Journal of
Data Mining in Genomics & Proteomics
ISSN: 2153-0602

Mehmood et al., J Data Mining Genomics Proteomics 2014, 5:2
DOI: 10.4172/2153-0602.1000158

Review Article Open Access

Use of Bioinformatics Tools in Different Spheres of Life Sciences

Muhammad Aamer Mehmood[1], Ujala Sehar[1] and Niaz Ahmad[2]*

[1]Bioenergy Research Centre, Department of Bioinformatics and Biotechnology, Government College University Faisalabad, Faisalabad-38000, Pakistan
[2]Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering (NIBGE), Jhang Road, Faisalabad-38000, Pakistan

Abstract

The pace, by which scientific knowledge is being produced and shared today, was never been so fast in the past. Different areas of science are getting closer to each other to give rise new disciplines. Bioinformatics is one of such newly emerging fields, which makes use of computer, mathematics and statistics in molecular biology to archive, retrieve, and analyse biological data. Although yet at infancy, it has become one of the fastest growing fields, and quickly established itself as an integral component of any biological research activity. It is getting popular due to its ability to analyse huge amount of biological data quickly and cost-effectively. Bioinformatics can assist a biologist to extract valuable information from biological data providing various web- and/or computer-based tools, the majority of which are freely available. The present review gives a comprehensive summary of some of these tools available to a life scientist to analyse biological data. Exclusively this review will focus on those areas of biological research, which can be greatly assisted by such tools like analysing a DNA and protein sequence to identify various features, prediction of 3D structure of protein molecules, to study molecular interactions, and to perform simulations to mimic a biological phenomenon to extract useful information from the biological data.

Keywords: Bioinformatics; Life sciences; Sequence analysis; Phylogeny; Structure prediction; Molecular interaction; Molecular dynamic simulations

Abbreviations: ADMET: Absorption Distribution Metabolism Excretion and Toxicity; ANN: Artificial Neural Network; BLAST: Basic Local Alignment Search Tool; CADD: Compute Aided Drug Design; cDNA: Complementary DNA; CDS: Coding Sequence; ESTs: Expressed Sequence Tags; GWSA: Genome Wide Sequence Analysis; HMM: Hidden Markov Model; HTS: High Throughput Screening; MSA: Multiple Sequence Alignment; NCBI: National Centre for Biotechnology Information; NJ: Neighbour Joining; NMR: Nuclear Magnetic Resonance; ORF: Open Reading Frame; PDB: Protein Data Bank; SNP: Single Nucleotide Polymorphism; UPGMA: Unweighted Pair Group Method with Arithmetic Mean; XRD: X-ray Crystallography Diffraction

Introduction

Bioinformatics is an interdisciplinary science, emerged by the combination of various other disciplines like biology, mathematics, computer science, and statistics, to develop methods for storage, retrieval and analyses of biological data [1]. Paulien Hogeweg, a Dutch system-biologist, was the first person who used the term "Bioinformatics" in 1970, referring to the use of information technology for studying biological systems [2,3]. The launch of user-friendly interactive automated modeling along with the creation of SWISS-MODEL server around 18 years ago [4] resulted in massive growth of this discipline. Since then, it has become an essential part of biological sciences to process biological data at a much faster rate with the databases and informatics working at the backend.

Computational tools are routinely used for characterization of genes, determining structural and physiochemical properties of proteins, phylogenetic analyses, and performing simulations to study how biomolecule interact in a living cell. Although these tools cannot generate information as reliable as experimentation, which is expensive, time consuming and tedious, however, the *in silico* analyses can still facilitate to reach an informed decision for conducting a costly experiment. For example, a druggable molecule must have certain ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties to pass through clinical trials. If a compound does not have required ADMETs, it is likely to be rejected. To avoid such failures, different bioinformatics tools have been developed to predict ADMET properties, which allow researchers to screen a large number of compounds to select most druggable molecule before launching of clinical trials [5]. Earlier, a number of reviews on various specialized aspects of bioinformatics have been written [6-8]. However, none of these articles makes it suitable for a scientist who does not belong to computational biology. Here, we take the opportunity to introduce various tools of bioinformatics to a non-specialist reader to help extract useful information regarding his/her project. Therefore, we have selected only those areas where these tools could be highly useful to obtain useful information from biological data. These areas include analyses of DNA/protein sequences, phylogenetic studies, predicting 3D structures of protein molecules, molecular interactions and simulations as well as drug designing. The organization of text in each section starts from a simplistic overview of each area followed by key reports from literature and a tabulated summary of related tools, where necessary, towards the end of each section.

Gene Identification and Sequence Analyses

Sequence analyses refer to the understanding of different features of a biomolecule like nucleic acid or protein, which give to it its unique function(s). First, the sequences of corresponding molecule(s) are retrieved from public databases. After refinement, if needed, they are subjected to various tools that enable prediction of their features related to their function, structure, evolutionary history or identification of homologues with a great accuracy. Which tool should be used for what

***Corresponding author:** Niaz Ahmad, Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering (NIBGE), Jhang Road, Faisalabad-38000, Pakistan, Tel: 92 (0)333 662; E-mail: niazbloch@yahoo.com

Received June 10, 2014; **Accepted** August 30, 2014; **Published** September 03, 2014

Citation: Mehmood MA, Sehar U, Ahmad N (2014) Use of Bioinformatics Tools in Different Spheres of Life Sciences. J Data Mining Genomics Proteomics 5: 158. doi:10.4172/2153-0602.1000158

Copyright: © 2014 Mehmood MA, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Tool	Description	References
BLAST	It is a search tool, used for DNA or protein sequence search based on identity.	[109]
HMMER	Homologous protein sequences may be searched from the respective databases using this tool.	[110]
Clustal Omega	Multiple sequence alignments may be performed using this program.	[111]
Sequerome	Used for sequence profiling.	[112]
ProtParam	Used to predict the physico-chemical properties of proteins.	[113]
JIGSAW	To find genes, and to predict the splicing sites in the selected DNA sequences.	[114]
novoSNP	Used to find the single nucleotide variation in the DNA sequence.	[115]
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 the gene	http://bioinformatics. biol.rug.nl/ websoftware/ppp/ppp_start.php
Virtual Foorprint	Whole prokaryotic genome (with one regular pattern) may be analysed using this program along with promoter regions with several regulator patterns.	[116]
WebGeSTer	This is a database containing sequences of transcription terminator sequences and is used to predict the termination sites of the genes during transcription.	[117]
Genscan	Used to predict the exon-intron sites in genomic sequences.	[118]
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.	www.softberry.com

Table 1: Selected tools for primary sequence analyses.

depends on the very nature of analysis to be carried out . For example, data retrieval tools such as *Entrez* of PubMed [9] allows one to search and retrieve data from a wide range of data domains. Similarly, pattern discovery tools such as Expression Profiler [10], Gene Quiz [11] allow researchers to search out different patterns in the given data. Another set of tools is dedicated to carry out sequence comparison. These tools such as BLAST (Basic Local Alignment Search Tool) [12], ClustalW [13] enable one to compare gene or protein sequences to study their evolutionary history or origin. The data visualization tools such as Jalview [14], GeneView [15], TreeView [16], Genes-Graphs [17] allow researchers to view data in graphic representation. These tools use advanced mathematical modelling and statistical inferences such as dynamic programming, Hidden Markov Model (HMM), Regression analysis, Artificial Neural Network (ANN), Clustering and Sequence Mining to analyse the given sequence.

These analyses are popular due to their huge applications in biological sciences, the simplicity, and the capacity to generate a wealth of knowledge about the gene/protein in question. These types of analyses are particularly useful for identification of promoter, terminator, or un translated regions involved in the expression regulations, recognition of a transit peptide, introns, exons or an open reading frame (ORF), and identification of certain variable regions to be used as signatures for diagnostic purposes. Therefore, sequence analyses are one of the frequently performed analyses of bioinformatics. For example, Stoilov et al. [18] used sequence analysis coupled with homology modelling to investigate the genetic basis of primary congenital glaucoma (PCG) [18]. The authors were able to underpin mutations that impair the proper folding and haeme-binding ability of CYP1B1 peptides. Similarly, a genome-wide sequence analysis (GWSA) of *Mycobacterium tuberculosis* H37Rv revealed that majority of the bacterium's proteins were the result of repetitive gene duplication or exon–shuffling events [19]. In a recent study, the gene *cbp*50 from *Bacillus thuringiensis* serovar konkukian was predicted to encode protein that features multiple chitin-binding domains [20,21]. Similarly, Rho-independent transcription terminators form a collection of 343 prokaryotic genomes were predicted quite accurately (<6% false positive prediction) using various computational tools [22].

