Applications of Bio-molecular Databases in Bioinformatics

Archana Kumari, Swarna Kanchan, Rajeshwar P. Sinha and Minu Kesheri

Abstract Discovery of genome as well as protein sequencing aroused interest in bioinformatics and propelled the necessity to create databases of biological sequences. These data are processed in useful knowledge/information by data mining before storing into databases. This book chapter aims to present a detailed overview of different types of database called as primary, secondary and composite databases along with many specialized biological databases for RNA molecules, protein-protein interaction, genome information, metabolic pathways, phylogenetic information etc. Attempt has also been made to focus on drawbacks of present biological databases. Moreover, this book chapter provides an elaborate and illustrative discussion about various bioinformatics tools used for gene prediction, sequence analysis, phylogenetic analysis, protein structure as well as function prediction, molecular interactions prediction for several purposes including discovery of new gene as well as conserved regions in protein families, estimation of evolutionary relationships among organisms, 3D structure prediction of drug targets for exploring the mechanism as well as new drug discovery and protein-protein interactions for exploring the signaling pathways.

Keywords Bioinformatics · Data mining · Biological databases · Gene prediction · Sequence analysis · Phylogeny · Structure prediction · Molecular interaction

A. Kumari
Department of Bio-Engineering, Birla Institute of Technology Mesra, Ranchi, India

S. Kanchan
Department of Biological Sciences, Birla Institute of Technology and Science, Pilani 333031, Rajasthan, India

R.P. Sinha · M. Kesheri (✉)
Laboratory of Photobiology and Molecular Microbiology, Centre of Advanced Study in Botany, Banaras Hindu University, Varanasi 221005, India
e-mail: minubhu@gmail.com

N. Dey et al. (eds.), *Medical Imaging in Clinical Applications*,
Studies in Computational Intelligence 651, DOI 10.1007/978-3-319-33793-7_15

1 Introduction

Bioinformatics presents one of best example of interdisciplinary science. Actually, it is the mixture of various disciplines such as biology, mathematics, computer science and statistics. The term 'Bioinformatics' was given by a Dutch system-biologist Paulien Hogeweg, in the year of 1970 [1]. For the last few decades, it has become an important part of biological research to process the biological data quickly. Nowadays, bioinformatics tools are regularly used for identification of novel genes and their characterization. Bioinformatics is also used for calculating physiochemical properties, prediction of tertiary structure of proteins, evolutionary relationship and biomolecular interactions. Although these bioinformatics tools cannot generate as reliable information as those generated by experimentation. But the experimental techniques are difficult, costly and time consuming. Therefore the in silico approach facilitates in reaching an approximate informed decision for conducting the wet lab experiment. The role of bioinformatics is not only limited to generation of data but also extended to storage of large amount of biological data, retrieval, and sharing of data among researchers. The design of databases, development of tools to retrieve data from the databases and creation of user web interfaces are the major roles of bioinformatics scientists. Life sciences researchers are using these databases since 1960s [2]. In mid 1980s, bioinformatics came into existence and National Center for Biotechnology Information in 1988 was established by USA government.

There are many types of biological databases which are called primary, secondary and composite databases. Primary databases contain gene and protein sequence information as well as structure information only. Secondary databases contain derived information from primary databases and composite databases contain criteria for searching multiple resources. Along with theses databases Literature databases, Structural databases, Metabolic pathway databases, Genome databases for specific organisms, protein-protein interaction databases, phylogenetic information databases, RNA molecules databases and protein signalling databases are also discussed in detail.

Bioinformatics is also used to integrate the data mining techniques e.g. Genetic algorithms, Support vector machines, Artificial intelligence, Hidden Markov model etc. for developing software for sequence, structure and function based analysis.

Due to flooding of genome sequencing projects, vast amount of data have been accumulated at very high rate. However, pure data are not meaningful because knowledge/information in such data is hidden. Knowledge/information is much more valuable than data many times. Thus, a new technology field has emerged in mid 1990s to extract knowledge/information from raw data which is called *knowledge discovery in databases* (*KDD*) or simply *data mining* (*DM*) [3, 4]. First of all, Data which is the raw material and related to some specific problem are

collected, checked and finally selected. After careful selection of data, preprocessing of data is required i.e. erroneous data which is also called outliers are identified and removed. After preprocessing, algorithms or mathematical models are applied to extract the useful patterns, which are called data mining. These patterns are interpreted by experts and thereafter evaluated for by their novelty, correctness, comprehensibility and usefulness. Finally, information in graphics or presentable form is available to the end user in databases. The systematic diagram of KDD process in the form of flow chart is shown in Fig. 1.

A number of reviews and scientific articles related to databases have been published in the specialized area of Bioinformatics [5, 6]. However, none of these articles prove useful for a scientist who is not from bioinformatics/computational biology discipline. Therefore in the present chapter, we proceed to introduce various bioinformatics databases to a non-specialist reader to help extract useful information regarding his/her project. In this chapter every section contains a basic idea of each area supported by the literature and a tabulated summary of related databases, where necessary, towards the end of each section.

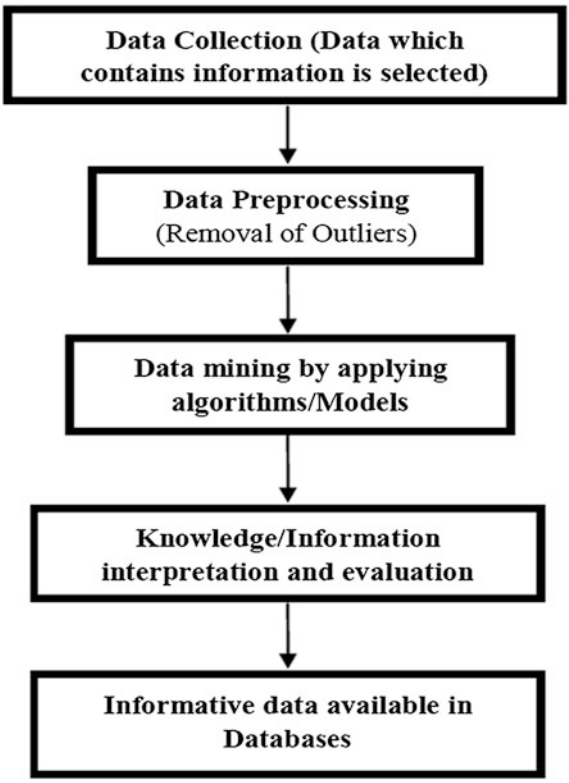

Fig. 1 Systematic diagram of KDD process

2 Biological Databases

Due to advancement in the high throughput sequencing techniques, sequencing of whole genome sequence of organisms are quite easy today and thereby leading to production of massive amount of data. Storage of large amount of biological data, retrieval and sharing of data among researchers are efficiently done by creation of databases which is a large, organized body of persistent data in a meaningful way. These databases were usually associated with computerized software designed to update, query, and retrieve the various components of the data stored in databases. There are different types of databases, which is based on the nature of the information being stored (e.g., sequences, structures, 2D gel or 3D structure images, and so on) as well as on the manner of data storage (e.g., whether in flat-files, tables in a relational database, or objects in an object-oriented databases. The sequence submission and storage of this information turn out to be freely accessible to the scientific world has directed to develop a number of databases worldwide. Respectively, every database becomes an autonomous illustration of a molecular unit of life.

