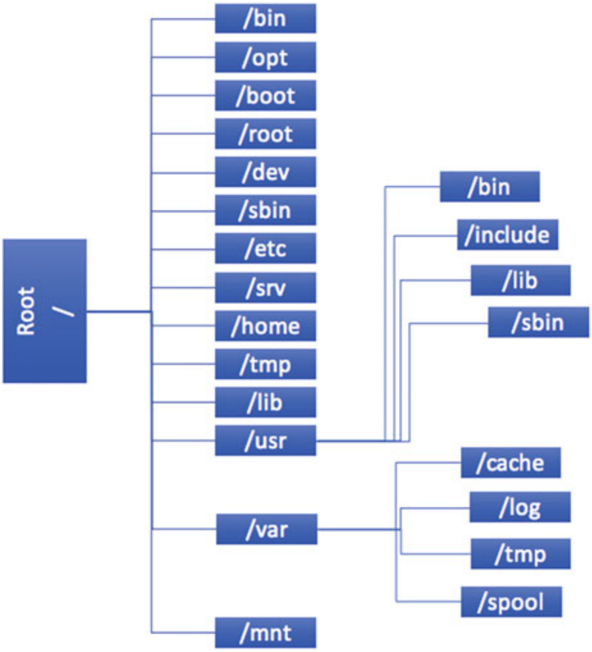


Fig. 17.1 Hierarchy of a Linux operating system

Table 17.1 Some commonly used Linux commands

Command	Description
cd	To change the current directory
ls	Displays a list of files in the current working directory
pwd	Print/display present working directory
exit or **logout**	Leave this session
man	Read manual pages on command
cat text file	Throws content of text file on the screen

For the sake of generalization and simplicity, one can say everything in a Linux system is a file; however there are some special files which are more than just a file, e.g. pipes and sockets.

The regular files which contain normal text data are executable files or program files or output of a program execution. There are some exceptions to say while keeping in mind that everything is a file in a Linux system, and these are as follows:

- Directories: files that are lists of other files.
- Special files: the mechanism used for input and output. Most special files are in / dev.
- Links:
  a system to make a file or directory visible in multiple parts of the system's file tree.

Table 17.2 File type displayed by Linux command ls -l

Symbol	Description
-	Regular file
d	Directory
l	Link
c	Special file
s	Socket
p	Names pipe
b	Block device

- (Domain) sockets: a special file type, similar to TCP/IP sockets, providing inter-process networking protected by the file system access control.
- Named pipes: act more or less like sockets and form a way for processes to communicate with each other, without using network socket semantics.

List of these file can be listed by giving the following option with ls command. The -l option to ls displays the file type (Table 17.2), using the first character of each input line:

17.1.1 Advantages of Linux

1. The very first and important advantage of Linux is that it comes with zero cost.
2. It is portable to any hardware platform.
3. Linux was made to keep on running which means that it is supposed to run without rebooting it every time.
4. Linux is scalable, secure, and versatile.

17.1.2 Disadvantages of Linux

1. Linux is considered difficult for beginners.
2. There are too many distributions available for Linux OS.

17.2 Partitioning in Linux: Why and How?

It is necessary to have a dual booting hard drive if the user wants to have both Linux and Windows working on the same machine. Partitioning of the hard drive is needed in that case. The whole drive can be allocated to a single partition, or multiple ones in case of dual-booting, maintaining a swap partition. The layout of partition is described through the partition table which are of two types, namely, Master Boot Record (MBR) and GUID Partition Table (GPT). MBR is also called as MS-DOS, whereas GPT is the more recent one. MBR's two major limitations led to the development of GPT. These limitations are as follows:

- More than four main partitions are not allowed in MBR which are considered as the primary partitions.
- 2 TB is the limit for disk partitions.

After the partition the Linux installed workstation has many more things to explore and make the OS comfortable according to user needs and workstyle. One such advantage is editing the .bashrc file which is a shell script that Bash runs whenever it is started interactively. It initializes an interactive shell session. One can write the commands here for the particular environment and customize it according to the preferences needed. A common thing to put in .bashrc are aliases that one wants to be always available. The .bashrc runs on every interactive shell launch. The user can locate the .bashrc file in the /home directory.

17.3 Advanced Linux Commands

The next level of advanced commands is for experienced Linux user, and these should be used carefully with the specified option. A list of commonly used advanced commands is given in Table 17.3.

Table 17.3 Some commonly used advanced Linux commands

Command	Description
find	Search for files in mentioned directory, hierarchically from the parent directory moving to sub-directories
grep	Search for the specific string/words for a line in a given file
ps	Give the status of the running processes with their unique id known as PID
kill	Use to kill the process which is not relevant or not responding using pid of that particular process
pkill	Kill the process using pattern specifying job/process. PID is not needed in this case
alias	Use to make complicated commands to use in a simple way by assigning an alias comfortable for the user
df	Track on disk usage by the file system
passwd	To change or modify the password in terminal
wget	Directly download the file from web. It supports HTTP, HTTPS, FTP protocols, and HTTP proxies
sftp	File transfer protocol to transfer file on a remote host
scp	Copy file from or to the remote host
ssh	Use to connect to the remote host

17.4 File Permissions

To prevent the accidental deletion or manipulation of a file by others Linux provides a very safe option to allow its user to prevent their data. There are three types of file permissions available in Linux:

Read permission (r): content of the file can only be read.
Write permission (w): content of the file can be read and edited.
Execute permission (x): indicates that file can be executed as a program.
To change the permission of a given file a command called **chmod** is used.

The above explanation gives the flavour of Linux, its history, hierarchical system, and some basic and advanced commands. To make the user comfortable for handling large data files, a very good application called Vi/Vim editor is used.