Mostly predictions rely on complementary DNA (cDNA) and Expressed Sequence Tags (ESTs). However, the cDNA/ESTs information is often scarce and incomplete, and therefore makes the task of finding new genes hugely difficult. Computational scientists have developed another technique referred as an *ab initio* gene-identification. The potential of this technique was demonstrated in a study, which was able to predict 88% of already verified exons and 90% of the coding nucleotides from *Drosophila melanogaster* with very low rate of false-positive identification [23]. Keeping in view the accuracy (~90%) delivered by this approach, it could be a reliable tool for annotating lengthy genomic sequences and prediction of new genes. Recently, Lencz et al. [24] were able to identify an inter-genic Single Nucleotide Polymorphic (SNP; rs11098403) at chromosome 4q26 linked with schizophrenia and bipolar disorder by performing a genome-wide association study (GWAS) coupled with cDNA and RNA Seq on a set of 23,191 individuals (5,415 schizophrenic, 4,785 bipolar and 12,991 controls) [24]. The rs11098403 was found to be linked with the expression of neighbouring enzyme, NDST3, involved in the metabolism of heparan sulphate (HS) in the brain tissues. Similarly, Peng and co-workers (2013) predicted the function of 31,987 genes from the draft genome of a forest species *Phyllostachys heterocycla* using gene prediction modelling approaches based on FgeneSH++ [25]. Please refer to Table 1 for a list of tools used in primary sequence analyses.

Phylogenetic Analyses

Phylogenetic analyses are procedures used to reconstruct the evolutionary relationship among a group of related molecules or organisms, to predict certain features of a molecule with unknown functions, to track gene flow, and to determine genetic relatedness [26]. This all could be represented on a genealogic tree or tree of life. The underlying principle of phylogeny is to group living organisms according to the degree of similarity: greater the similarity, closer the organisms would appear on a tree. A phylogenetic comparative analysis is widely used to control for the lack of statistical independence among species [27].

The methods to construct a phylogenetic tree are divided into three major groups: distance methods, parsimony methods, and likelihood methods. None of the methods is perfect; each one has its own particular strengths and weaknesses. For example, the distance-based trees are easy to set up but not that accurate. The maximum parsimony and maximum likelihood methods are (in theory) the most accurate, but they take more time to run [28]. The distance-matrix methods such as Neighbour Joining (NJ) or Unweighted Pair Group Method with Arithmetic mean (UPGMA) are the simplest. The experts believe that

Tool	Description	Reference
MEGA (Molecular Evolutionary Genetics Analysis)	Builds phylogenetic tress to study the evolutionary closeness.	[119]
MOLPHY	It is molecular phylogenetic analysis tool based on maximum likelihood method.	[120]
PAML	A phylogenetic analysis tool based on maximum likelihood.	[121]
PHYLIP	A package for phylogenetic studies.	[122]
JStree	An open-source library for viewing and editing phylogenetic trees for presentation improvement.	[123]
TreeView	Software to view the phylogenetic trees, with the provision of changing view.	[124]
Jalview	It is an alignment editor and is used to refine the alignment	[125]

Table 2: Some popular tools used for phylogenetic analyses.

the neighbour joining method provides a very good trade-off between the available methods.

Since discussing the details of each bit for performing MSA, building trees, and testing best fits is beyond the scope of this article, therefore, the reader is referred to the detailed protocol published by the Molecular Genetics Laboratory, Central University of Punjab [29] on this issue. Table 2 lists some widely used tools in phylogenetic analyses.

Phylogenetic tools are commonly used to test various evolutionary hypotheses and have become indispensable for functional genomics, particularly when the functions of a gene are not known. For example, prior to the expression of an algal membrane protein, plastid terminal oxidase 1 (PTOX1), in tobacco chloroplasts, authors conducted a phylogenetic analysis to construct the evolutionary history and determine essential features of that particular polypeptide [30]. The phylogenetic analysis revealed that the *Chlamydomonas reinhardtii* PTOX1 (Cr-PTOX1) has typical signatures of higher plant PTOX such as iron-binding sites, a conserved exon and various blocks of amino acids to act as plastoquinol terminal oxidase [30]. Similarly, Chen et al. [31] used phylogenetic analysis to study the evolutionary history of respiratory mechanisms in the deep-sea bacterium *Shewanella piezotolerans* WP3 [31]. The phylogenetic analyses coupled with reverse genetic studies revealed that out of two nitrate reductases, NAP-α and NAP-β, the hallmark of the genus *Shewanella*, the NAP-β evolved long before NAP-α molecules.

Sequence Databases

Biological sequence database refers to a vast collection of information about biological molecules such as nucleic acids, proteins and polymers, each molecule to be identified by a unique key. The stored information is not only important for future use but also serves as a tool for primary sequence analyses. With the advancement of high throughput sequencing techniques, the sequencing has reached to a whole-genome scale, which is generating a massive amount of data every day. The submission and storage of this information to become freely available to the scientific community has led to the development of various databases worldwide. Each database has become an autonomous representation of a molecular unit of life. This section deals with such databases, as an understanding of these databases will help to retrieve important information from these data collections relevant to one's project.

Databases contain a variety of information; and therefore are classified into *Primary*, *Secondary*, or *Composite* databases, depending upon the information stored in them. For example, the data in a primary database is obtained through experimentation such as yeast-two hybrid assay, affinity chromatography, XRD or NMR approaches such as related to sequence or structure. SWISS-PROT [32], UniProt [33] and PIR [34], GenBank [35], EMBL [36], DDBJ [37] and the Protein Databank PDB [38] are examples of primary databases. A secondary database contains information that is derived from the analysis of data stored in primary databases like conserved sequences, active sites of a protein family or conserved secondary motifs of protein molecules [39,40]. Examples of secondary databases include SCOP [41], CATH [42], PROSITE [43] eMOTIF [44]. Consequently, the primary databases are of archival nature while secondary databases are termed as curated databases. A composite database contains information derived from different primary sources. Examples of composite databases include NRDB (non-redundant database), which contains data obtained from GenBank (CDS translations), PDB, SWISS-PROT, PIR, and PRF. Similarly, the INSD (International Nucleotide Sequence Database) is another example of composite database, which is collection of nucleic acid sequences from EMBL, GenBank, and DDBJ. The UniProt (universal protein sequence database) [45] represents another example, which is also a collation of sequences derived from various other databases PIR-PSD, Swiss-Prot, and TrEMBL. Similarly, wwPDB (worldwide PDB) is a composite of 3D structures in the RCSB (Research Collaboratory for Structural Bioinformatics), PDB, MSD, and PDBj [46].

Genome Sequence Databases

The GenBank, built by the NCBI [35], is a vast collection of genome sequences of over 250,000 species. The data from GenBank can be accessed through the NCBI's integrated retrieval system, *Entrez*, while the literature is accessible via PubMed [47]. Each sequence carries information about the literature, bibliography, organism, and a set of various other features, which include coding regions, promoters, untranslated regions, terminators, exons, introns, repeat regions, and translations. The sequence information stored in GenBank is obtained through submission both by the individual laboratories as well as by large-scale genome sequencing projects. Similarly, the Xenbase is an updated resource of genomic and biological data on the frogs including *Xenopus laevis* and *Xenopus tropicalis* [48], where *Xenopus* spp. are considered as model providing new knowledge in the field of developmental biology which may exploited to modelling and simulation studies of the human diseases.

The *Saccharomyces* Genome Database (SGD) contains comprehensive information of the yeast (*Saccharomyces cerevisiae*) and also provides bioinformatics tools to explore and analyse the data available in SGD. The SGD may be used to study functional relationships among gene sequence and gene products in other fungi and eukaryotes (http://www.yeastgenome.org/). Similarly, another genome database called "WormBase" is developed and maintained by international consortium of computer scientists and molecular biologists to provide precise, recent and reachable information related to the molecular biology of *C. elegans* and other related nematodes (http://www.wormbase.org). The webpage for this database also host several tools for the precise analyses of the stored information. Another up-to-date database is "FlyBase" dedicated to provide information on the genes and genomes of *Drosophila melanogaster* along with the tools to search gene sequences, alleles, genetic aberrations, different phenotypes, and images of the *Drosophila* species [49]. Similarly, the wFleaBase (http://