Biological sequence database refers to a massive collection of data which have biological significance. Each biological molecule such as nucleic acids, proteins and polymers is identified by a unique key. The stored information can be used for future but also serves as an important input which for sequence and structure analyses. Biological Databases are mainly categorized into primary, secondary and composite databases and are discussed in detail in following sections.

2.1 Primary Databases

In primary database, the data related to sequence or structure are obtained through experiments such as yeast-two hybrid assay, affinity chromatography, XRD or NMR approaches. SWISS-PROT [7], UniProt [8–10], PIR [11], TrEMBL (translation of DNA sequences in EMBL) [7], GenBank [12], EMBL [13], DDBJ [14], Protein Databank PDB [15] and wwPDB (worldwide Protein DataBank) [16] are the well known examples of primary databases. A primary database is basically a collection of gene, protein sequence and structure information only. GenBank (USA), EMBL (Europe) and DDBJ (Japan) exchange data on a daily basis to ensure comprehensive coverage of these databases. SWISS-PROT is a protein sequence database which was established in 1986, collaboratively by University of Geneva and the EMBL [17]. SWISS-PROT includes annotations which has made it the database of choice for most of the researchers. The SWISS-PROT [17] contains information of its entries, which has been produced both by wet lab work as well dry lab. It is also interconnected to several other databases such as GenBank, EMBL, DDBJ, PDB and several other secondary protein databases. The protein data in SWISS-PROT mainly focuses only on model organisms and human only. On the other hand, the TrEMBL provides information on proteins from all

organisms [7]. Similarly, the PIR is one more inclusive collection of protein sequences which provides its user several attractive features like enabling to search for a protein molecule through text search. PIR also provides facility for web based analyses such as sequence alignment, identification of peptide molecules and peptide mass calculations [11, 17, 18]. The PIR Protein Sequence Database was developed by National Biomedical Research Foundation (NBRF) in 1960 s by Margaret Dayhoff. PIR is a database of protein sequences for investigating evolutionary relationships among proteins [11, 17, 18]. UniProt is another comprehensive collection of protein sequence which is available freely. The UniProt database is combination of SWISS-PROT, PIR and TrEMBL [8–10]. The worldwide Protein Data Bank (wwPDB) contains over 83,000 structures and they planned to provide each single 3D structure of protein molecules freely to the scientific community.

2.2 Secondary Databases

A secondary database is based on derived information from the primary database i.e. it contains information about the conserved sequence, active site residues of the protein families, patterns and motifs [19, 20]. Examples of secondary databases are SCOP [21], CATH [22], PROSITE [23], PRINTS [24] and eMOTIF [25]. The first secondary database to be developed was PROSITE, which is maintained by Swiss Institute of Bioinformatics. Within PROSITE, motifs are encoded as regular expressions which are also called patterns. PRINTS fingerprint database is another secondary database, which is maintained in University College London (UCL) and contains motifs as ungapped, unweighted local alignments [24]. The SCOP (Structural Classification of Proteins) database is maintained by MRC Laboratory and Centre for Protein Engineering which describes structural and evolutionary relationships among proteins for which structure are known [21]. In SCOP, proteins are classified in a hierarchical fashion to reflect their structural and evolutionary relationship. This hierarchy basically describes the family, superfamily and fold. The CATH (Class, Architecture, Topology, and Homology) is another secondary database which is a hierarchical classification of protein structures maintained at UCL [26]. CATH includes five levels within the hierarchy which are as follows:

- Class includes secondary structure content and packing. Four classes of domain are recognised: (i) mainly-α, (ii) mainly-β, (iii) α−β, which includes both alternating α/β and α + β structures, and (iv) Protein structures with low secondary structure content.
- Architecture includes arrangement of secondary structures, without connectivities; (e.g., barrel, roll, sandwich, etc.).
- Topology describes the overall shape and the connectivity of secondary structures.

- Homology includes domains that share 35 % sequence identity and are thought to share a common ancestor, i.e. are homologous.
- Sequence is the last level, where structures are clustered on the basis of sequence identity.

2.3 Composite Databases

Composite database is basically an amalgamation of variety of different primary database sources, which are meant to search multiple resources by putting different criteria in their search algorithm. Example of composite database is National Center for Biotechnology Information (NCBI) which is an extensive collection of nucleotide, protein sequence and many other databases providing free access to research community. NCBI provides interconnections between genetic sequence data, protein sequence data, structure data, phylogenetic tree based data, Genomes data and literature references. These links may also be between the same types of records in different databases, for example, literature articles in literature database Pubmed provide gene sequences, protein sequences, 3D structure, genome information and their links. Links between genetic sequences records are based on Blast sequence comparisons [27] while structure records are based on Vast structure comparisons [28]. NCBI includes one of the literature database called PubMed contains citations for biomedical literature from MEDLINE, journals and online books. NCBI also includes nucleotide sequence database called GenBank [12] which is collection of genome sequences of more than 2,50,000 species and these data can be retrieved by the NCBI's integrated retrieval system, i.e. Entrez, whereas the literature is easily accessible via PubMed [12, 29, 109]. It provides the information for related literature, organism, untranslated regions, exons/introns, repeat regions, coding regions, terminators, translations, promoters, bibliography etc. for each sequence. Sequence submission in GenBank can be done by individual laboratories along with large-scale genome sequencing projects. Protein sequence database in NCBI contains sequences from several sources which includes translations from annotated coding regions in GenBank, RefSeq. It also contains data records from SwissProt, PIR, PRF and PDB. The genome database in NCBI contains information on genomes which includes sequences, maps, chromosomes, assemblies as well as annotations. Protein structure databases at NCBI is called Molecular Modeling Database (MMDB) which contains data from experimentally resolved proteins structures, RNA and DNA molecules which are derived from the Protein Data Bank (PDB). MMDB also aid value-added features such as computationally identified 3D domains which can be used to identify similar 3D structures, as well as links to literature and information about chemicals/drug bound to the structures. Small chemical structure database integrated with NCBI is called Pubchem which includes small chemical structure and their biological activity (Fig. 2).