17.5 Vi/Vim

Vi is a **vi**sual text editor, developed by William Joy in the late 1970s. Vim is an improved version ("Vi IMproved") of vi, which incorporates all the utilities of vi along with many additional features. Bram Moolenaar made most of vim with the help of many others. Vim is described as compatible with vi. It is included as "vi" with most UNIX systems and with Apple OS X. Vim is a charityware whose licence is GPL-compatible and is distributed freely. Vim is based on commands given in a text user interface and it also has a GUI mode called gVim. Vim has six modes:

1. Normal mode
2. Command mode
3. Insert mode
4. Visual mode
5. Select mode
6. Ex mode

Vim always starts in normal mode, and insertion mode begins with entering an insertion or change command. [ESC] or Ctrl + C returns the editor to normal mode. In the command mode, a single line of text can be entered at the bottom of the window.

Recent command history can be looked up by typing :history command in the normal mode. Help on vim can be obtained by typing :help in the normal mode. This shows the help menu, which can then be used to select help on a particular topic. A menu number (n) can be provided with help (:help n) to open the help on the given topic.

17.5.1 Inserting Text

i	Insert before cursor
a	Append after cursor
I	Insert at beginning of line
A	Append at the end of line
o	Open a new line after current line
O	Open a new line before the current line
r	Replace one character under cursor
R	Replace many characters
gi	Return to insert mode with the cursor at its previous location

17.5.2 Saving and Quitting

:w	Save the edited file
:x	Exit, saving changes
:q	Exit as long as there have been no changes
:q!	Exit without saving any changes
ZZ	If any changes have been made, exit and save them
:wq!	Exit and save changes
:saveas file or wq! file	Save the file with new filename “file”

17.5.3 Copying and Pasting Text

yy or :y	Copy the current line
yw	Copy the current word
y$	Copy to the end of the line
p	Put after the current line
P	Put before the current line

17.5.4 Deleting Text

x	Delete a character to the right of cursor
X	Delete a character to the left of cursor
dw	Delete the word from the right of the cursor
db	Delete the word from the left of the cursor
:d or dd	Delete current line

(continued)

ndd	Delete “n” number of lines from the current line (e.g. 5dd – deletes five lines starting from the current line)
d^	Delete to the beginning of the line
D or d$	Delete to the end of the line

17.5.5 Undo and Redo

.	Repeat last text-changing command
u	Undo last change
U	Undo all changes to line
Ctrl + r	Redo last change

17.5.6 Changing Text

The change command is basically a deletion command, but it leaves the editor in insert mode.

C	Change to the end of the line
cw	Change the word
cc	Change the whole line

17.5.7 Read and Write Files

:r file.txt	Insert the contents of “file.txt” in current file
:10r file.txt	Insert the contents of “file.txt” in current file after tenth line
:1,40 w test	Write the lines 1 to 40 in file named “test”
:1,40 w >> test	Append lines 1 to 40 in file named “test”

17.5.8 Find and Replace Strings

The search and replace function is accomplished with the :s command. It is commonly used in combination with ranges.

/Chr	Search “Chr” in the forward direction
?Chr	Search “Chr” in the backward direction
n	Search for next instance of matching pattern

(continued)

N	Search for previous instance of matching pattern
/Chr/;/AA	Searches for “AA” that is preceded by “Chr”
:s/X/Y/flags	Replace X with Y according to flags
g	Flag – Replace all occurrences of matching pattern
c	Flag – Ask for confirmation of the replacement (single replacement)
0-9	Flag – Number of instances to be replaced
:2,10s/X/Y/g	Replace all occurrences of X with Y between line 2 and 10
:%s/AA//n	Counts the number of occurrences of “AA” in the file
:%s/^$/d	Delete all blank lines
:g/Chr/d	Delete all lines containing “Chr”
:v/Chr/d	Delete all lines except the ones containing “Chr”
&	Repeat last :s command

17.5.9 Cursor Movement

All these commands work only in the normal mode.

h	Move left
j	Move down
k	Move up
l	Move right
w	Move to next word
W	Move only to the next blank delimited word
b	Move to the beginning of the word
B	Move only to the beginning of a blank delimited word
e	Move to the end of the word
E	Move to the end of blank delimited word
(	Move a sentence back
)	Move a sentence forward
{	Move a paragraph back
}	Move a paragraph forward
0	Move to the beginning of the line
$	Move to the end of the line
1G	Move to the first line of the file
gg	Move to the top of file
G	Move to the bottom of file
nG or :n	Move to nth line of the file
fx	Move forward to a single character “x”
Fx	Move back to a single character “x”
H	Move to top of screen
M	Move to middle of screen

(continued)

L	Move to bottom of screen
Ctrl + f	Move forward full screen
Ctrl + b	Move backward full screen
Ctrl + u	Move up half screen
Ctrl + d	Move down half screen

17.5.10 Regular Expressions

All the expressions to be matched are preceded by a forward slash (/).

. (dot)	Any single character including except newline
*	One or more occurrences of any character
<	Beginning of a word
>	End of a word
[...]	Any character specified in the set
[^...]	Any character not specified in the set
^	Beginning of the line marker
$	End of line marker
\d	Any digit
\D or ^[0-9]	Any non-digit
\a	Alphabetic character [a-zA-Z]
\A	Non-alphabetic character [^a-zA-Z]
\l	Lowercase character
\u	Uppercase character
\c	Ignore case while matching
\s	Whitespace character
\t	Tab character
\n	End of line character
\%nl	Match in a particular line
\%>al\%<cl	Match between lines numbered “a” and “c”

If the characters mentioned in the options above need to be used literally, a backlash (\) should always precede them.