Database	Description	Reference
	Nucleotide Databases	
DNA Data Bank of Japan	It is the member of International Nucleotide Sequence Databases (INSD) and is one of the biggest resources for nucleotide sequences.	[37]
European Nucleotide Archive	It captures and presents information relating to experimental workflows that are based around nucleotide sequencing.	[126]
GenBank	It is the member of International Nucleotide Sequence Databases (INSD) and is a nucleotide sequence resource.	[47]
Rfam	A collection of RNA families, represented by multiple sequence alignments	[54]
	Protein Databases	
Uniprot	One of the largest collection of protein sequences.	[45]
Protein Data Bank	This is another major resource of proteins containing information of experimentally-determined structures of nucleic acids, proteins, and other complex assemblies.	[38]
Prosite	Provides information on protein families, conserved domains and actives sites of the proteins.	[43]
Pfam	Collection of protein families	[39]
SWISS PROT	A section of the UniProt Knowledgebase containing the manually annotated protein sequences	[32].
InterPro	Describes the protein families, conserved domains and actives sites	[127]
Proteomics Identifications Database	A public source, containing supporting evidence for functional characterization and post-translation modification of proteins and peptides.	[128]
	Genome databases	
Ensembl	It is a database containing annotated genomes of eukaryotes including human, mouse and other vertebrates.	[129]
PIR	An integrated public resource to support genomic and proteomic research	[34]
	Miscellaneous Databases	
Medherb	Resource database for medicinally important herbs	[130]
Reactome	A peer-reviewed resource of human biological processes	[131]
TextPresso	This database provides full text literature searches of model organism research, 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/
TAIR	The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular data for the model plant *Arabidopsis thaliana*. It provides information on gene structure, gene product, gene expression, DNA and seed stocks, genome maps, genetic and physical markers.	http://www.arabidopsis.org/
dictyBase	dictyBase is an online bioinformatics database for *Dictyostelium discoideum*.	[132]
	Signalling & Metabolic Pathway Databases	
KEGG	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.	[133]
CMAP	Complement Map Database is a novel and easily accessible research tool to assist the complement community and scientists from related disciplines in exploring the complement network and discovering new connections.	[134]
SGMP	The Signaling Gateway Molecule Pages (SGMP) database provides highly structured data on proteins which exist in different functional states participating in signal transduction pathways.	[135]
PID	The Pathway Interaction Database (PID) is a collection of curated and peer-reviewed pathways composed of human molecular signaling and regulatory events and key cellular processes. It serves as a research to study the cellular pathways with a special emphasis on cancer.	[136]
HMDB	The Human Metabolome Database (HMDB) is the most comprehensive curated collection of human metabolite and human metabolism data in the world. It contains records for more than 2180 endogenous metabolites with information gathered from thousands of books, journal articles and electronic databases along an extensive collection of experimental metabolite concentration data compiled from hundreds of mass spectra (MS) and Nuclear Magnetic resonance (NMR) from the analyses performed on urine, blood and cerebrospinal fluid samples. The HMDB is designed to address the broad needs of biochemists, clinical chemists, physicians, medical geneticists, nutritionists and members of the metabolomics community.	[137]

Table 3: List of some popular databases.

wfleabase.org/) provides information on genes and genomes for species of the genus *Daphnia* (water flea) where *Daphnia* is considered as a model system to study and understand the complex interplay between genome structure, gene expression, individual fitness, and population-level responses to chemical contaminants and environmental change. Although, the wFleaBase contains data from all species of the genus yet the primary species are *D. pulex* and *D. magna.* Please refer to Table 3 for further information on genome databases.

Protein Sequence Databases

The most significant protein sequence databases include SWISS-PROT (Swiss Protein) Databank [50], TrEMBL (translation of DNA sequences in EMBL) [32], UniProt (Universal Protein Resource) [33], PIR (Protein Information Resource) [34] and wwPDB (worldwide Protein DataBank). The SWISS-PROT [32] represents one of the comprehensive protein sequence databases. The SWISS-PROT provides information of its entries, which has been generated both experimental as well computational studies. It also provides links to several other data sources such as GenBank, EMBL, DDBJ, PDB and various other secondary protein databases namely domains, post-translational modifications, species-specific data collections. The protein information in SWISS-PROT mainly concentrates on model organisms and human. The TrEMBL by contrast provides information on proteins from all organisms [32].

Similarly, the PIR is another comprehensive collection of protein sequences. It provides user several attractive features for example to search for a protein molecule via an 'interactive text search' and to perform various web-based analyses such as sequence alignment, matching of peptide molecules and peptide mass calculations [51].

The UniProt is one of the comprehensive collections of protein sequence resources, which are open to free access. The UniProt database

emerged by combining SWISS-PROT, PIR and TrEMBL collections. It provides all sorts of protein information ranging from sequence to function [52]

The worldwide Protein Data Bank (wwPDB) has been exclusively designed to archive each single 3D structure of protein molecules to become freely available to the scientific community. The databank now contains over 83,000 experimentally generated structures. The PDB also constantly develops tools for the users to provide better access to the data [53].

Miscellaneous Databases

The Rfam database contains comprehensive information about RNA molecules and their various features like secondary structures and gene expression modulating elements. The Rfam databases are hosted by the Wellcome Trust Sanger Institute and it is similar to the Pfam database for annotating protein families [54]. As there are number of curated databases available, one of such databases is IntAct, which contains data on protein interactions. All data manually curated by MINT (Molecular INTeraction database) curators has been shifted onto the IntAct database at EMBL-EBI and have been merged with the existing IntAct data collections [55].

MINT is another database that stores information about protein-protein interactions derived from already published data in literature [56]. Curated databases for information on complex metabolic pathways have also been built. For example, the Reactome is one such curated database that represents a range of diverse human processes ranging from metabolism to signal transduction. The Reactome is an open source platform, which is freely available to be used and re-distributed [57].

The Transporters Classification Database (TCDB) is a collection of membrane transporters [58]. It uses an internationally approved Transport Classification (TC) system for the classification of protein, which is similar to that of Enzyme Commission (EC). However, it also has some differences from EC system; it provides functional and phylogenetic information as well, for example. The information of more than 600 families of transporters is available in this database. A TC number to sequenced homologues of unknown function is assigned only if it belongs to rare or under-represented family. Various subunits are represented by 'S' followed by a number such as S1, S2, S3 and so on. Whereas the proteins which act as accessory transporters as well as those whose characterization is not complete yet are represented by number 8 and 9, respectively [59]. Similarly, the Carbohydrate-Active enzyme Database (CAZy) contains comprehensive information about carbohydrate-modifying enzymes and other information relevant to them. The enzymes are classified into distinct families on the basis of amino acid similarities in their sequences or the presence of various catalytic domains [60]. The databases about the structure, classification and ontology of the lipid molecules have been discussed in detail elsewhere [61]. Since, comprehending all the databases is beyond the scope of this article, we have listed some popular databases in Table 3.

Predicting Protein Structure and Function

Protein molecules begin their life as shapeless amino acid strings, which ultimately fold up into a three-dimensional (3D) structure to become biologically active. The folding of the protein into a correct topology is a pre-requisite for any protein to perform its biological functions. Therefore, information of 3D structure of a protein is necessary to gain an insight into the function of a specific protein. Usually, 3D structures are determined by X-ray crystallography or related techniques like NMR. However, these techniques are expensive, difficult and time consuming and are often hampered by the poor heterologous expression, and attempts to obtain good crystals [62]. Therefore, very few structures (~250) using XRD and NMR spectroscopy are submitted compared to nearly a million monthly submissions to NCBI. Information of tertiary structures on genome-scale level for many proteins is therefore lacking. Alternatively, a protein's 3D structure can be predicted using various bioinformatics tools, and consequently has become one of the hot topics in the field of bioinformatics [62].

Bioinformatics approaches can easily identify secondary structure elements in a protein sequence such as helices, sheets, domains, strands and coils. Proteins adopt a specific structure due to the presence of weaker electrostatic forces such as hydrogen bonds between these elements. Therefore, the propensity of appearing certain residues in a particular region of protein such as sheets or coils can be useful to predict a secondary structure of a protein. The most straightforward approach to predict a 3D structure of a protein molecule is comparative modelling. In this approach, a related template (at least 30% sequence identity with target protein) is selected to predict the unknown structure. Since, the 3D structural information is scant, and it is not possible to infer tertiary structures from similarity searches. Therefore, targets not having sufficient identity can be modelled by using different approaches such as known as threading or fold-recognition [63]. In case where these tools fail to generate a reliable structure, then a combination of various physical principles is applied.

The most commonly used method is homology modelling for predicting template-based structure of target protein. However, relatively low number of structures available in PDB hampers this approach [64]. A variety of procedures is available to model the target protein if a homologue of an unknown protein is available such as COMPOSER [65] or 3D –JIGSAW [66] and MODELLER [67] to name a few.

The majority of secondary structure prediction tools use the frequency of observed amino acids at a certain position, which is guessed from the 3D structures determined experimentally. Therefore, earliest methods used observed periodicity of residues to predict secondary structures. However, advanced methods to predict secondary structures or identify secondary structure elements in a given protein sequence use neural networks such as NNPredict. The NNPredict is a multilayer neural network-based method, which predicts the position of each amino acid by letters 'H' or 'E' for residues appearing in helices or coils respectively. Similarly, the PredictProtein is another automated server, which is based on neural network. It uses multiple sequence alignment to predict various structural and functional annotations of a protein molecule. JPred is another neural-network based method that uses a combination of various methods to predict secondary structure. Since it uses different methods to predict a structure, therefore, its predictions are usually of higher accuracy. To have a snapshot of various protein prediction tools, visit this page http://www.biologie.uni-hamburg.de/b-online/library/genomeweb/GenomeWeb/prot-2-struct.html.