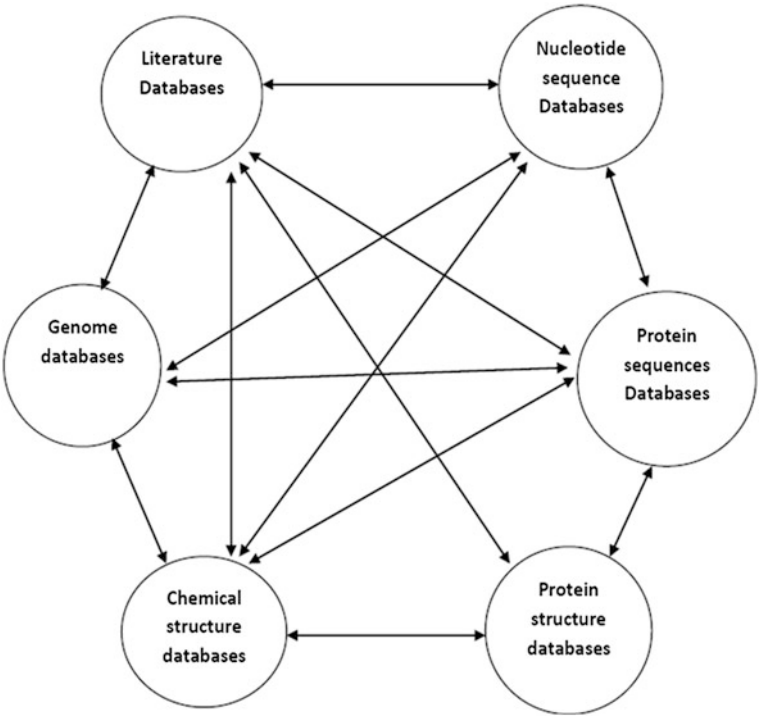

Fig. 2 *Circles* represent various databases; *straight* lines between *circles* represent links between different data types among different databases

NRDB (Non-Redundant DataBase) is another composite database which contains data from GenPept (derived from automatic GenBank CDS translations), SWISS-PROT, PIR, GenPeptupdate (the daily updates of GenPept), SPupdate (the weekly updates of SWISS-PROT) and PDB sequences. Similarly, INSD (International Nucleotide Sequence Database) is another composite database, which is collection of nucleic acid sequences from EMBL, GenBank and DDBJ. The UniProt (universal protein sequence database) which is also a composite database which contains sequences derived from various other databases such as PIRPSD, Swiss-Prot, and TrEMBL. In the same way, wwPDB (worldwide PDB) is a composite database of 3D structures which is maintained by RCSB (Research Collaboratory for Structural Bioinformatics), PDB, MSD and PDBj [30].

2.4 Specialized Databases

The Rfam database contains secondary structure of RNA molecules and their gene expression pattern. This database is introduced by the Wellcome Trust Sanger

Institute and it is similar to the Pfam database for annotating protein families [31]. There are numerous curated databases which are accessible worldwide such as IntAct contains data of various protein interactions. MINT (Molecular INTeraction database) is another curated database which is merged with IntAct database maintained by EMBL-EBI [32]. MINT is basically a database that stores information about protein-protein interactions derived from published articles [33]. For the metabolic pathway analysis in human, Reactome is one of the freely available databases which provide the diverse information regarding metabolic pathway and signal transduction pathways in human [34].

The Transporters Classification Database (TCDB) is a database of membrane transporters [35] which is based on Transport Classification (TC) system for the classification of protein similar to that of Enzyme Commission [36]. Similarly, the Carbohydrate-Active enzyme Database (CAZy) is a database of carbohydrate modifying enzymes and relevant information related to them. These enzymes are classified into different families which are based on the amino acid similarities or the presence of various catalytic domains [37].

Xenbase is a specialized database which contains genomic and biological informations about frogs [38], whereas the Saccharomyces Genome Database (SGD) provides complete information about yeast (*Saccharomyces cerevisiae*) which also offers web based bioinformatics tools to analyse the data available in SGD [39]. The SGD may be used to study functional interactions among gene sequence and gene products in other fungi including eukaryotes. Likewise, WormBase is a specialized database which is developed and maintained by an international consortium to make available precise, recent data related to the molecular biology of *C. elegans* and other associated nematodes. Wormbase also provides some web based tools for analysis of the stored information. Another up-to-date database is "FlyBase" devoted to provide information on the genes and genomes of *Drosophila melanogaster* along with various web based bioinformatics tools to search gene sequences, alleles, different phenotypes as well as images of the *Drosophila* species [40]. Similarly, wFleaBase provides information on genes and genomes for species of the genus *Daphnia* (water flea). *Daphnia* is considered as a model organism to study and understand the complex interplay between genome structures, gene expression and population level responses to chemicals and environmental changes. Although, wFleaBase contains data for all *Daphnia* species but the primary species are *D. pulex* and *D. magna*.

MetaCyc is a curated database of metabolic pathways which were taken from published literatures from all domains of life. It contains 2260 pathways from 2600 different organisms. MetaCyc contains pathways which are involved in both primary and secondary metabolism. It also includes their reactions, enzymes and associated genes [41]. PANTHER (Protein ANalysis THrough Evolutionary Relationships) is another metabolic pathway which consists of over 177 primarily signaling pathways. It contains different pathway components where component is basically a single protein/group of proteins in a given organism [42]. Pathway diagrams are interactive which also includes bioinformatics tools for visualizing gene expression data. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database of many

databases which was developed and maintained by Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo [43]. KEGG covers metabolic pathways in yeast, mouse and human etc. KEGG has expanded these days by the addition of signaling pathways for cell cycles and apoptosis. Reactome is a collection of metabolic and signaling pathways and their reactions [34]. Theses pathways and reactions can be viewed but not edited through a web browser. Although humans are the main organism catalogued, but this database also contains data for 22 other species such as mouse and rat.

TreeBASE is a collection of phylogenetic trees and the data associated to construct them. TreeBASE accepts all types of phylogenetic data for species tree as well as gene tree from all domains of life [44]. PhylomeDB [45] is another public database of phylogenetic information based on genes which allows users to explore evolutionary history of genes. Moreover, phylomeDB provides automated pipeline used to reconstruct trees of different genomes based on phylogenetic trees. Table 1 illustrates a list of genomic, protein sequences and specialized databases.

Table 1 List of gene and protein based databases, their description along with their webpage's URL

Databases	Description	Web link
Nucleotide databases		
DDBJ [14]	It is the member of International Nucleotide Sequence Databases (INSD) and is one of the biggest resources for nucleotide sequences	http://www.ddbj.nig.ac.jp/
European Nucleotide Archive	It captures and presents information relating to experimental workflows that are based around nucleotide sequencing	http://www.ebi.ac.uk/ena
GenBank [29]	It is the member of International Nucleotide Sequence Databases (INSD) and is a nucleotide sequence resource	http://www.ncbi.nlm.nih.gov/genbank/
Rfam [31]	A collection of RNA families, represented by multiple sequence alignments	http://rfam.xfam.org/
Protein databases		
InterPro [46]	Describes the protein families, conserved domains and actives sites	http://www.ebi.ac.uk/interpro/
Pfam [19]	Collection of protein families	http://pfam.xfam.org/
Prosite [23]	Provides information on protein families, conserved domains and actives sites of the proteins	http://prosite.expasy.org/
Protein Data Bank 2000 [15]	This is the most popular database of experimentally-determined structures of nucleic acids, proteins, and other complex assemblies	http://www.rcsb.org/pdb/home/home.do
Proteomics Identifications Database [47]	A public source, contain supporting evidence for post-translation modification and functional characterization of proteins and peptides	http://www.ebi.ac.uk/pride/archive/