17.5.11 Regular Expression Matching

/^Chr10$/	First occurrence of the line containing only “Chr10”
/Chr/+5	Position the cursor five lines after the first match of “Chr”
/^[a-zA-Z]/	Search for lines starting with any alphabet (both upper and lower case)
/^[a-z].*/	Lines where first character is a-z which is followed by at least a single character
/TTTT$/	Search for line ending with “TTTT”
/\(GC\CG)/	Search for line containing either “GC” or “CG”

(continued)

/[0-9]*/	Matches if there are zero or more numbers in the line
/\<\d\d\d\>	Matches exactly three numbers (digits)
/^[^#]/	Matches all the lines where the first character is not a #

17.5.12 Other Useful Commands

:set nu	Display the line numbers
:set nonu	Turn off the line number display
:set hlsearch	Highlight all the matches of the given pattern
:set noh	Remove highlight from current search
:sort	Sort contents of the file
:set ignorecase	Ignore the case during search
J	Join the line below to the current one
ggguG	Change all the text to lowercase
Ctrl + n	Auto completes the word or gives suggestions to choose in case of multiple words matching the pattern
Ctrl + xl	Auto complete a line
g;	To go to previous edited positions
Ctrl + i	Move forward in jump history
Ctrl + o	Move backward in jump history
:>>	Shift current line to right by two indents
:<<	Shift current line to left by three indents
:set shiftwidth = 5	Sets indent size as five spaces
:set autoindent	Turns on auto indentation
mx	Bookmark current location (x = any key for assigning the bookmark)
`x	Jump to the bookmark with the assigned key (x)
:marks	Show all the bookmarks with line and column information
:delm x	Delete the bookmark assigned to key x
:delm!	Delete all bookmarks
:split	Split screen horizontally
:vsplit	Split screen vertically
:new file.txt or :split file.txt	Open a new file with name file.txt in a horizontal split screen
:vnew file.txt	Open a new file with name file.txt in a vertical split screen
Ctrl + ww	Move between screens
:set scrollbind	Set it up in both files on a split screen for to scroll both of them together
:tabnew	Opens a new tab
:tabedit file.txt	Open a new tab with the file "file.txt"
gt	Move between tabs
:tabfirst	Move to first tab
:tablast	Move to last tab
:tabdo %s/X/Y/g	Executes the substitute command in all the tabs
:wqa	Write and quit all open tabs

(continued)

:sh	Returns to Linux prompt temporarily. Typing exit on the command prompt will bring it back to the vim
:!pwd	Execute pwd command in Linux prompt and returns to vim by pressing return
:browse oldfiles	To get the list of old files edited using vim

All the commands in vim are case sensitive, including pattern matching (unless set ignorecase is executed).

17.5.13 Counts

Some commands can be performed multiple times by preceding the command with a number that specifies how many times a command is to be performed. For example, 5dw will delete five words, and 6fm will move the cursor forward to the sixth occurrence of the letter “m”.

17.5.14 Visual Mode

As seen above, most of the commands are performed in the normal mode, while some can be executed in the insert mode. Visual mode lets the user select a block of text and execute commands on them. User can enter visual mode by typing “v” in normal mode and can quit using ESC. Arrow keys can be used to move in visual mode. Some useful commands for visual mode are provided below (these are executed only on the selected text).

v	Enter and select in visual mode
V	Select complete row
Ctrl + v	Select blocks/columns
o	Move to the other end of the selected area
aw	Select a word
>	Indent text right
<	Indent text left
Y or Y	Copy text
d	Delete text
~	Toggle case
gu	Convert to lower case

The Vi/Vim editor is used to edit the document/codes from within the files, but there are commands like tools that can manipulate files from the shell prompt. One of these tools is “AWK” which is explained in detail in the next section.

17.6 AWK

Awk is an interpreted programming language that allows easy manipulation of structured data. 'AWK' is derived from the names of its creators – "**A**ho, **W**einberger, and **K**ernighan". Awk is mostly used for:

1. Pattern matching and processing text files
2. Generating formatted files
3. Performing string and arithmetic operations
4. Filtering text

Awk reads from standard input or a file and writes to the standard output, which can also be redirected to a file by using ">" symbol. The basic schema of awk is to search lines or files for the specified pattern and perform the desired action as shown below:

```
awk 'pattern {action}' infile > outfile
```

Awk reads one line at a time from input, performs an action based on a pattern if provided, and repeats it till the end of input. By default awk splits input lines into fields, based on spaces and tabs, and each column is assigned to variables as $1, $2 for column 1 and column 2 and so on. The variable $0 is assigned to the whole line. The default option of splitting the input on white spaces can be changed by providing a field separator (FS) using the –F option. The default output separator in awk is a white space, which can be changed with output field separator (OFS) option. Other built-in variables in awk are record separator (RS), output record separator (ORS), number of records (NR), number of fields (NF), name of current file (FILENAME), and number of records in current file (FNR). Usage for some of the in-built awk variables is described in the examples below.

17.6.1 Print a File

In this case no pattern is specified, so the action is performed on all the lines of file

```
awk '{print}' file OR awk '{print $0}' file
```

17.6.2 Print a Particular Column (Second Column) from a File

This command can be used to print the second column from a file separated with white spaces.

```
awk '{print $2}' file
```

17.6.3 Print Second and Third Column from a Comma-Separated File

Note the use of –F option followed by the field separator (,) to split the lines on comma instead of white space.

```
awk –F, '{print $2, $3}' file
```

17.6.4 Print the Second Last Column (If the Number of Columns Is Not Known)

The total number of columns in each line can be obtained by using number of field (NF) variable as - awk '{print NF}' file. Note the difference with the use of “$” for obtaining the value stored in the variable, rather than the number of fields.

```
awk '{print $(NF - 1)}' file
```

17.6.5 Print the Total Number of Lines in a File

Just like NF, number of records can be obtained by using NR variable. As NR contains the number of records, when it is used with the “$” sign, it returns the nth field of the nth record, for example, awk ' {print $NR}' file will print first field of the first record, second field of the second record, and so on.

```
awk '{print NR}' file
```