For predicting structures, computer simulations include energy calculations based on physio-chemical principles, thermodynamic equilibrium with a minimum free energy and global minimum free energy of protein surface. A number of tools are available to predict the secondary structure of a protein molecule. One of the most important tools is ExPASy (the Expert Protein Analysis System), powered by the Swiss Institute of Bioinformatics (SIB). The Expasy provides access to a number of web-based sources such as SWISS-PROT, TrEMBL,

Tool	Description	References
CATH	A semi-automatic tool for the categorized organization of proteins.	[138]
RaptorX	It facilitates the user to predict protein structure based on either a single- or multi-template threading.	[139]
JPRED	Used to predict secondary structures of proteins.	[140]
PHD	Used to predict neural network structure.	[141]
HMMSTR	A hidden Markov model for the prediction of sequence-structure correlations in proteins.	[142]
APSSP2	Predicts the secondary structure of proteins.	[143]
MODELLER	Predicts 3D structure of protein based on comparative modelling	[70]
Phyre and Phyre2	Web-based servers for protein structure prediction	[144]

Table 4: Selected tools used to perform structure-function analyses of proteins.

SWISS-2DPAGE, PROSITE, ENZYME and the SWISS-MODEL to perform a protein's structure- as well as function-related studies. The Expasy also provides several additional tools to determine similarity, pattern identification, and studying posttranslational modifications [68]. The iterative threading assembly refinement (I-TASSER) is a web-based tool, which generates automated protein structure and makes functional predictions. The server generates 3D models of a target protein via multiple threading using templates from PDB [69].

These approaches have been successfully applied to predict the structure of chitin-binding proteins CBP50 and CBP24 from *B. thuringiensis* serovar konkukian S4 using Modeller v9.0 [70] and Auto Dock vina [21,71,72]. Another study explored the pathogenicity of *Mycoplasma genitalium* strain G37 in sexually transmitted diseases by modelling the hypothetical proteins of the selected strain using (PS)2v2 sever [73]. Similarly, a putative gene (deactylase or xylanase) *cda*1 was subjected to functional annotation, and it was confirmed that the enzyme encoded by *cda*1 gene is a chitin deacetylase gene and may not have any xylanase activity [74]. Table 4 lists some commonly used tools to predict secondary structure of protein molecules.

Molecular Interactions

Proteins seldom perform their functions in isolation, and therefore often interact with other molecules all the time to execute a certain process. Understanding how biomolecules interact with other molecules holds numerous implications, for example, for protein folding, drug design and purification techniques [75] and therefore has become one of the mostly pursued research area using either experimental or bioinformatics approaches. Understanding of molecular interactions is also essential to elucidate the biological functions of a molecule. For example, protein-protein interactions play a key role in cellular activities such as signalling, transportation, homeostasis, cellular metabolism and various biochemical processes [76].

Bioinformatics in this regard becomes quite handy to predict protein-protein interactions without resorting to costly, and time-consuming physical approaches such as X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. Often crystal structure coordinates give misleadingly static views of interactions as a complex cannot be represented by a single structure. Therefore, it has been realized that 3D structure of a molecule cannot produce a complete picture of each and every individual interaction. Therefore, computational approaches capable to predict reliable protein-protein interactions have become essential. Nevertheless, such studies generate useful information, which enable scientists to determine a specific pathway to be manipulated in order to achieve required change(s) in the cell.

The parameters governing protein-protein interactions include interface size, amino acid composition at interface, types of chemical groups, complementarity between surfaces, hydrophobicity, hydrogen bonds, and conformational changes whilst complex formation takes place. These properties are studied using various protein datasets. The *in-silico* approaches used to study molecular interactions fall into two groups: homology based, and non-homology based. The homology-based methods as the name implies work on direct comparison of protein sequences. The non-homology approaches take into account functional interactions collectively. Although the homology-based methods remain the most preferred methodology, the non-homology-based methods are powerful to assign functions to those genes whose homologues have not been characterized yet. For example, Lu et al. [77] developed 2,865 protein-protein interactions using a multimeric threading approach [77] and out of these 1138 were confirmed in the DIP (Database of Interacting Proteins) [78]. Recently, Hosur et al. [79] developed a new three-step algorithm, Coev2Net, to predict protein-protein interactions. The algorithm is capable to predict interactions with a high-level of performance as compared to prevalent methods. Similarly, Zhang et al. [80] used PrePPI algorithm to predict a large number of reliable interactions from both yeast (30,000) and human (>300,000) [80]. However, all methods have their own limitations. For instance, they make use of fewer examples, which cannot be applied to all species or all proteins. In an attempt to develop a universal methods, Valente and co-workers [81] designed a method called Universal *In silico* Predictor of Protein-Protein Interactions (UNISPPI), which can be applied to a range of diverse species, hence it was termed 'universal'. Other useful features of the model include its capability to differentiate instances of a complete proteome or even parasite-host associations.

Apart from prediction of protein structures, the molecular modelling can also assist in choosing one unique conformations, which governs the activity of a biomolecule. Other applications include spotting residues at 'hot-spot' of protein interfaces by docking a protein onto a small molecule called ligand. There are a large number of softwares available to perform docking calculations; only few, which are most widely used, will be discussed here.

The best-known docking program is DOCK, which is able to assign ligand site on the receptor quite reliably. It also performs evaluation of the quality of the fit. Another program GRID [82] uses a 3D grid to find out protein binding sites for ligands. AUTODOCK is another commonly used suit and is perhaps one of the most cited platform for the prediction of protein-ligand docking studies (http://autodock.scripps.edu/). It is run and maintained jointly by The Scripps Research Institute (TSRI) and Olson Laboratory. The High Ambiguity Driven protein-protein Docking (HADDOCK) (http://haddock.science.uu.nl/) is another docking approach for the modelling of bio-molecular complexes [83]. For a comprehensive comparison of various docking programs, reader is referred to the review article published by [84]. Similarly, an algorithm, IsoRank, was developed to perform the global alignment of multiple protein-protein interaction networks to maximize the overall match across all input networks. The IsoRank was used to compute the first known global alignment of PPI networks using five

Tool	Description	Reference
SMART	A Simple Modular Architecture Retrieval Tool; describes multiple information about the protein query.	[145]
AutoDock	Predicts protein-ligand interaction and is considered as reliable tool.	[146]
HADDOCK	Describes the modelling and interaction of bio-molecular complexes such as protein-protein, protein-DNA	[83]
BIND	A database that provides access to molecular interaction and bio-complexes	[89]
MOE	An integrated package of tools used for drug discovery. It combines visualization, modelling, and drug discovery on one plate-form.	[147]
STRING	A database of both known and predicted protein interactions.	[148]
MIMO	A dynamics graph-matching tool for the comparison of biological pathways in an efficient manner.	[149]
IntAct	It is an open source database system and provides analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available.	[55]
Graemlin	It is capable of scalable multiple network alignment with its functional evolution model that allows both the generalization of existing alignment scoring schemes and the location of conserved network topologies other than protein complexes and metabolic pathways.	[150]
PathBLAST	It is meant to search protein-protein interaction network of the any selected organism and extracts all interaction pathways that align with the query.	[151]
CFinder	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. It is efficient for clustering data represented by genetic or social networks and microarray.	[152]
MCODE	It is suited for both computationally and biologically oriented researchers. Its features include; Fast network clustering, Fine-tuning of results with numerous node-scoring and cluster-finding parameters, Interactive cluster boundary and content exploration, Multiple result set management, Cluster sub-network creation and plain text export	[153]

Table 5: Selected tools to study the molecular interactions.

species: yeast, fly, worm, mouse and human. It was revealed functional orthologs across these species [85]. Later, IsoRankN (IsoRank-Nibble) was developed based on spectral methods and is error-tolerant and computationally efficient [86].

Computer based techniques could assist and accelerate the discovery of biological mechanism and lead molecules needed for new drug. For example, virtual screening of flavonoids from *Amelanchier alnifolia* against Hepatitis C virus (HCV)'s non-structural NS_3/4A protease /helicase discovered a high binding affinity of Quercitin 3- galactoside and 3-glucoside with HCV/NS_3/4A [87]. The study suggested that Quercitin 3- galactoside and 3-glucoside might be good candidates for inhibition of HCV NS_3/4A. Similarly, Sehar and co-workers predicted a chitin-binding site in CBP24 [72] and CBP50 [21] of *Bacillus thuringiensis* using molecular modelling to study chitin-degradation pathways to be used in engineering fungal resistance mechanisms in plants.