(continued)

Table 1 (continued)

Databases	Description	Web link
SWISS-PROT [7]	A section of the UniProt Knowledgebase containing the manually annotated protein sequences	www.ebi.ac.uk/swissprot/
Uniprot [8–10]	One of the largest collection of protein sequences which contains curated as well as automated generated entries about protein sequences	http://www.uniprot.org/
PIR [11]	An integrated public resource to support genomic and proteomic research	http://pir.georgetown.edu/
Specialized databases		
Ensembl [48]	It is a database containing annotated genomes of eukaryotes, including human, mouse and other vertebrates	http://www.ensembl.org/index.html
DictyBase [49]	DictyBase is an online bioinformatics database for *Dictyostelium discoideum*	http://www.dictybase.org/
Medherb [50]	An important resource database for medicinally important herbs	https://www.medherbs.de/site/
TAIR [51]	The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular data for the model plant Arabidopsis thaliana. This database also provides information on gene structure, gene product, gene expression, genome maps, genetic and physical markers	http://www.arabidopsis.org/
TextPresso [52]	This database provides full text literature searches of model organism research which helps database curators to identify and extract biological entities which include new allele and gene names and human disease gene orthologs	http://www.textpresso.org/
Reactome [34]	A peer-reviewed resource of human biological processes i.e. metabolic pathways	http://www.reactome.org/
CMAP [53]	Complement Map Database is a novel and easily accessible research tool to assist the complement community and scientists. This database explores the complement network and discovers new connections	http://www.complement.us/labweb/cmap/
HMDB [54]	The Human Metabolome Database (HMDB) is the most comprehensive curated collection of human metabolite and human metabolism data in the world	http://www.hmdb.ca/
KEGG [43]	KEGG is a suite of databases and associated software for understanding and simulating higher-order functional behaviours of the cell or the organism from its genome information	http://www.genome.jp/kegg/

(continued)

Table 1 (continued)

Databases	Description	Web link
PID [55]	The Pathway Interaction Database (PID) is a collection of curated and peer-reviewed pathways. It is mainly composed of human molecular signaling and regulatory events and key cellular processes related to cancer	http://pid.nci.nih.gov/
SGMP [56]	The Signaling Gateway Molecule Pages (SGMP) database which provides highly structured data on proteins. It also identifies different functional states of the proteins which participate in signal transduction pathways	www.signaling-gateway.org/molecule

2.5 Drawbacks of Biological Databases

Many times life sciences researchers are interested not only in a few entries of a database but in huge amount of entries or large amount of data, which needs to be processed further, searching through web interfaces are not good options. To support large amount of data, the collection of relevant data and its processing must be automated. Therefore, each database should have programming options which make bioinformatics software developers to query and search databases from their own programs [57]. Modern database management systems such as ODBC (Open Database Connectivity) and JDBC (Java Database Connectivity) provide these web interfaces for bioinformatics programmers to query, search and analyze data. But the major limitation is that database providers allow public access only sometimes to these interfaces. These databases are DDBJ (DNA Data Bank of Japan) and KEGG (Kyoto Encyclopedia of Genes and Genomes). Apart from lacking in providing the programming interfaces, present biological databases also contain other limitations/drawbacks such as description of the database contents and authenticity of data produced and data sources. One of the drawbacks associated with these web interfaces is that these interfaces don't allow to be searched by using all fields in a database. These search modes such as 'and', 'or' and 'not' are not supported at all. Mostly theses modes are supported only in a limited way. Hence desired results for the query are not obtained. It is observed that in primary nucleotide and sequence databases, redundancy of many nucleotide and protein sequences are present, which should be removed. Biologists are usually not familiar with the principles of database design languages. Biologists are mostly ignorant about database query languages also and they should have knowledge of the database schema. In biological databases, in most of the cases flat files are used for data exchange which does not have standardized format. There are many formats for thousands of

biological databases which create problem in biological data preprocessing. Many types of information are often missing in biological databases which include functional annotations of genes and proteins. Many biological databases also provide missing information in terms of genotype phenotype relationships along with detailed pathway information and their reactions.

3 Gene Identification and Sequence Analyses

Due to lack of genome annotation and high-throughput experimental approaches computational gene prediction has become challenging and interesting area for bioinformatics/computational biology scientists. Gene prediction is very crucial especially for disease identification in human. Computational gene prediction can be categorized into two major classes which are ab initio methods and similarity/homology based methods [58]. These types of analyses are mainly useful for gene expression analysis. Gene expression is directly or indirectly related to the identification of promoter, terminator and untranslated regions. These regions are involved in the expression regulations, recognition of a transit peptide, introns/exons as well as an open reading frame (ORF). These regions are also involved in identification of variable regions which are used as signatures for various diagnostic purposes. Therefore, sequence analyses are one of the commonly used analyses for gene prediction in bioinformatics.

Gene prediction is relatively more difficult in eukaryotes as compared to prokaryotes due to presence of introns. Homology based gene predictions depend on complementary DNA (cDNA) and Expressed Sequence Tags (ESTs), though, the cDNA/ESTs information is often unusual and incomplete, and thus makes the task of finding novel genes extremely difficult. Homology based on sequence based information has been shown to be useful for better identification of prokaryotic and eukaryotic genes with higher accuracy. Local and global sequence alignments are performed by BLAST and NEEDLE respectively which is widely used in homology/similarity based gene prediction. These days protein homology has been introduced in many gene prediction programmes such as GENEWISE [59] and GENOMESCAN [60] GeneParser [61] and GRAIL [62]. Novel gene finding is often possible by ab initio gene identification. Examples of ab initio programs are GENSCAN [63], GENIE [64], HMMGene [65] and GENEID [66]. These methods were used in *Drosophila melanogaster* where it showed a very low rate of false-positive. These methods also predict 88 % of exons (already verified) and 90 % of the coding sequences [67]. Due to high accuracy, this methodology could be used for annotating large genomic sequences and prediction of new genes.

Recently, Lencz et al. identified an intergenic single nucleotide polymorphism (SNP; rs11098403) at chromosome 4q26 which was linked with schizophrenia and bipolar disorder. They performed a genome wide association study (GWAS) which was coupled with cDNA as well as RNA Seq on a set of 23,191 individuals [68]. Similarly, Peng and co-workers predicted 31,987 genes from *Phyllostachys heterocycle* draft genome by gene prediction approaches based on FgeneSH ++ [69]. Sequence analyses refer to the understanding of various features of biomolecules like nucleic acids and proteins, which are responsible for providing unique function(s). The first step is retrieval of sequences from public databases which are subjected to analysis by various tools which help in the prediction of specific features which might be associated to their function, structure, evolutionary relationship or identification of homologs with high accuracy. The database should be selected depending upon the nature of analysis. For example, Entrez of PubMed [70] allows one to search about different patterns in the given data. Also, pattern discovery can be performed by Expression Profiler [71], GeneQ [72] which allow scientists to search out different patterns in the given data. A different set of databases are dedicated to carry out sequence comparison like BLAST (Basic Local Alignment Search Tool) [27], ClustalW [73], for data visualization Jalview [74], GeneView [75], TreeView [76] and Genes-Graphs [77] allowing researchers to visualize data in graphic representation. Table 2 illustrates a list of databases used in primary sequence analyses.