17.6.6 Print Lines Greater than 70 Characters

The length function can be used to obtain the length of the string (including white spaces), which can then be used to specify any pattern. The example given below calculates the length of each line ($0) and returns only those lines having more than 70 characters.

```
awk 'length($0) > 70' file
```

17.6.7 Print Lines Where Second Column Is Greater than 50

This will check the length of string in column 2 and will only print the lines which satisfy the pattern.

```
awk 'length($2) > 50' file
```

17.6.8 Print Every Non-empty Line

Every non-empty line will have at least one field, so NF can be used as a quick solution to remove any blank lines from the output.

```
awk 'NF > 0' file
```

17.6.9 Print Even-Numbered Lines from a File

This can also be altered to print all the odd-numbered lines by using 'NR % 2 != 0' (note: blank lines are counted as a record).

```
awk 'NR % 2 == 0' file
```

17.6.10 Calculate and Print Sum of the First Two Columns for Each Line

Simple arithmetic operations can easily be performed in awk on numerical values.

```
awk '{print $1+$2}' file
```

17.6.11 Calculate and Print Sum of Second Column from a File

BEGIN and AND are special patterns in awk that match the start and end of file. A BEGIN rule is executed only once, used for performing actions before reading the first input, such as initializing counters. Similarly an END rule is also executed only once, after all the input is read, such as printing the final calculations. In the example given below, we first calculate the total of second column into a variable named "total", finish the calculation on the file using END construct, and then print the final total. When END block is not used, total is printed after reading each record.

```
awk '{total += $2} END {print total}' file
```

17.6.12 Calculate and Print the Average of the First Column

Simply calculate as above using NR as the counter. A separate counter can also be used (note: blank lines are also counted as records).

```
awk '{total += $1} END {print total/NR}' file
```

17.6.13 Sort and Print File on the Basis of Length of Each Line

Other Linux commands including sort, grep, and rev can also be combined within awk to get the desired output. In the example below, the output of print command is piped to sort command that will output the file in increasing order of the length of each line.

```
awk '{ print $0 | "sort -n" }' file
```

17.6.14 Print Lines with "Chr" (Regular Expression Matching Is Case-Sensitive)

Awk can also be used for regular expression matching as grep.

```
awk '/Chr/' file
```

17.6.15 Print All Lines with "RNA" in Third Column

Regular expression matching can also be performed on a particular column by restricting the search space.

```
awk '$3 ~ /RNA/' file
```

17.6.16 Print Only Second Column from the Lines that Have Either of the Two Regular Expressions ("Chr" or "Human")

AND/OR operators can be used with awk to combine regular expression patterns.

```
awk '/Chr/ || /Human/ {print $2}' file
```

17.6.17 Print All the Lines that Are Between the Line Starting with Chr and Human

Some commands can be combined without the use of any of the logical operators. The following command will print all the instances of blocks of lines occurring between "Chr" and "Human", along with all the occurrences of the first match (Chr) as it looks for more blocks in the file.

```
awk '$1 == "Chr", $1 == "Human"' file
```

17.6.18 Count the Number of Lines with Chr in a File

This command only prints the number of lines with the matched pattern (the line is counted once even if the pattern is present multiple times in the same line).

```
awk '/Chr/ {count++} END {print count}' file
```

17.6.19 Adding a Column at the End of File Based on the Calculation of Other Columns (Difference Between Value in Fourth Column and Third Column) in the File

Quick calculations and file processing can be performed in awk as shown by a simple example:

```
awk '{print $0"\t"($4-$3)}' file
```

17.6.20 Add a New Line After Every Two Lines

Another quick text processing example where blank lines are added after every second line, where the first print command prints the record, NR % 2 == 0 checks for the pattern, and second print command adds a blank line.

```
awk '{print;} NR % 2 == 0 { print ""; }' file
```

17.6.21 Quartile Calculations (Second Quartile) on Second Column of the File

Using sort and awk, where sort command arranges the file in increasing order; piping the output to awk where every value is stored in an array (all[NR]) and then after the END of the file, print the record at the second quartile position.

```
sort –n –k2 file | awk '{all[NR] = $2} END {print all[int(NR*0.5)]}'
```

17.6.22 Count Number of Reads in a fastq File

As each read has four lines of information associated with it, dividing the total number of records by 4 provides the answer. Note the use of END at the start of the awk, which means that the command is to be executed only after the file reading is completed.

```
awk ‘END {print NR/4}’ file.fastq
```

17.6.23 Convert fastq to fasta

Simply get the read sequence from the fastq file, and add the “>” sign along with the read information for the fasta format.

```
awk 'NR % 4 == 1 {print ">" $0 } NR % 4 == 2 {print $0}' file.fastq > file.fasta
```

17.6.24 Get Reads Matching a Sequence Pattern (ATGCGCCC) and Print Them

Use of NR, AND logical operator (&&), and pattern matching in a fastq file for the required result.

```
awk 'NR%4 == 2 && $1~/ATGCGCCC/ {print $0}' file.fastq
```

17.6.25 Separate Reads Based on Their Length (Print Reads > = 35 Base Pairs with All of Their Information) from a fastq File

When we need to store information for the output, multiple variables could be used in the awk command to store temporary information.

```
awk 'NR%4==1{a=$0} NR%4==2{b=$0} NR%4==3{c=$0} NR%4==0 &&
   length(b)>=35 {print a"\n"b"\n"c"\n"$0;}' file.fastq
```

17.6.26 Extract the Amino Acids from a pdb (1ata.pdb) File

This gives us the amino acid sequence making the pdb file in a three-letter code (if the amino acid sequence needs to be converted into a single-letter code, just pipe in the output to a sed command containing the substitutions for all the three-letter codes to a single-letter code).

```
awk '/ATOM/ && $3 == "CA" && $5 == "A" {print $4}' 1ata.pdb
```

17.6.27 Remove Hydrogen from a pdb (1ata.pdb) File

This edits a pdb file to remove all the hydrogen atoms and their coordinates ($12 has the element value for each atom).

```
awk '/ATOM/ && $3 == "CA" && $5 == "A" {print $4}' 1ata.pdb
```

Awk as a text processor can be used to edit any type to text/data file even though most of the examples shown above deal with sequence data. Also, multiple awk commands can be used together separated by pipes.