The massive generation of data has led to the development of various databases to organize and facilitate study on molecular interactions. For example, signal transduction pathways databases may include protein-protein, protein-DNA, Protein-RNA, DNA-RNA, DNA-substrate interactions [88]. The Biomolecular Interaction Network Database (BIND) is one of the largest available information resources that provide access to pairwise molecular interaction and complexes [89]. Similarly, MINT is another database, which stores information of functional interactions of biological molecules [90]. A list of selected tools to study protein-protein interactions is given in Table 5.

Drug Designing

Drug discovery is a process by which new drug molecules are discovered or designed to cure different diseases. Before the advent bioinformatics tools, scientists used chemistry, pharmacology and clinical sciences to discover new compounds. However, the traditional process is quite slow and expensive as well. The market pressure to find new drugs in a short period with minimum risks has fuelled the interest in alternative ways of designing drugs such as bioinformatics. Bioinformatics has greatly facilitated this complex process and is playing a vital role in advancing the process of drug discovery/designing, since it is faster to analyse molecules on computer as compared to experimental approaches. In fact, a completely new and dedicated field known as Computer Aided Drug Design (CADD) has come into existence to discover novel drug molecules [91]. The whole process of discovering and designing new drug molecules is quite complicated and is quite challenging. The entire process can be divided into four different steps: identification of drug target, validation of target, lead identification, and lead optimization [92]. In this section, we will briefly discuss how bioinformatics is useful in discovering new drugs.

Since drug molecules always act on a target to deliver therapeutic benefit to the patient. The *target* is a small key biomolecule that allows the drug molecule to produce a desired effect on metabolic or signalling pathway pertinent to the disease under study without interfering the normal functioning of the cell. Therefore, the very first step in the drug designing process is to identify a target involved in that disease. This demands a full knowledge of metabolic processes in normal as well as diseased conditions. The sequencing of human genome provided over 30,000 genes to researchers to include them in their search for new drug targets [93]. Since then the number of potential drug targets is increasing day-by-day [92]. Understanding how a gene functions is indeed a key to choose a gene as a target. A number of databases have been developed to facilitate the search of new drug targets (Table 6).

After selecting potential targets, the involvement of those targets in a particular disease is studied. This is target validation. The targets are compared to analyse for their ability to influence that disease. This is also necessary to determine the likelihood of success in next phase. Bioinformatics approaches such as modelling enable scientists to tailor compounds to bind at a particular site (Predicting protein structure and function for detail on modelling). Next scientists have to find a certain compound - lead compound - capable to alter the action of target. A number of bioinformatics tools allow virtual screening of a large number of compounds that could bind/inhibit or activate a protein. The virtual High Throughput Screening (vHTS) enables identification of promising molecules as early as possible; one of the most needed process in the entire drug discovery process. Often the identified compounds do not have required properties, and therefore they are

Database	Description	Reference/URL
Potential Drug Target Database (PDTD)	It is a dual function, wide-range database of drug targets that is globally accessible via internet, containing 1207 entries including 842 with known structures.	[154]
Drug Bank	It is an exclusive resource that interconnects the drug-target data. It contains 7681 drug entries including 1545 FDA-approved, 155 FDA-approved biotech-based drugs, 89 nutraceuticals and over 6000 experimental drugs along with 4218 non-redundant protein sequences which are associated to these drug entries.	[155]
Therapeutic Target Database (TTD)	It is a collection of known and explored therapeutic proteins and DNA targets, the disease, pathways involved in the disease and the corresponding drugs directed at each of these targets. It also provides links to relevant databases about target function.	[156]
TDR Target Database	It is a database as well as a tool. It is meant to identify and prioritize the genes of interest from the ignored disease pathogens by running simple queries, assigning them numerical values and combining the output to produce a ranked list of candidate targets. The TDR here is abbreviated for Tropical Disease Research which is a special program within the World Health Organization (WHO) agenda.	[157]
MATADOR: Manually Annotated Targets and Drugs Online Resource	MATADOR is a resource for protein-chemical interactions. In contrast to DrugBank which usually contains only the main mode of interaction, the MATADOR provides manually annotated list of direct (binding) as well as protein-chemical interaction. Each interaction is linked to PubMed or OMIM entries that were used to deduce the interaction. The user can choose either to trust only the direct interactions (with a known mechanism) or also indirect interactions.	[158]
TB Drug Target Database	This is a specialized database contains information on drugs and target proteins for the treatment of tuberculosis (TB) only including the structural details of inhibitors.	http://www.bioinformatics.org/tbdtdb/
DrugPort	It provides the structural information available in the Protein Data Bank (PDB) related to drug molecules and their targets based on the latest version of DrugBank database.	http://www.ebi.ac.uk/thornton-srv/databases/drugport/
ChEMBL	It is a collection of drug-like bioactive molecules, along with their 2-D structures, calculated and abstracted properties such as; logP, molecular mass, Lipinski Parameters, binding constants, pharmaco-kinetics etc.	[159]

Table 6: Some popular drug target databases.

Tool	Description	Reference/URL
Abalone	Abalone is a general purpose molecular modeling program which is meant for biomolecular dynamics simulations of proteins, DNA, ligands. It has by-default ability to interact with external quantum programs NWChem, CP2K and PC GAMESS/Firefly).	http://www.biomolecular-modeling.com/Abalone/index.html
Ascalaph	Similar to Abalone, Ascalaph is also a general purpose molecular modeling tool to perform quantum mechanics calculations for model development, molecular mechanics and dynamics simulations of DNA, proteins and hydrocarbons, either in the gas or in condensed phase. It has a built-in ability to interact with external molecular modeling packages such as, (MDynaMix, ORCA, NWChem, CP2K, PC GAMESS/Firefly and DelPhi).	http://www.biomolecular-modeling.com/Products.html
Discovery Studio	Discovery Studio is a comprehensive modeling and simulation package focused on optimizing the drug discovery process including the capabilities of small molecule simulations, pharmacophore modelling, protein-ligand docking, protein homology modelling, sequence analyses, protein-protein docking and antibody modelling, etc.	http://accelrys.com/products/discovery-studio/
Amber	Amber is the collection of programs that facilitate users to perform molecular dynamics simulations with an emphasis on biomolecules.	[160]
FoldX	FoldX provides quick and quantitative estimation of molecular interactions which are contributing towards the stability of either single protein or protein complexes.	http://foldx.crg.es/

Table 7: Molecular dynamics simulation tool.

'refined' to produce more specific effect with reduced number of side effects. This process is 'lead optimization'.

A number of computer techniques are capable enough to give a compound a higher specificity and fewer side effects. Since lead optimization is the most expensive step in the entire drug discovery process, therefore, often scientists have to develop chemical analogues of such compounds with desired properties. Following the identification and refinement of the lead molecule, scientists conduct pre-clinical animal safety tests. If the lead molecules do not have required binding properties, the drug discovery project is likely to fail [5].

One of the challenges for researchers for developing a new drug is the prediction of drug-like properties of the lead compound. These properties include charge distribution, solubility, hydrophobicity, pKa, refractivity, molecular weight, and ClogP/LogD. Therefore, the initial evaluation with the help of bioinformatics techniques can significantly influence the ultimate success of the project [94].

Many tools are available to predict drug-like properties and ADMET (absorption, distribution, metabolism, excretion, and toxicity). For example, the OSIRIS Property Explorer is a web-based tool, which predicts a number of ADMET properties including cLogP, solubility, Toxicity, and Overall Drug-Score. The values are displayed in different colours like green for good, red for bad and yellow for irrelevant. Although it is very user-friendly platform, however, independent assessment is needed of its quality. Similarly, ChemSilico is neural net based prediction method, which calculates ADMET properties. This tool has been validated and trained using over 35,000 tested compounds. The Pre-ADME is another freely available web-based utility that predicts ADMET properties of a druggable compound. Similarly PASS Online (Prediction of Activity Spectra for Substances) is a Bayesian-based tool that calculates over 4000 different biological activities including pharmacological effects, mutagenicity, mode of action, toxicity, interaction with metabolic enzymes and transporters, influence on gene expression, and embryotoxicity. PASS is able to predict with a validation of 85%. It is freely available, but the user has to register first. DREADD (designer receptors exclusively activated by designer drugs) is a recent addition in the toolkit of a computational biologist for the identification of druggable targets for both known and orphan GPCRs (G-protein coupled receptors). It also allows one