4 Phylogenetic Analyses

Phylogenetic analysis help to determine the evolutionary relationship among a group of related organism or related genes and proteins [83, 84], to predict the specific feature of a molecule with unknown functions, to track the gene flow and also to determine genetic relatedness [85]. Phylogenetic analysis is mainly based similarity at sequence level i.e. higher is the similarity; the closer will be the organisms on a tree. Phylogenetic tree is constructed by various methods which are distance, parsimony and maximum likelihood methods. None of the methods is perfect; each one has its own strengths and weaknesses. For example, the distance based methods performs average whereas the maximum parsimony and maximum likelihood methods are accurate. The major disadvantage of maximum parsimony and maximum likelihood methods is these methods takes more time to run as compared to distance based methods [86]. Among the distance-matrix methods Neighbour Joining (NJ) or Unweighted Pair Group Method with Arithmetic mean (UPGMA) are the simplest. Table 2 illustrates a list of phylogenetic analyses programmes. (Table 3).

Table 2 List of gene identification and sequence analyses programmes and their description along with their webpage's URL

Software tools	Description	Web link
BLAST [27]	Used for database sequence searching for protein and DNA homologs	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Clustal Omega [78]	Used for Multiple sequence alignments of DNA and protein sequences	http://www.ebi.ac.uk/Tools/msa/clustalo/
Genscan [63]	Used to predict the exon-intron sites in genomic sequences	http://genes.mit.edu/GENSCAN.html
HMMER [79]	A tool which is used for homologous protein sequence search	http://hmmer.janelia.org/
JIGSAW [80]	Used for identification of gene and their splice site prediction in the selected DNA sequences	http://cbcb.umd.edu/software/jigsaw
novoSNP [81]	Used to find the single nucleotide variation in the DNA sequence	http://www.molgen.ua.ac.be/bioinfo/novosnp/
ORF Finder	The putative genes may be subjected to this tool to find Open Reading Frame (ORF)	http://www.ncbi.nlm.nih.gov/projects/gorf/
PPP	Prokaryotic promoter prediction tool used to predict the promoter sequences present up-stream in the gene	http://bioinformatics.biol.rug.nl/websoftware/ppp/ppp_start.php
ProtParam [82]	Used to predict the physico-chemical properties of proteins	http://web.expasy.org/protparam/
Sequerome	The tool which is mainly used for sequence profiling	www.bioinformatics.org/sequerome/
Softberry Tools	Several tools are specialized in annotation of animal, plant, and bacterial genomes along with the structure and function prediction of RNA and proteins	http://www.softberry.com/
Virtual Foorprint	Whole prokaryotic genome (with one regular pattern) may be analyzed using this program along with promoter regions with several regulator patterns	http://www.prodoric.de/vfp/
WebGeSTer	The database which is composed of sequences of transcription terminator sequences and is used to predict the termination sites of the genes during transcription	http://pallab.serc.iisc.ernet.in/gester/dbsearch.php

In functional genomics where a function of a particular gene is not known phylogenetic analysis is used to find their relative genes which ultimately help to the identification their function and other features of that particular gene.

Table 3 List of phylogenetic analyses programmes and their description along with their webpage's URL

Software tools	Description	Web link
JStree	An open-source library for viewing and editing phylogenetic trees for presentation improvement	www.jstree.com/
MEGA [87]	A tool to construct phylogenetic trees to by parsimony, distance based and maximum likehood based tree construction and to study the evolutionary relationships	http://www.megasoftware.net/
MOLPHY	Phylogenetic analysis tool based on maximum likelihood method	http://www.ism.ac.jp/ismlib/softother.e.html
PAML	A phylogenetic analysis tool based on maximum likelihood	http://abacus.gene.ucl.ac.uk/software/paml.html
PHYLIP	A complete software package for phylogenetic tree generation for DNA and protein sequences	http://evolution.genetics.washington.edu/phylip.html
TreeView [76]	It is a tool for visualisation of the phylogenetic trees	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

5 Predicting Protein Structure and Function

Protein molecules initiate their life as amorphous amino acid strings, which finally fold up into a three-dimensional (3D) structure. The folding of the protein into a correct topology is needed for proteins to perform its biological functions. Usually, 3D structures are mostly determined by X-ray crystallography and NMR which are costly, difficult and time taking. X ray crystallography method fails if we do not get good crystals. Moreover NMR is limited to small proteins [88]. There are very few structures submitted monthly using NMR and XRD in NCBI. Correct prediction of secondary and tertiary structure of proteins is one of the challenging tasks for bioinformatics/computational biologist till date. Predicting the correct secondary structure is the key to predict not only a good/satisfactory tertiary structure of the protein but also helps in prediction of protein function [88]. Protein structure prediction is classified into three categories: (i) Ab initio modeling [89] (ii) Threading or Fold recognition [90] and (iii) Homology or Comparative modeling (Šali and Blundell 1993 [91]. Threading and comparative modeling build protein models by aligning query sequences with known structures which are determined by X-ray crystallography or NMR. When templates having identity ≥ 30 % are found, high resolution models could be built by the template-based methods. If templates are not available from the protein data bank (PDB), these models are built from scratch, i.e. ab initio modeling [92]. Homology modeling is the most accurate prediction method so far and it is used frequently. In one of our studies good quality homology models of superoxide dismutase (SOD) has been obtained by Modeller software package in antarctic cyanobacterium *Nostoc*

Table 4 List of protein structure and function predictions programmes their description along with their webpage's URL

Software tools	Description	Web link
RaptorX [94]	It facilitates the user to predict protein structure based on either a single- or multi-template threading	http://raptorx.uchicago.edu/
JPRED [95]	Used to predict secondary structures of proteins	http://www.compbio.dundee.ac.uk/www-jpred/
HMMSTR [96]	A hidden Markov model for the prediction of sequence-structure correlations in proteins	http://www.bioinfo.rpi.edu/bystrc/hmmstr/server.php
APSSP2 [97]	Predicts the secondary structure of proteins	http://omictools.com/apssp2-s7458.html
MODELLER [98]	Predicts 3D structure of protein based on comparative modelling	https://salilab.org/modeller/
Phyre and Phyre2 [99]	Web-based servers for protein structure prediction by threading algorithm	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index

commune which aids to cope with environmental stresses prevailing at its natural habitat [93]. Bioinformatics tools can also identify secondary structure elements such as helices, sheets and coils. Protein tertiary structures are stabilized by the presence of helices, sheets and coils which play an important role in establishing weaker electrostatic forces. Table 4 illustrates a list of tools to predict the secondary structure of protein molecules.