17.7 Conclusion

It might seem difficult to work with the Linux command-line environment initially, but it is well worth to learn it if one is going to deal with large data. The handling and analyses of large data sets generated in the field of genomics, proteomics, and bioinformatics can be easily done with Linux command-line and editing tools – Vi/Vim and AWK. These commands and tools help user to easily manage the data available at hand in a timely manner.

Acknowledgement We would like to thank our teacher, Dr. Asheesh Shanker, for giving us the opportunity to contribute towards this book and our friend Dr. Vasantika Singh for her helpful suggestions.

References

http://glaciated.org/vi/
http://linuxcommand.org/lc3_adv_awk.php
http://manpages.ubuntu.com/manpages/xenial/man1/nvi.1.html
http://vimdoc.sourceforge.net/htmldoc/help.html
http://web.mit.edu/gnu/doc/html/gawk_5.html
http://www.grymoire.com/Unix/Awk.html
https://cs.stanford.edu/~miles/vi.html
https://likegeeks.com/awk-command/
https://linuxconfig.org/learning-linux-commands-awk?lang=en_gb
https://www.ccsf.edu/Pub/Fac/vi.html
https://www.computerhope.com/unix/uawk.htm
https://www.cs.colostate.edu/helpdocs/vi.html
https://www.geeksforgeeks.org/awk-command-unixlinux-examples/
https://www.gnu.org/software/gawk/manual/gawk.html
https://www.washington.edu/computing/unix/vi.html
Robbins A (2015) Effective awk programming: universal text processing and pattern matching. O’Reilly Media, Sebastopol
Robbins A, Lamb L, Hannah E (2009) Learning the Vi and Vim editors, 7th edn. O’Reilly Media, Sebastopol

Glossary

Ab initio Latin terms meaning “from the beginning.”

Ab initio modeling Uses fundamental principles of physical sciences like statistical thermodynamics and quantum mechanics to predict the 3D structure of macromolecule.

Accuracy It is the fraction of the correct predictions among the total number of instances to be examined.

ADMET A set of parameters for assessing a molecule to qualify for drug candidate and stands for adsorption, distribution, metabolism, excretion, and toxicity.

Algorithm A procedure for solving a problem.

Alignment See sequence alignment.

Alignment score It represents the number of matches, substitutions, insertions, and deletions (gaps) within an alignment. Alignment scores are often reported in log odds units and higher scores denote better alignments.

Allele frequency It is the proportion of a particular allele among all the possible alleles at the same locus in the population.

Alpha helix A secondary structure in protein having spiral conformation (helix), in which every backbone N–H group donates a hydrogen bond to the backbone C=O group of the amino acid located three or four residues earlier along the protein sequence.

Amino acid residues In a polypeptide chain, *two* amino acids combine to form a peptide bond by removal of a water molecule. Each amino acid in the polypeptide chain is referred to as an amino acid residue.

Amplified fragment length polymorphism Amplified fragment length polymorphism (AFLP) is a PCR-based genetic technique used to selectively amplify DNA fragments to genotype individuals based on the differences in their alleles.

Attribute It is a quantifiable property of the objects to be classified. Also known as feature.

Beta sheet It consists of beta strands connected laterally by hydrogen bonds, forming a pleated sheet. A beta strand has three to ten amino acids in an extended conformation.

Binning Clustering sequences based on their nucleotide composition or similarity to a reference database.

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6

BLOSUM matrices BLOck SUbstitution Matrices, computed using local multiple alignments of more distantly related sequences, as compared to PAM matrices where the dataset consisted of closely related protein families.

Bootstrap analysis One of the most popular resampling procedures used to assess the reliability of branches in a phylogenetic tree. A bootstrap value denotes the confidence for each unit or taxon of the tree.

Branch length The number of sequence changes along a branch of a phylogenetic tree.

CASP Critical Assessment of protein Structure Prediction (CASP) is a community-wide, worldwide biennial experiment for protein structure prediction.

ChIP-seq data Chromatin immunoprecipitation dataset incorporates sequence information for protein binding regions (transcription factor) on DNA.

Classifier It is a mathematical system based on machine learning algorithm (e.g., decision-tree based) that categorizes the unlabeled data to distinct output classes.

Clustering It is a grouping of objects such that the most similar objects are in the same group. For example, genes can be grouped based on similar structure or function.

Computer-aided drug designing Computer-aided drug design uses computational algorithms to discover, enhance, or study drugs and related biologically active molecules.

Conformations Alternative structures of the same molecule.

Copy number variations A particular gene can have multiple copies which can lead to the change in genotype and phenotype of an individual.

Correlation spectroscopy Correlation spectroscopy (COSY) is a 2D NMR that transfers magnetization through chemical bonds between adjacent atoms.

Coverage (in sequencing) The mean number of times a nucleotide is sequenced in a genome.

Cross-validation It is a validation method used to assess the classifier's performance**.**

De novo See ab initio.

Deoxyribonucleic acid Deoxyribonucleic acid (DNA) is a biopolymer made up of repeating units of nucleotides containing a sugar moiety (deoxyribose), nitrogenous bases (purines, adenine, guanine; pyrimidines, thymine, cytosine), and a phosphate group.