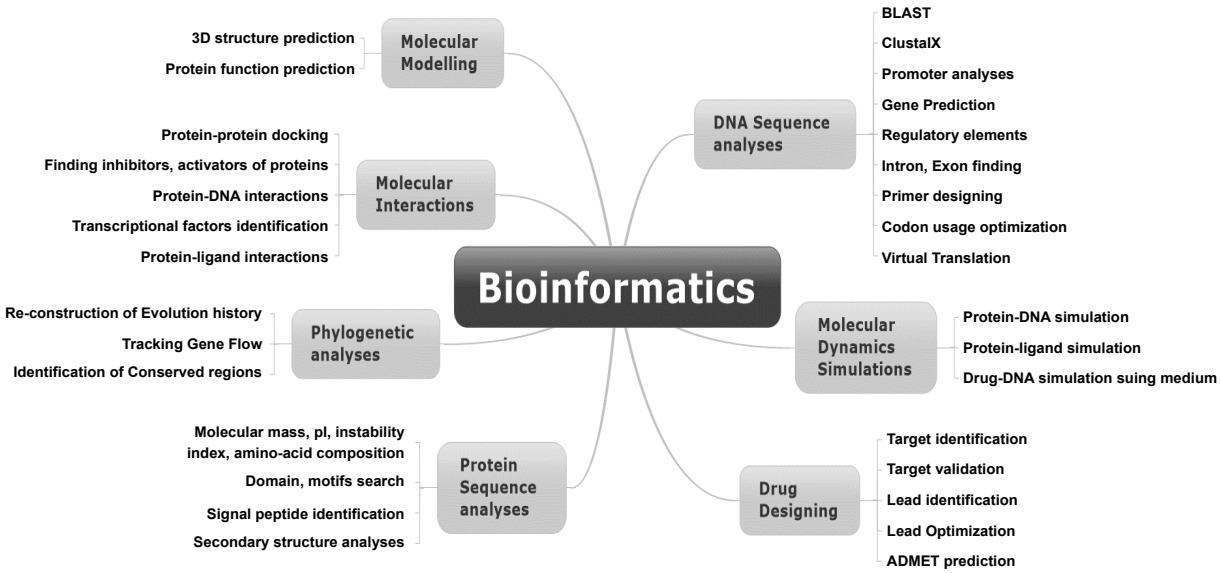

Figure 1: Application of bioinformatics tools in various areas of biological sciences.

to study the activities of novel drugs as well. For more on DREADDs, please refer to the comprehensive review published by [95].

Due to space limitations, it is not possible to include every ADMET prediction tool here. For more details on how drug metabolism is studied using various bioinformatics tools, kindly refer to [96]. Table 7 gives list of various ADMET prediction tools.

It is interesting to note that structure-based predictions are considered more accurate compared to empirically derived predictions due to the detailed understanding of the structure of the substrates and active sites. The generation of more and more structural data is now making it possible to design tools generating mechanistic prediction for ADMET. Even when a structure of a protein is not available, various bioinformatics tools such as modelling allow predicting the structure of an unknown protein. However, having a 3D structure of a protein is not enough to carry out ADMET work. To comprehend mechanistic details of an enzyme, understanding the enzyme-substrate interaction is also essential. This is carried out using another set of computation tools collectively termed as docking. These tools allow one to have study active sites at a molecular level by dock a protein onto a substrate molecule.

Each New Chemical Entity (NCE) should have acceptable ADMET properties to pass through the clinical trials. ADMET data is necessary to determine the feasibility and safety of the drug in human. ADMET deficiencies are the leading cause of failure of most of the drug candidates. The Swiss Institute of Bioinformatics (SIB) has developed an interactive directory of different *in silico* tools used at each stage of drug discovery. The directory is accessible from this URL http://www.click2drug.org/citations.html

There are a number of drugs whose development was assisted by structure-based design and screening strategies. The discovery of the HIV protease inhibitors is one of the examples [97]. Similarly, Reddy et al. [98] used computational tools to develop cyclooxygenase (COXs)-based anti-inflammatory drug with no gastric side effects [98]. In order to identify various mutant forms of H-Ras (Harvey-Ras) polypeptides in cancer patients, Jayakanthan and co-workers performed virtual screening of lead compounds. The authors were able to identify two novel leads, 3-aminopropanesulphonic acid and hydroxyurea. The docking analysis revealed that Ile-36, Glu-37, Asp-38 and Ser-39 were involved in the interaction with the ligand [99]. In a study aimed at finding novel targets for globlastoma, use of bioinformatics tools led to the discovery of several novel genes related to the disease [100]. The study was also be to discover a regulatory feedback loop mediated by cyclin-dependent kinase 1 (CDK1) and WEE1. Similarly, Wu et al. [101] developed tumour-specific networks to identify targets from differentially expressed tumour genes for breast, colon and lung cancer [101]. By using this approach, authors were able to identify several new targets for cancer of which two, Calcium/calmodulin-dependent serine protein kinase (CASK) and RuvB-like1, have recently been verified by experimental approaches [102]. In another study, McDermott and co-workers [103] applied protein-protein interaction approach to newly discovered differentially expressed proteins from a cell culture model of HCV (Hepatitis C Virus) to identify novel targets [103] (Figure 1).

Molecular dynamic simulations

As we know, biological activities are the result of molecular interactions that occur in a time-dependent manner. This time dependent behaviour of a molecule could be studied using another set of bioinformatics tools collectively referred as Molecular Dynamics Simulations (MDS). The MDS techniques aim to provide detailed information on the fluctuations, dynamic processes such as ion transport, and small- and large-scale conformational changes of proteins, nucleic acids, and their complexes occurring in biological systems. They also assist determination of structures from experimental approaches like XRD and NMR spectroscopy. The MDS tools could also be useful in gaining insights into situations where use of experimental means is not possible.

For example, to determine which serine residues take part in phosphorylation of starch branching enzyme IIb (SBEIIb) authors used MDS together with site-directed mutagenesis and mass spectroscopy. The study was able to determine that phospho-Ser297 forms a stable salt bridge with Arg665, part of a conserved Cys-containing domain in plant branching enzymes. This study hold numerous implications

for elucidating biological role of the enzyme in starch biosynthesis in higher plants from yield-improvement perspectives [104].

In another study, MDS were used to study the selectivity of two membrane bound transporter proteins aquaporin-1 (AQP1) and aquaglyceroporins (glycerol facilitator; GlpF) for various solutes like ammonia, urea, water and glycerol. The study observed that unlike GlpF, the selectivity of AQP1 was dependent on the hydrophobicity of the solute particles and therefore could act as a filter for *in vivo* filtering of small molecules [105].

In a recent study, MDS were employed to investigate the folding process of the Trp-cage protein molecules. The Trp-cage protein is a small (~20 amino acids) protein, which enters into a stable state after folding and forms a hydrophobic core around a central Trp residue. Several experimental and simulation approaches failed to understand the underlying mechanism of folding. Authors used a series of advanced simulation techniques to discover that the central Trp6 residue is critical for folding process. The single chain interacts with itself and becomes a barrier for controlling transitions to a near native folded structure [106].

Similarly, Isin et al. [107] used MDS to study different conformations of various ligands with β_2-adrenergic receptor and discovered a novel binding site is involved in binding with high molecular weight molecules [107]. The study suggested that modelling different active conformations for identification of novel binding sites could be used in refining mechanisms of action of various drugs [108]. Table 7 lists some popular platforms widely used for MDS analyses.

Conclusion and Future Prospects

Bioinformatics is a comparatively young discipline and has progressed very fast in the last few years. It has made it possible to test our hypotheses virtually and therefore allows to take a better and an informed decision before launching costly experimentations. Although, more and more tools for analysing genomes, proteomes, predicting structures, rational drug designing and molecular simulations are being developed; none of them is 'perfect'. Therefore, the hunt for finding a better package for solving the given problems will continue. One thing is clear that the future research will be guided largely by the availability of databases, which could be either generic or specific. It can also be safely assumed, based on the developments in the field of bioinformatics, that the bioinformatics tools and software packages would be able to give results that are more accurate and thus more reliable interpretations. Prospects in the field of bioinformatics include its future contribution to functional understanding of the human genome, leading to enhanced discovery of drug targets and individualised therapy. Thus, bioinformatics and other scientific disciplines have to move hand in hand to flourish for the welfare of humanity.

Acknowledgment

Authors thank Higher Education Commission (HEC), Pakistan, for funding their research work and apologise to all those colleagues whose work could not be discussed here due to space constraints.