6 Predicting Molecular Interactions

Biomolecules interacting with each other affect various biological activities which has nowadays become one of the popular areas for research [100]. For example, protein-protein interaction, protein-DNA or protein-RNA interaction etc. Protein-protein interactions play an essential role in various cellular activities like signalling and transportation. Protein-protein interactions also play major role in homeostasis, cellular metabolism etc. [101]. In this regard, bioinformatics helps to predict the 3D structure of proteins and also helps in predicting the interaction pattern between different biomolecules. These predictions are based on various parameters such as interface size, amino acid position, types of chemical groups involved. These predictions are also based on vander wall forces, electrostatic interaction and hydrogen bonds. Table 5 illustrates a list of tools to study protein-protein interactions.

Table 5 List of molecular interactions database and programmes, their description along with their webpage's URL

Software tools	Description	Web link
SMART [102]	A Simple Modular Architecture Retrieval Tool; describes multiple information about the protein query	http://smart.embl-heidelberg.de/
AutoDock [103]	Predicts protein-ligand interaction and is considered as reliable tool	http://autodock.scripps.edu/
HADDOCK [104]	Describes the modelling and interaction of bio-molecular complexes such as protein-protein, protein-DNA	http://haddock.science.uu.nl/
STRING [105]	A database of both known and predicted protein interactions	http://string-db.org/
IntAct [32]	It is an open source database system which is used for molecular interaction data	http://www.ebi.ac.uk/intact/
Graemlin [106]	It is capable of scalable multiple network alignment with its functional evolution model that allows both the generalization of existing alignment scoring schemes. This tool also model the location of conserved network topologies other than protein complexes and metabolic pathways	http://graemlin.stanford.edu/
PathBLAST [107]	This tool is to search protein-protein interaction network of the any selected organism and extracts all interaction pathways that align with the query	http://www.pathblast.org/
CFinder [108]	This tool is capable of finding and visualizing the overlapping dense groups of nodes in networks, and quantitative description of the evolution of social groups	http://www.cfinder.org/

7 Discussion, Conclusion and Future Prospects

Bioinformatics has emerged as a challenging discipline which has developed very fast in the last few years due to generation of large amount of data generated by various genome sequencing projects. Such a large amount of data needs pre-processing to extract useful knowledge/information by data mining techniques. These processed data are not only stored but also retrieved in a meaningful manner from biological databases. These biological databases containing nucleotide and protein sequences are called primary databases. These primary databases have a drawback that these databases contain redundant sequences. Secondary database has solved this issue to a greater extent which contain derived information from primary databases and redundancy is also minimized at lowest in Swiss-Prot database. Composite databases e.g. NCBI provides better search criteria to search multiple primary resources at a time. NCBI also provides the linking with literature, structure, chemical molecules, genome information, gene and protein sequences databases. Apart from these databases, various specialized databases are also available these days which provide informations about protein-protein interactions,

protein families, experimentally known metabolic pathways, genome sequence, protein structure and phylogenetic tree for evolutionary relationship. These databases also have few drawbacks e.g. lack of description of data contained, redundancy of sequences etc. One of the major drawbacks of most of the databases is that they don't provide the programming interface so that researchers can write their programmes to download and process huge amount of stored data from the database. Bioinformatics is not only used in designing the biological databases but also used in developing software tools for sequence, structure and evolutionary analysis of genes/proteins etc. which save our time, energy and cost in biological research. A number of bioinformatics softwares were designed to predict the correct genes in genomic sequences which use various machine learning approaches like artificial intelligence, genetic algorithm, support vector machine, hidden markov model, dynamic programming etc. However, the best predictors are based on hybrid methods which use more than one machine learning approaches to predict the correct genes. Bioinformatics tools were also developed to construct parsimony, distance based and maximum likelihood based trees to explore the evolutionary relationship among species. Parsimony method is successful when sequence identity is high while maximum likelihood performs well when sequence variation is high. Bioinformatics have proved to be a boon in structure based drug design by predicting the structure of drug targets immaterial of whether template structure are available in PDB or not by different approaches. Homology modelling proved the best predictor among all the methods. Moreover, bioinformatics tools also predict protein-protein interactions which play an essential role in various cellular activities like signalling, transportation, homeostasis, cellular metabolism and also various biochemical processes. It can also be expected, based on the developments in the field of bioinformatics, that the bioinformatics tools and software packages would be able to give more specific, more accurate and more reliable in upcoming years. In future the field of bioinformatics will contribute in functional understanding of whole genome of organisms which will lead to enhanced discovery of gene expression, their interaction pattern, individualised gene therapy and new drug discovery. Thus, bioinformatics and other scientific disciplines should move together in order to flourish for the welfare of humanity.

References

1. Hogeweg, P.: The roots of bioinformatics in theoretical biology. PLoS Comput. Biol. **7**, e1002021 (2011)
2. Neufeld, L., Cornog, M.: Database history: from dinosaurs to compact discs. J Am. Soc. Inf. Sci. **37**, 183–190 (1999)
3. Chen, M.-S., Han, J., Yu, P.S.: Data mining: an overview from a database perspective. IEEE Trans. Knowl. Data Eng. **8**(6), 866–883 (1996)
4. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., et al. (eds.): Advance in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, Cambridge (1996)

5. Ouzounis, C.A., Valencia, A.: Early bioinformatics: the birth of a discipline-a personal view. Bioinformatics **19**, 2176–2190 (2003)
6. Hassanie, A.E.: Classification and feature selection of breast cancer data based on decision tree algorithm. Stud. Inform. Control **12**(1), 33–40
7. Boeckmann, B., Bairoch, A., Apweiler, R., et al.: The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**, 365–370 (2003)
8. UniProt Consortium: The universal protein resource (UniProt). Nucleic Acids Res. **36**, D190–D195 (2008)
9. UniProt Consortium: The universal protein resource (UniProt) in 2010. Nucleic Acids Res. **38**, D142–D148 (2010)
10. UniProt Consortium: Activities at the universal protein resource (UniProt). Nucleic Acids Res. **42**, D191–D198 (2014)
11. Wu, C.H., Yeh, L.S., Huang, H., et al.: The protein information resource. Nucleic Acids Res. **31**(1), 345–347 (2003)
12. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., et al.: GenBank. Nucleic Acids Res. **36**, D25–D30 (2008)
13. Kanz, C., Aldebert, P., Althorpe, N., et al.: The EMBL nucleotide sequence database. Nucleic Acids Res. **33**, D29–D33 (2005)
14. Miyazaki, S., Sugawara, H., Gojobori, T., et al.: DNA data bank of Japan (DDBJ) in xml. Nucleic Acids Res. **31**, 13–16 (2003)
15. Berman, H.M., Westbrook, J., Feng, Z., et al.: The protein data bank. Nucleic Acids Res. **28**, 235–242 (2000)
16. Berman, H., Henrick, K., Nakamura, H.: Announcing the worldwide protein data bank. Nat. Struct. Mol. Biol. **10**, 980 (2003)
17. Barker, W.C., Garavelli, J.S., Huang, H., et al.: The protein information resource (PIR). Nucleic Acids Res. **28**, 41–44 (2000)
18. Barker, W.C., Garavelli, J.S., Haft, D.H., et al.: The PIR-international protein sequence database. Nucleic Acids Res. **26**(1), 27–32 (1998)
19. Finn, R.D., Bateman, A., Clements, J., et al.: Pfam: the protein families database. Nucleic Acids Res. **42**, D222–D230 (2014)
20. Gonzalez, S., Binato, R., Guida, L., et al.: Conserved transcription factor binding sites suggest an activator basal promoter and a distal inhibitor in the galanin gene promoter in mouse ES cells. Gene **538**, 228–234 (2014)
21. Murzin, A.G., Brenner, S.E., Hubbard, T., et al.: Scop: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**, 536–540 (1995)
22. Pearl, F., Todd, A., Sillitoe, I., et al.: The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res. **33**, D247–D251 (2005)
23. Sigrist, C.J., de Castro, E., Cerutti, L., et al.: New and continuing developments at PROSITE. Nucleic Acids Res. **41**, D344–D347 (2013)
24. Attwood, T.K., Beck, M.E., Flower, D.R., et al.: The PRINTS protein fingerprint database in its fifth year. Nucleic Acids Res. **26**(1), 304–308 (1998)
25. Huang, J.Y., Brutlag, D.L.: The EMOTIF database. Nucleic Acids Res. **29**, 202–204 (2001)
26. Orengo, C.A., Michie, A.D., Jones, S., et al.: CATH—a hierarchic classification of protein domain structures. Structure **5**(8), 1093–1108 (1997)
27. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389–3402 (1997)
28. Gibrat, J.F., Madej, T., Bryant, S.H.: Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. **6**, 377–385 (1996)
29. Benson, D.A., Karsch-Mizrachi, I., Clark, K., et al.: GenBank. Nucleic Acids Res. **40**, D48–D53 (2012)
30. Kinjo, A.R., Suzuki, H., Yamashita, R., et al.: Protein data bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Res. **40**, D453–D460 (2012)