Discrete optimized protein energy Discrete optimized protein energy (DOPE) is used to assess the quality of the model via a statistical potential optimized for model assessment.

Distance The number of observed changes in an optimal alignment of two sequences, usually not counting gaps.

Docking A method that predicts the preferred orientation of one molecule with respect to another when bound to each other to form a stable complex. Knowledge of the preferred orientation is used to predict binding affinity between two molecules.

Domain It is a compact, conserved 3D part of protein structure that can evolve, function, and exist independently from the rest of the protein chain.

Dot matrix A graphical method for comparing two sequences where one is written horizontally across the top and the other along the left hand side. Dots are placed within the graph to indicate matches of characters appearing in both sequences.

Drug discovery Drug discovery is a process through which potential new medicines are identified. It involves a wide range of scientific disciplines, including biology, chemistry, and pharmacology.

Dynamic programming It involves dividing a large problem into smaller subproblems and combines their solutions to find the solution of the larger problem.

Effect size In simple terms, it is a statistical measure to quantify the strength or effect of a phenomenon. In the context of GWAS, it is the contributory effect of multiple variations toward disease association.

Encyclopedia of DNA Elements ENCODE is a consortia initiative that aims to identify the functional elements in the human genome.

Ensemble methods These are meta-algorithms which use more than one machine learning techniques to achieve higher accuracy of prediction.

European Patent Office (EPO) It is a patent office for Europe and delivers services under the European Patent Convention.

Exome Refers to the portion of the genome that is the complement of all the exons.

False positives (FPs) The false positives are the proportion of all negatives that still yield positive test outcomes.

Feature See Attribute.

Feature vector In machine learning a feature vector is an n-dimensional vector comprising of numerical features of the input objects.

Fold recognition A knowledge-based method that uses existing information of folds from already known structures to build the structure of a sequence. It identifies distant relationship among proteins.

Format (file) Sequences can be available in different formats. GenBank and EMBL have their own individual flat file formats. Additionally, the other commonly used formats for nucleotide sequences are plain text format and FASTA formats.

Frameshift As a result of an insertion or deletion of nucleotides in any number which is not a multiple of three, a shift occurs in the codon reading frame and hence changes the amino acid.

Gap Gaps are a result of either insertion or deletion, jointly refer as indels, in sequences. Gaps are introduced to maximize the matches in any column to obtain the most optimal alignment (see indel).

Gap penalty A numeric score to penalize gap opening and gap extension in an alignment.

GDT score Global distance test (GDT) score is another measure of structural similarity which gives the percentage of residues predicted accurately within given cutoff.

Gene flow Introduction/loss of new alleles into/from the gene pool of a population when organisms move in/out of it. Migration leads to change in the gene pool of the new population, due to gene flow.

Genetic drift Change in allelic frequency owing to random or chance events.

Genome-wide association study GWAS is an approach for scanning genetic markers across the complete sets of DNA, or genomes, and identifies genetic variations linked to any particular disease.

Global alignment Covers the entire length of sequences involved and is used when the sequences are reasonably similar with almost same length.

Haplotypes Haplotypes constitute group of genes on a chromosome which are inherited from single parent. It also refers to inheritance of cluster of single nucleotide polymorphism in an organism.

Hardy-Weinberg equilibrium A principle of population genetics given by GH Hardy and Wilhelm Weinberg independently in 1908, which states that in a panmictic (randomly mating) infinitely large population, the allele frequencies will remain constant over generations in the absence of external perturbations like natural selection, genetic drift, gene flow, etc.

Heterozygosity Presence of different alleles at a genomic locus.

Hidden Markov model (HMM) A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states.

High-throughput virtual screening HTVS is a computational screening method which is widely applied to screen in silico collection of compound libraries to check the binding affinity of the target receptor with the library compounds.

Homolog A set of sequences that share certain level of similarity due to a common ancestor of the two or more organisms during evolution. Homologs may apply to the relationship between genes separated by the event of speciation (ortholog) or to the relationship between genes separated by the event of gene duplication (paralog).

Homology modeling A comparative modeling method which uses a framework of already known 3D structure and predicts the structure of a homologous sequence.

Human Genome Project HGP is an international collaborative project undertaken to sequence all the 3 billion nucleotides that make up the haploid human genome, with an aim to understand its structure, organization, and function. Draft of the human genome was published in 2001, and subsequently the HGP was completed in 2003. Human genome sequencing was also carried out in parallel by Celera Genomics, a private company.

In silico In silico is an expression used to mean "performed on computer or via computer simulation."

In vitro Studies that are performed with cells or biological molecules studied outside their normal biological context.

In vivo Studies in which the effects of various biological entities are tested on whole, living organisms usually animals including humans and plants as opposed to a partial or dead organism.

Indel An insertion or deletion in a sequence alignment.

Indian Patent Office (IPO) This is a subordinate office of the Government of India and administers the Indian law of Patents, Designs, and Trade Marks.

Instance It is an input object of a study which can be a chemical compound or anything which is to be classified.

International Patent Classification (IPC) IPC provides a hierarchical system of language-independent symbols for the classification of patents and utility models according to the different areas of technology to which they pertain.

Irregular secondary structure Secondary structures that lack a regular pattern of hydrogen bonding.

K-tup A parameter to specify the size of the word and control the sensitivity and speed of the search in programs like FASTA.

Lead A chemical compound that has biological activity likely to be therapeutically useful but may still have suboptimal structure that requires modification to fit better to the target.

Linkage disequilibrium LD is defined as nonrandom association between two or more loci.

Local alignment Covers parts of the sequence and are used to compare small segments of all possible lengths when sequences have domains or regions of similarity and have different overall lengths.

Log odds score The logarithm of an odds score. Usually substitution matrices are populated with log odds scores.