References

1. Mount DW (2004) Sequence and genome analysis. New York: Cold Spring.
2. Hesper B, Hogeweg P (1970) Bioinformatica:een werkconcept. Kameleon 1: 28-9.
3. Hogeweg P (2011) The roots of bioinformatics in theoretical biology. PLoS Comput Biol 7: e1002021.
4. Peitsch MC (1996) ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. Biochem Soc Trans 24: 274-279.
5. Dibyajyoti S, Bin ET, Swati P (2013) Bioinformatics: The effects on the cost of drug discovery. Galle Med J 18: 44-50.
6. Ouzounis CA, Valencia A (2003) Early bioinformatics: the birth of a discipline--a personal view. Bioinformatics 19: 2176-2190.
7. Molatudi M, Molotja N, Pouris A (2009) A bibliometric study of bioinformatics research in South Africa. Scientometrics 81: 47-59.
8. Ouzounis CA (2012) Rise and demise of bioinformatics? Promise and progress. PLoS Comput Biol 8: e1002487.
9. Geer RC, Sayers EW (2003) Entrez: making use of its power. Brief Bioinform 4: 179-184.
10. Parmigiani G, Garrett ES, Irizarry RA, Zeger SL (2003) The analysis of gene expression data: an overview of methods and software, Springer, New York.
11. Hoersch S, Leroy C, Brown NP, Andrade MA, Sander C (2000) The GeneQuiz web server: protein functional analysis through the Web. Trends Biochem Sci 25: 33-35.
12. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
13. 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.
14. Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. Bioinformatics 20: 426-427.
15. Thomas P, Starlinger J, Vowinkel A, Arzt S, Leser U (2012) GeneView: a comprehensive semantic search engine for PubMed. Nucleic Acids Res 40: W585-591.
16. Page RD (2001) TreeView. Glasgow University, Glasgow, UK.
17. Zhang Y, Phillips CA, Rogers GL, Baker EJ, Chesler EJ, et al. (2014) On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. BMC Bioinformatics 15: 110.
18. Stoilov I, Akarsu AN, Alozie I, Child A, Barsoum-Homsy M, et al. (1998) Sequence analysis and homology modeling suggest that primary congenital glaucoma on 2p21 results from mutations disrupting either the hinge region or the conserved core structures of cytochrome P4501B1. Am J Hum Genet 62: 573-584.
19. Tekaia F, Gordon SV, Garnier T, Brosch R, Barrell BG, et al. (1999) Analysis of the proteome of Mycobacterium tuberculosis in silico. Tuber Lung Dis 79: 329-342.
20. Mehmood MA, Xiao X, Hafeez FY, Gai Y, Wang F (2011) Molecular characterization of the modular chitin binding protein Cbp50 from Bacillus thuringiensis serovar konkukian. Antonie Van Leeuwenhoek 100: 445-453.
21. Sehar U, Mehmood MA, Hussain K, Nawaz S, Nadeem S, et al. (2013) Domain wise docking analyses of the modular chitin binding protein CBP50 from Bacillus thuringiensis serovar konkukian S4. Bioinformation 9: 901-907.
22. Kingsford CL, Ayanbule K, Salzberg SL (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol 8: R22.
23. Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 10: 516-522.
24. Lencz T, Guha S, Liu C, Rosenfeld J, Mukherjee S, et al. (2013) Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. Nat Commun 4: 2739.
25. Peng Z, Lu Y, Li L, Zhao Q, Feng Q, et al. (2013) The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla). Nat Genet 45: 456-461, 461e1-2.
26. Khan FA, Phillips CD, Baker RJ (2014) Timeframes of speciation, reticulation, and hybridization in the bulldog bat explained through phylogenetic analyses of all genetic transmission elements. Syst Biol 63: 96-110.
27. Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160: 712-726.
28. Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490.

29. Bast F (2013) Sequence similarity search, multiple sequence alignment, model selection, distance matrix and phylogeny reconstruction. Nat Protoc.

30. Ahmad N, Michoux F, Nixon PJ (2012) Investigating the production of foreign membrane proteins in tobacco chloroplasts: expression of an algal plastid terminal oxidase. PLoS One 7: e41722.

31. Chen Y, Wang F, Xu J, Mehmood MA, Xiao X (2011) Physiological and evolutionary studies of NAP systems in Shewanella piezotolerans WP3. ISME J 5: 843-855.

32. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365-370.

33. UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 42: D191-198.

34. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, et al. (2003) The Protein Information Resource. Nucleic Acids Res 31: 345-347.

35. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. Nucleic Acids Res 36: D25-30.

36. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, et al. (2005) The EMBL Nucleotide Sequence Database. Nucleic Acids Res 33: D29-33.

37. Miyazaki S, Sugawara H, Gojobori T, Tateno Y (2003) DNA Data Bank of Japan (DDBJ) in XML. Nucleic Acids Res 31: 13-16.

38. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.

39. Finn RD, Bateman A, Clements J, Marco Punta, Penny C Coggill, et al. (2014) Pfam: the protein families database. Nucl Acids Res 42: D222-D230.

40. Gonzalez S, Binato R, Guida L, Mencalha AL, Abdelhay E4 (2014) 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.

41. Fox NK, Brenner SE, Chandonia JM (2014) SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42: D304-309.

42. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, et al. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 33: D247-251.

43. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, et al. (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41: D344-347.

44. Huang JY, Brutlag DL (2001) The EMOTIF database. Nucleic Acids Res 29: 202-204.

45. UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36: D190-195.

46. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, et al. (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Res 40: D453-460.

47. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, et al. (2012) GenBank. Nucleic Acids Res 40: D48-53.

48. Bowes JB, Snyder KA, Segerdell E, Jarabek CJ, Azam K, et al. (2010) Xenbase: gene expression and improved integration. Nucleic Acids Res 38: D607-612.

49. St Pierre SE, Ponting L, Stefancsik R, McQuilton P; FlyBase Consortium (2014) FlyBase 102--advanced approaches to interrogating FlyBase. Nucleic Acids Res 42: D780-788.

50. Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Eng 12: 107-118.

51. Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, et al. (2000) The protein information resource (PIR). Nucleic Acids Res 28: 41-44.

52. UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142-148.

53. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35: D301-303.

54. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, et al. (2013) Rfam 11.0: 10 years of RNA families. Nucleic Acids Res 41: D226-232.

55. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, et al. (2014) The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42: D358-363.

56. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, et al. (2002) MINT: a Molecular INTeraction database. FEBS Lett 513: 135-140.

57. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, et al. (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33: D428-432.

58. Saier MH Jr, Tran CV, Barabote RD (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res 34: D181-186.

59. Saier MH Jr, Reddy VS, Tamang DG, Västermark A (2014) The transporter classification database. Nucleic Acids Res 42: D251-258.

60. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42: D490-495.

61. Subramaniam S, Fahy E, Gupta S, Sud M, Byrnes RW, et al. (2011) Bioinformatics and systems biology of the lipidome. Chem Rev 111: 6452-6490.

62. Huang T, He ZS, Cui WR, Cai YD, Shi XH, et al. (2013) A sequence-based approach for predicting protein disordered regions. Protein Pept Lett 20: 243-248.

63. Zakeri P, Jeuris B, Vandebril R, Moreau Y (2014) Protein fold recognition using geometric kernel data fusion. Bioinformatics 30: 1850-1857.

64. Yao L, Evans JA, Rzhetsky A (2010) Novel opportunities for computational biology and sociology in drug discovery. Trends Biotechnol 28: 161-170.

65. Sutcliffe MJ, Haneef I, Carney D, Blundell TL (1987) Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng 1: 377-384.

66. Bates PA, Sternberg MJ (1999) Model building by comparison at CASP3: using expert knowledge and computer automation. Proteins Suppl 3: 47-54.

67. Sali A, Blundell T (1994) Comparative protein modelling by satisfaction of spatial restraints. Protein Struc Dist Anal 64: 86.

68. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, et al. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31: 3784-3788.

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

70. Eswar N, Webb B, Martin-Renom MA, Shen MY, Pieper U, et al. (2006) Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics.

71. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31: 455-461.

72. Sehar U, Mehmood MA, Nawaz S, Nadeem S, Hussain K, et al. (2013) Three dimensional (3D) structure prediction and substrate-protein interaction study of the chitin binding protein CBP24 from B. thuringiensis. Bioinformation 9: 725-729.

73. Butt AM, Batool M, Tong Y (2011) Homology modeling, comparative genomics and functional annotation of Mycoplasma genitalium hypothetical protein MG_237. Bioinformation 7: 299-303.

74. Ali M, Mehmood MA, Hussain K (2013) Functional annotation of the cda1 gene from Bacillus thuringiensis through homology modeling and molecular docking. Pak J Life Soc Sci 11:190-195.

75. Wang L, Huang C, Yang MQ, Yang JY (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol 4 Suppl 1: S3.

76. Vinayagam A, Zirin J, Roesel C, Hu Y, Yilmazel B, et al. (2014) Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. Nat Methods 11: 94-99.

77. Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins 49: 350-364.

78. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30: 303-305.

79. Hosur R, Xu J, Bienkowska J, Berger B (2011) iWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions. J Mol Biol 405: 1295-1310.

80. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, et al. (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490: 556-560.

81. Valente GT, Acencio ML, Martins C, Lemke N (2013) The development of a universal in silico predictor of protein-protein interactions. PLoS One 8: e65587.

82. Zhang Z, Li Y, Lin B, Schroeder M, Huang B (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. Bioinformatics 27: 2083-2088.

83. de Vries SJ, van Dijk M, Bonvin AM (2010) The HADDOCK web server for data-driven biomolecular docking. Nat Protoc 5: 883-897.