31. Burge, S.W., Daub, J., Eberhardt, R., et al.: Rfam 11.0: 10 years of RNA famlies. Nucleic Acids Res. **4**, D226–D232 (2013)
32. Orchard, S., Ammari, M., Aranda, B., et al.: The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. **42**, D358–D363 (2014)
33. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., et al.: MINT: a molecular INTeraction database. FEBS Lett. **513**, 135–140 (2002)
34. Joshi-Tope, G., Gillespie, M., Vastrik, I., et al.: Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. **33**, D428–D432 (2005)
35. Saier Jr, M.H., Tran, C.V., Barabote, R.D.: TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Res. **34**, D181–D186 (2006)
36. Saier Jr, M.H., Reddy, V.S., Tamang, D.G., et al.: The transporter classification database. Nucleic Acids Res. **42**, D251–D258 (2014)
37. Lombard, V., Golaconda, H., Drula, R.E., et al.: The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. **42**(D1), D490–D495 (2014)
38. Bowes, J.B., Snyder, K.A., Segerdell, E., et al.: Xenbase: gene expression and improved integration. Nucleic Acids Res. **38**, D607–D612 (2010)
39. Cherry, J.M., Hong, E.L., Amundsen, C., et al.: Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res. **40**(D): D700–705 (2012)
40. St. Pierre, S.E., Ponting, L., Stefancsik, R., et al.: FlyBase 102–advanced approaches to interrogating FlyBase. Nucleic Acids Res. **42**: D780–788 (2014)
41. Caspi, R., Altman, T., Billington, R.: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. **42**, D459–D471 (2014)
42. Thomas, P.D., Campbell, M.J., Kejariwal, A., et al.: PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. **13**(9), 2129–2141 (2003)
43. Kanehisa, M.: The KEGG database. Silico Simul. Biological Process. **247**, 91–103 (2002)
44. Morell, V.: TreeBASE: the roots of phylogeny. Science **273**, 569 (1996)
45. Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L.P., et al.: PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. **42**, D897–D902 (2014)
46. Mitchell, A., Chang, H.-Y., Daugherty, L., et al.: InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. **43**, D213–D221 (2015)
47. Martens, L., Hermjakob, H., Jones, P., et al.: PRIDE: the proteomics identifications database. Proteomics **5**, 3537–3545 (2005)
48. Flicek, P., Amode, M.R., Barrell, D., et al.: Ensembl Nucleic Acids Res. **40**, D84–D90 (2012)
49. Gaudet, P., Fey, P., Basu, S., et al.: dictyBase update 2011: web 2.0 functionality and the initial steps towards a genome portal for the Amoebozoa. Nucleic Acids Res. **39**, D620–D624 (2011)
50. Rajoka, M.I., Idrees, S., Khalid, S., et al.: Medherb: an interactive bioinformatics database and analysis resource for medicinally important herbs. Curr. Bioinformatics **9**, 23–27 (2014)
51. Lamesch, P., Berardini, T.Z., Li, D., et al.: The arabidopsis information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. **2011**, 1–9 (2011)
52. Muller, H.M., Kenny, E.E., Sternberg, P.W.: Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. **2**(11), e309 (2004)
53. Yang, K., Dinasarapu, A.R., Reis, E.S., et al.: CMAP: complement map database. Bioinformatics **29**(14), 1832–1833 (2013)
54. Wishart, D.S., Tzur, D., Knox, C., et al.: HMDB: the human metabolome database. Nucleic Acids Res. **35**, D521–D526 (2007)
55. Schaefer, C.F., Anthony, K., Krupa, S., et al.: PID: the pathway interaction database. Nucleic Acids Res. **37**, D674–D679 (2009)