Loops A secondary structure in protein that has irregular structure, which connects other secondary structure elements. It is present on protein surface and contains hydrophilic residues.

Machine learning To use algorithms for observing and exploring predictive relationships by learning from data and performing predictions on previously unseen data.

Macromolecular crystallography It is a technique to determine atomic 3D structure of biological molecules such as proteins and nucleic acids (RNA and DNA).

Metabarcoding Metabarcoding is a rapid method of biodiversity assessment that combines two technologies: DNA-based identification and high-throughput DNA sequencing. It uses universal PCR primers to mass-amplify DNA barcodes from mass collections of organisms or from environmental DNA.

Metadata Definitional data that provide information about other data.

Metagenome The DNA representing all cells of organisms obtained from an environmental sample.

Metagenomics The study or community analysis of genomic DNA obtained from environmental samples.

Microsatellites Repetitive regions of one to six nucleotides.

Molecular chaperones Protein molecules which assist other proteins to fold.

Molecular clock hypothesis It suggests that molecular sequences change at the same rate in the branches of an evolutionary tree.

Molecular descriptors These are a result of a procedure that transforms the information encoded in input objects into numerical values.

Motif A supersecondary structure in a protein that describes the connectivity between secondary structural elements and has a particular pattern.

Multiple sequence alignment The alignment of more than two sequences and is used to detect similarity between several homologs.

Native state of proteins The completely folded three-dimensional functional form of proteins.

Needleman-Wunsch algorithm A dynamic programming algorithm to generate global alignment of nucleotide/protein sequences.

Neighbor joining method Involves bottom-up clustering and creates a phylogenetic tree whose branches reflect the degrees of difference among the objects.

Next-generation sequencing It refers to non-Sanger-based high-throughput DNA sequencing technologies where millions and billions of DNA strands can be sequenced in parallel.

Nuclear magnetic resonance NMR is a method to determine the structure of macromolecule at the atomic resolution in solution. It is based on the magnetization of the nuclei, which aligns it in the magnetic field. The perturbations using radio-frequency pulse are recorded.

NOESY A method to detect the correlation between two nuclei, which are not bonded but are closely placed in space.

Non-synonymous changes Codon changes which are accompanied by changes in the amino acid, leading to alteration in the sequence of the polypeptide.

Operational taxonomic unit Species distinction in microbiology represents a species or group of species that are operationally different.

ORFans Open reading frames with no homologs in other organisms.

Pairwise sequence alignment An alignment between two sequences.

PAM scoring matrices Developed using closely related protein sequences to calculate mutation rates. These matrices are widely used to infer evolutionary relationships.

Pharmacogenetics The study of the effect of genetic factors (or variations) that are associated with differential drug responses among individuals in a population.

Pharmacophore A pharmacophore is an abstract description of molecular features which are necessary for molecular recognition of a ligand by a biological macromolecule.

Phylogenetic tree A phylogenetic tree or evolutionary tree is a branching diagram or “tree” showing the inferred evolutionary relationships among various biological species or other entities—their phylogeny—based upon similarities and differences in their physical or genetic characteristics.

Physicochemical scoring metric It helps to identify native/native-like structures via an integration of physicochemical features.

Polymorphism A heritable variation which occurs at a frequency of >1% in the population and may act as the substrate for adaptation/maladaptation.

Population bottleneck An event that leads to a drastic reduction in the size of the population due to any sudden environmental changes during natural calamity or epidemic, etc. is referred to as a population bottleneck. The effects last for at least one generation. It results in overall reduced variability or heterozygosity due to loss of alleles from the gene pool.

Positive selection When the number of non-synonymous changes exceeds that of synonymous, the coding sequence is said to be under positive selection. It is usually seen in a genomic region which is not essential for the organism and is likely to be evolving.

Primary structure of protein It is the linear sequence of amino acids linked by peptide bonds.

Proband The individual through whom a family with a genetic disorder is ascertained [Source: NIH-National Cancer Institute].

Probes Probe is a single-stranded 25bp oligonucleotide sequence that is complementary to the target DNA. There are two probes for each SNP to be genotyped. They differ only at the site of the SNP, with one probe complementary to the wild-type allele and the second probe to the mutant allele.

Propeller twist In a base pair in nucleic acid structure, the rotation of one base with respect to the other in the same base pair is called propeller twist.

Protein Protein is a biopolymer made up of repeating units of amino acids.

Protein Data Bank PDB is a structural repository of large biological molecules solved experimentally.

Protein folding problem The quest to understand the mechanism by which a protein spontaneously adapts its native structure from its primary sequence, within the biologically relevant timescale.

ProTSAV Protein tertiary structure analysis and validation (ProTSAV) is a meta-server approach for evaluating the quality of a protein.

PSI-BLAST Position-Specific Iterative Basic Local Alignment Search Tool is used to generate 1D sequence profiles between the target and template. It derives a position-specific scoring matrix or profile from the multiple sequence alignment of sequences detected above a given score threshold.

Position-specific scoring matrix PSSM is a commonly used representation of motifs in biological sequences. It is derived from a set of aligned sequences that are considered as functionally related. It has 1 row for each symbol of the alphabet, 4 rows for nucleotides in nucleic acid sequences, or 20 rows for amino acids in protein sequences and 1 column for each position in the alignment.

Purifying selection When the number of non-synonymous changes is less than that of synonymous changes, the region of the genome is said to be under purifying selection. This indicates that the sequence is vital for the organism and any changes in the protein are deleterious; hence such changes are purged out (or “selected against”) from the population.

Purines A heterocyclic aromatic organic compound consisting of a pyrimidine ring fused to an imidazole ring.

Pyrimidines An aromatic heterocyclic organic compound having single ring with nitrogen.

Qualitative Model Energy Analysis QMEAN is a composite scoring function describing the major geometrical aspects of protein structures.