84. Plewczynski D, Łaźniewski M, Augustyniak R, Ginalski K (2011) Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. J Comput Chem 32: 742-755.

85. Singh R, Xu J, Berger B (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Natl Acad Sci U S A 105: 12763-12768.

86. Liao CS, Lu K, Baym M, Singh R, Berger B (2009) IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics 25: i253-258.

87. Khan M, Masoud MS, Qasim M, Khan MA, Zubair M, et al. (2013) Molecular screening of phytochemicals from Amelanchier Alnifolia against HCV NS3 protease/helicase using computational docking techniques. Bioinformation 9: 978-982.

88. Klingström T, Plewczynski D (2011) Protein-protein interaction and pathway databases, a graphical review. Brief Bioinform 12: 702-713.

89. Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31: 248-250.

90. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40: D857-861.

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

92. Katara P (2013) Role of bioinformatics and pharmacogenomics in drug discovery and development process. Network Modeling Analysis in Health Informatics and Bioinformatics 2: 225-30.

93. Yamanishi Y, Kotera M, Kanehisa M, Goto S (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 26: i246-254.

94. Boruah L, Das A, Nainwal LM, Agarwal N, Shankar B (2013) In-Silico Drug Design: A revolutionary approach to change the concept of current Drug Discovery Process. Ind J Pharm Biologi Res 1: 60-73.

95. Lee HM, Giguere PM, Roth BL (2014) DREADDs: novel tools for drug discovery and development. Drug Discov Today 19: 469-473.

96. Wishart DS (2005) Bioinformatics in drug development and assessment. Drug Metab Rev 37: 279-310.

97. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3: 935-949.

98. Reddy RN, Mutyala R, Aparoy P, Reddanna P, Reddy MR (2007) Computer aided drug design approaches to develop cyclooxygenase based novel anti-inflammatory and anti-cancer drugs. Curr Pharm Des 13: 3505-3517.

99. Jayakanthan M, Wadhwa G, Mohan TM, Ponnusamy B, Durai S, et al. (2009) Computer-aided drug design for cancer-causing H-Ras p21 mutant protein. Lett Drug Des Discov 6: 14-20.

100. Kim J, Gao L, Tan K (2012) Multi-analyte network markers for tumor prognosis. PLoS One 7: e52973.

101. Wu CC, D'Argenio D, Asgharzadeh S, Triche T (2012) TARGETgene: a tool for identification of potential therapeutic targets in cancer. PLoS One 7: e43305.

102. Berg EL (2014) Systems biology in drug discovery and development. Drug Discov Today 19: 113-125.

103. McDermott JE, Diamond DL, Corley C, Rasmussen AL, Katze MG, et al. (2012) Topological analysis of protein co-abundance networks identifies novel host targets important for HCV infection and pathogenesis. BMC Syst Biol 6: 28.

104. Makhmoudova A, Williams D, Brewer D, Massey S, Patterson J, et al. (2014) Identification of multiple phosphorylation sites on maize endosperm starch branching enzyme IIb, a key enzyme in amylopectin biosynthesis. J Biol Chem 289: 9233-9246.

105. Hub JS, de Groot BL (2008) Mechanism of selectivity in aquaporins and aquaglyceroporins. Proc Natl Acad Sci U S A 105: 1198-1203.

106. Kannan S, Zacharias M (2014) Role of tryptophan side chain dynamics on the Trp-cage mini-protein folding studied by molecular dynamics simulations. PLoS One 9: e88383.

107. Isin B, Estiu G, Wiest O, Oltvai ZN (2012) Identifying ligand binding conformations of the Î²2-adrenergic receptor by using its agonists as computational probes. PLoS One 7: e50186.

108. Csermely P, Korcsmáros T, Kiss HJ, London G, Nussinov R (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol Ther 138: 333-408.

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

110. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29-37.

111. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7: 539.

112. Ganesan N, Bennett NF, Velauthapillai M, Pattabiraman N, Squier R, et al. (2005) Web-based interface facilitating sequence-to-structure analysis of BLAST alignment reports. Biotechniques 39: 186, 188.

113. Gasteiger E, Hoogland C, Gattiker A, Ron D Appel, Amos Bairoch, et al. (2005) In: The proteomics protocols handbook; Protein identification and analysis tools on the ExPASy server. Springer 571-607.

114. Allen JE, Salzberg SL (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21: 3596-3603.

115. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, et al. (2005) novoSNP, a novel computational tool for sequence variation discovery. Genome Res 15: 436-442.

116. Münch R, Hiller K, Grote A, Scheer M, Klein J, et al. (2005) Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. Bioinformatics 21: 4187-4189.

117. Unniraman S, Prakash R, Nagaraja V (2002) Conserved economics of transcription termination in eubacteria. Nucleic Acids Res 30: 675-684.

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

119. Kumar S, Tamura K, Nei M (1994) MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Comput Appl Biosci 10: 189-191.

120. Adachi J, Hasegawa M (1992) MOLPHY, programs for molecular phylogenetics I: PROTML, maximum likelihood inference of protein phylogeny. Institute of Statistical Mathematics, Tokyo.

121. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.

122. Felsenstein J (1993) PHYLIP phylogeny inference package. Department of Genetics, University of Washington, Seattle.

123. Boc A, Diallo AB, Makarenkov V (2012) T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Res 40: W573-579.

124. Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357-358.

125. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

126. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, et al. (2011) The European Nucleotide Archive. Nucleic Acids Res 39: D28-31.

127. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, et al. (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33: W116-120.

128. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, et al. (2005) PRIDE: the proteomics identifications database. Proteomics 5: 3537-3545.

129. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2012) Ensembl 2012. Nucleic Acids Res 40: D84-90.

130. Rajoka MI, Idrees S, Khalid S, Ehsan B (2014) Medherb: An Interactive Bioinformatics Database and Analysis Resource for Medicinally Important Herbs. Curr Bioinform 9: 23-27.

131. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619-622.

132. Gaudet P, Fey P, Basu S, Bushmanova YA, Dodson R, et al. (2011) dictyBase update 2011: web 2.0 functionality and the initial steps towards a genome portal for the Amoebozoa. Nucleic Acids Res 39: D620-624.

133. Kanehisa M (2002) The KEGG database. Silico Simulation of Biological Processes 247: 91-103.

134. Yang K, Dinasarapu AR, Reis ES, Deangelis RA, Ricklin D, et al. (2013) CMAP: Complement Map Database. Bioinformatics 29: 1832-1833.

135. Dinasarapu AR, Saunders B, Ozerlat I, Azam K, Subramaniam S (2011) Signaling gateway molecule pages--a data model perspective. Bioinformatics 27: 1736-1738.

136. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) PID: the Pathway Interaction Database. Nucleic Acids Res 37: D674-679.

137. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, et al. (2007) HMDB: the Human Metabolome Database. Nucleic Acids Res 35: D521-526.

138. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, et al. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res 41: D490-498.

139. Källberg M, Wang H, Wang S, Peng J, Wang Z, et al. (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7: 1511-1522.

140. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14: 892-893.

141. Rost B, Sander C, Schneider R (1994) PHD--an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 10: 53-60.

142. Bystroff C, Thorsson V, Baker D (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 301: 173-190.

143. Raghava G (2002) APSSP2: A combination method for protein secondary structure prediction based on neural network and example based learning. CASP5 A-132.

144. Kelley LA, Sternberg MJ (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 4: 363-371.

145. Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40: D302-305.

146. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, et al. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 30: 2785-2791.

147. Environment MO (2009) Chemical Computing Group. Montreal, Canada.

148. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41: D808-815.

149. Di Lena P, Wu G, Martelli PL, Casadio R, Nardini C (2013) MIMO: an efficient tool for molecular interaction maps overlap. BMC Bioinformatics 14: 159.

150. Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S (2006) Graemlin: general and robust alignment of multiple large interaction networks. Genome Res 16: 1169-1181.

151. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, et al. (2004) PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res 32: W83-88.

152. Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T (2006) CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 22: 1021-1023.

153. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.

154. Gao Z, Li H, Zhang H, Liu X, Kang L, et al. (2008) PDTD: a web-accessible protein database for drug target identification. BMC Bioinformatics 9: 104.

155. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, et al. (2014) DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 42: D1091-1097.

156. Chen X, Ji ZL, Chen YZ (2002) TTD: Therapeutic Target Database. Nucleic Acids Res 30: 412-415.

157. Magariños MP1, Carmona SJ, Crowther GJ, Ralph SA, Roos DS, et al. (2012) TDR Targets: a chemogenomics resource for neglected diseases. Nucleic Acids Res 40: D1118-1127.

158. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, et al. (2008) SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res 36: D919-922.

159. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, et al. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40: D1100-1107.

160. Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, et al. (2005) The Amber biomolecular simulation programs. J Comput Chem 26: 1668-1688.