56. Dinasarapu, A.R., Saunders, B., Ozerlat, I., et al.: Signaling gateway molecule pages–a data model perspective. Bioinformatics **27**(12), 1736–1738 (2011)
57. Philippi, S., Köhler, J.: Addressing the problems with life-science databases for traditional uses and systems biology. Nat. Rev. Genet. **7**(6), 482–488 (2000)
58. Lewis, S., Ashburner, M., Reese, M.G.: Annotating eukaryote genomes. Curr. Opin. Struct. Biol. **10**, 349–354 (2000)
59. Birney, E., Durbin, R.: Using GeneWise in the Drosophila annotation experiment. Genome Res. **10**, 547–548 (2000)
60. Yeh, R.-F., Lim, L.P., Burge, C.B.: Computational inference of homologous gene structures in the human genome. Genome Res. **11**, 803–816 (2001)
61. Snyder, E.E., Stormo, G.D.: Identification of protein coding regions in genomic DNA. J. Mol. Biol. **248**, 1–18 (1995)
62. Uberbacher, E.C., Mural, R.J.: Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc. Natl. Acad. Sci. USA **88**, 11261–11265 (1991)
63. Burge, C., Karlin, S.: Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268**, 78–94 (1997)
64. Kulp, D., Haussler, D., Reese, M.G., et al.: A generalized hidden Markov model for the recognition of human genes in DNA. In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology, vol. 4, pp. 134–142 (1996)
65. Krogh, A.: Two methods for improving performance of an HMM and their application for gene-finding. In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB‘97), vol. 5, pp. 179–186 (1997)
66. Parra, G., Blanco, E., Guig´o, R.: GeneID in Drosophila. Genome Res. **10**, 391–393 (2000)
67. Salamov, A.A., Solovyev, V.V.: Ab initio gene finding in Drosophila genomic DNA. Genome Res. **10**, 516–522 (2000)
68. Lencz, T., Guha, S., Liu, C., Rosenfeld, J., et al.: Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. Nat. Commun. **4**, 2739 (2013)
69. Peng, Z., Lu, Y., Li, L., et al.: The draft genome of the fastgrowing non-timber forest species moso bamboo (Phyllostachys heterocycla). Nat. Genet. **45**, 456–461 (2013)
70. Geer, R.C., Sayers, E.W.: Entrez: making use of its power. Brief Bioinform. **4**, 179–184 (2003)
71. Parmigiani, G., Garrett, E.S., Irizarry, R.A., et al.: The analysis of gene expression data: an overview of methods and software. Springer, New York (2003)
72. Hoersch, S., Leroy, C., Brown, N.P., et al.: The GeneQuiz web server: protein functional analysis through the web. Trends Biochem. Sci. **25**, 33–35 (2000)
73. Thompson, J.D., Higgins, D.G., Gibson, T.J.: 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 (1994)
74. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., et al.: Jalview version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics **25**, 1189–1191 (2009)
75. Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S., Leser, U.: GeneView: a comprehensive semantic search engine for PubMed. Nucleic Acids Res. **40**, W585–W591 (2012)
76. Page, R.D.M.: TREEVIEW: An application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12**, 357–358 (1996)
77. Zhang, Y., Phillips, C.A., Rogers, G.L., et al.: On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. BMC Bioinformatics **15**, 110 (2014)
78. Sievers, F., Wilm, A., Dineen, D.G., et al.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. **7**, 539 (2011)
79. Finn, R.D., Clements, J., Eddy, S.R.: HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. **39**, W29–W37 (2011)
80. Allen, J.E., Salzberg, S.L.: JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics **21**(18), 3596–3603 (2005)

81. Weckx, S., Del-Favero, J., Rademakers, R.: novoSNP, a novel computational tool for sequence variation discovery. Genome Res. **15**(3), 436–442 (2005)
82. Gasteiger, E., Hoogland, C., Gattiker, A., et al.: Protein identification and analysis tools on the expasy server. In: Walker, J.M. (ed.) The Proteomics Protocols Handbook. Humana Press, p 571–607 (2005)
83. Kanchan, S., Mehrotra, R., Chowdhury, S.: Evolutionary pattern of four representative DNA repair proteins across six model organisms: an in silico analysis. Netw. Model Anal. Health Inf. Bioinform **3**, 70 (2014)
84. Kanchan, S., Mehrotra, R., Chowdhury, S.: In Silico analysis of the Endonuclease III protein family identifies key residues and processes during evolution. J. Mol. Evol. **81**(1–2), 54–67 (2015)
85. Khan, F.A., Phillips, C.D., Baker, R.J.: Timeframes of speciation, reticulation, and hybridization in the bulldog bat explained through phylogenetic analyses of all genetic transmission elements. Syst. Biol. **63**, 96–110 (2014)
86. Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2–approximately maximum likelihood trees for large alignments. PLoS ONE **5**, e9490 (2010)
87. Kumar, S., Tamura, K., Nei, M.: MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput. Appl. Biosci. **10**, 189–191 (1994)
88. Huang, T., He, Z.S., Cui, W.R., et al.: A sequence-based approach for predicting protein disordered regions. Protein Pept. Lett. **20**, 243–248 (2013)
89. Liwo, A., Lee, J., Ripoll, D.R., et al.: Protein structure predictionby global optimization of a potential energy function. Proc. Natl. Acad. Sci. USA **96**(10), 5482–5485 (1999)
90. Bowie, J., Luthy, R., Eisenberg, D.: A method to identify protein sequences that fold into a known three-dimensional structure. Science **253**(5016), 164–170 (1991)
91. Šali, A., Blundell, T.L.: Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. **234**(3), 779–815 (1993)
92. Kesheri, M., Kanchan, S., Chowdhury, S., et al.: Secondary and tertiary structure prediction of proteins: a bioinformatic approach. In: Zhu, Q., Azar, A.T. (eds.) Complex System Modelling and Control Through Intelligent Soft Computations, pp. 541–569. Springer International Publishing, Switzerland (2015)
93. Kesheri, M., Kanchan, S., Richa, et al.: Isolation and in silico analysis of Fe-superoxide dismutase in the cyanobacterium Nostoc commune. Gene. **553**(2): 117–125 (2014)
94. Källberg, M., Wang, H., Wang, S., et al.: Template-based protein structure modeling using the RaptorX web server. Nat. Protoc. **7**, 1511–1522 (2012)
95. Cuff, J.A., Clamp, M.E., Siddiqui, A.S., et al.: JPred: a consensus secondary structure prediction server. Bioinformatics **14**, 892–893 (1998)
96. Bystroff, C., Thorsson, V., Baker, D.: HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J. Mol. Biol. **301**, 173–190 (2000)
97. Raghava, G.: APSSP2: a combination method for protein secondary structure prediction based on neural network and example based learning. CASP5 A-132 (2002)
98. Eswar, N., Eramian, D., Webb, B., et al.: Protein structure modeling with MODELLER. Methods Mol. Biol. **426**, 145–159 (2008)
99. Kelley, L.A., Sternberg, M.J.: Protein structure prediction on the web: a case study using the Phyre server. Nat. Protoc. **4**, 363–371 (2009)
100. Wang, L., Huang, C., Yang, M.Q., et al.: BindN + for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst. Biol. **4**(1), S3 (2010)
101. Vinayagam, A., Zirin, J., Roesel, C., et al.: Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. Nat. Methods **11**, 94–99 (2014)
102. Schultz, J., Copley, R.R., Doerks, T., et al.: SMART: A Web-based tool for the study of genetically mobile domains. Nucleic Acids Res. **28**, 231–234 (2000)
103. Morris, G.M., Huey, R., Lindstrom, W., et al.: AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J. Comput. Chem. **30**, 2785–2791 (2009)
104. De Vries, S.J., van Dijk, M., Bonvin, A.M.: The HADDOCK web server for data driven biomolecular docking. Nat. Protoc. **5**, 883–897 (2010)

105. Franceschini, A., Szklarczyk, D., Frankild, S., et al.: STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. **41**, D808–D815 (2013)
106. Flannick, J., Novak, A., Srinivasan, B.S., et al.: Graemlin: general and robust alignment of multiple large interaction networks. Genome Res. **16**, 1169–1181 (2006)
107. Kelley, B.P., Yuan, B., Lewitter, F., et al.: PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res. **32**, W83–W88 (2004)
108. Adamcsek, B., Palla, G., Farkas, I.J., et al.: CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics **22**, 1021–1023 (2006)
109. Fathy, M.E., Hussein, A.S., Tolba, M.F.: Fundamental matrix estimation: a study of error criteria. Pattern Recogn. Lett. **32**(2), 383–391 (2011)