Quality assessment It is evaluation of the quality of predicted protein model for distinguishing correctly modeled structures from others.

Quantitative estimate of drug-likeliness QED is a quantitative metric for assessing drug likeness.

Quantitative structure-activity relationship Quantitative structure-activity relationship is an analytical application that can be used to interpret the quantitative relationship between the biological activities of a particular molecule and its structure. It derives a correlation between calculated properties of molecules and their experimentally determined biological activity.

Quantitative trait loci They are the portion of DNA which contains genetic variations that can be associated with quantitative phenotype such as blood pressure, height, and weight.

Quaternary structure of protein It is the arrangement of protein subunits in a multi-subunit complex.

Recombination Recombination is the exchange of genetic information between two DNA molecules by the process of crossover. It results in new allelic arrangements and is the raw materials for genetic diversity.

Reference genome (human) The genome initially sequenced by HGP. It is taken as a reference for the purpose of comparison in different genomic studies.

Regular secondary structure Secondary structures that have a regular hydrogen bonding pattern.

Restriction fragment length polymorphism RFLP refer to the differences among individuals of the same species in the length of the DNA fragments obtained when it is cut using specific restriction enzymes.

Ribonucleic acid RNA is a biopolymer made up of repeating units of nucleotides containing a sugar moiety (ribose), nitrogenous bases (purines, adenine, guanine; pyrimidines, uracil, cytosine), and a phosphate group.

Roll A base pair geometry shows the rotation around the slide axis.

Root mean square deviation RMSD of atomic positions is the measure of the average distance between the atoms of superimposed proteins.

Sanger sequencing Sanger sequencing is a method of DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication.

Scanning electron microscopy SEM is a form of electron microscopy in which specimen is scanned with beams of electron to get the image.

Secondary structure of protein Secondary structure is the local substructure of proteins formed by a different pattern of backbone hydrogen bond interaction.

Sensitivity Sensitivity measures the proportion of positives that are correctly identified (see True positive rate).

Sequence alignment It involves arranging two or more sequences into rows, with characters aligned in successive columns. It is used on DNA, RNA, or protein sequences to identify regions of similarity/dissimilarity, which may be a consequence of functional and/or structural constraints. A good alignment shows evolutionary relationship between sequences.

Sequence identity (percent identity) The number of identical bases or residues (amino acids) in an alignment. Gaps in the aligned columns are not scored.

Sequence similarity (percent similarity) Similar residues/amino acids at corresponding positions (column of an alignment). In nucleotide sequences, sequence identity and sequence similarity mean the same. Gaps in the aligned columns are not scored. Often confused with homology.

Side A base pair geometry shows displacement along an axis in the plane of the base pair directed from one strand to the other.

Similarity score Sum of the number of identical matches and conservative substitutions divided by the total number of aligned sequence characters in an alignment. Gaps not considered.

Simple sequence repeats See Microsatellites.

Single nucleotide polymorphism A single nucleotide polymorphism, or SNP, is a variation at a single position in a DNA sequence of an individual, for example, nucleotide AAGCCTA can be mutated to AAGCGTA.

Smith-Waterman algorithm A dynamic programming approach to generate local alignments. All negative scores are changed to zero to assist in identifying local alignments.

Specificity Specificity measures the proportion of negatives that are correctly identified (see True negative rate).

Stop gain A mutation that converts a codon to a stop codon, resulting in premature termination of peptide. This leads to the production of a truncated protein product and is also known as nonsense mutation.

Structural bioinformatics A subdiscipline of bioinformatics that deals with structural data – representation, storage, retrieval, analysis, and visualization. Broadly divided into two areas – the development of methods to manipulate biological information to support structural biology and application of these methods to solve problems and elucidate new biological knowledge.

Structural biology A discipline of biology that tries to understand the molecular structure of biological macromolecules – proteins and nucleic acids. Addresses how they acquire a particular structure responsible for a particular function and how alterations in the structures can affect the function.

Structure Analysis and Verification Server SAVES unifies six quality assessment tools for checking and validating protein structures.

Substitution Matrices Used for scoring purpose such as BLOSUM and PAM matrices.

Supervised learning It is a type of machine learning based on labeled training data.

Synonymous changes Changes which alter the codon but the amino acid still remain the same. These mostly occur at the Wobble positions in a codon.

Target It is the molecule of interest under study.

TEM Transmission electron microscopy is a form of electron microscopy that uses electron transmission through ultrathin section of specimen to image the molecule.

Template modeling score TM score is a measure of similarity between two conformations of the same protein.

Tertiary structure of protein It is the three-dimensional structure of a protein.

Testing set A dataset used to evaluate the predictive ability of the classifiers.

The US Patent and Trademark Office (USPTO) It is the federal agency for granting patents and registering trademarks.

Threading A fold recognition method to model proteins that have same fold as the protein of known structure but do not have significant sequence similarity.

Tilt A base pair geometry shows rotation around the shift axis.

Torsion angle It is the dihedral angle between two planes. In proteins, φ and ψ angle define the rotation of the polypeptide chain.

Training set A dataset used to find the predictive relationships and train the classifiers.

True negative rate (specificity) It is the percentage of negative predictions.

True positive rate (recall, sensitivity) It is the percentage of positive predictions.

True positives (TPs) The true positives are the proportion of all positives that yield positive test outcomes.

Unsupervised learning It is a type of machine learning which draws interpretations from unlabeled training data.

Virtual screening A computational technique used in drug discovery to search libraries of small molecules in order to identify those which are most likely to bind to a drug target (a protein receptor or enzyme).

World Intellectual Property Organization WIPO is the global forum for intellectual property services, policy, information, and cooperation.

X-rays A form of electromagnetic radiation of high energy and very short wavelength, which is able to pass through many materials opaque to light. Wavelength ranges from 0.01 to 10 nanometers.