Chapter 9

Bioinformatics: A Way Forward to Explore "Plant Omics"

Mehboob-ur- Rahman, Tayyaba Shaheen, Mahmood-ur- Rahman, Muhammad Atif Iqbal and Yusuf Zafar

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/64043

Abstract

Bioinformatics, a computer-assisted science aiming at managing a huge volume of genomic data, is an emerging discipline that combines the power of computers, mathematical algorithms, and statistical concepts to solve multiple genetic/biological puzzles. This science has progressed parallel to the evolution of genome-sequencing tools, for example, the next-generation sequencing technologies, that resulted in arranging and analyzing the genome-sequencing information of large genomes. Synergism of "plant omics" and bioinformatics set a firm foundation for deducing ancestral karyotype of multiple plant families, predicting genes, etc. Second, the huge genomic data can be assembled to acquire maximum information from a voluminous "omics" data. The science of bioinformatics is handicapped due to lack of appropriate computational procedures in assembling sequencing reads of the homologs occurring in complex genomes like cotton (2n = 4x = 52), wheat (2n = 6x = 42), etc., and shortage of multidisciplinary-oriented trained manpower. In addition, the rapid expansion of sequencing data restricts the potential of acquisitioning, storing, distributing, and analyzing the genomic information. In future, inventions of high-tech computational tools and skills together with improved biological expertise would provide better insight into the genomes, and this information would be helpful in sustaining crop productivities on this planet.

Keywords: databases, data mining, comparative genomics, plant genomes, sequence analysis, structure prediction

1. Introduction

Sustainability in agriculture systems is largely challenged by a number of factors including human population increase, environmental changes, and tremendous demands for growing crops to produce biofuels worldwide [1, 2]. In this regard, exploring the plant genomes for determining the function of important genes involved in conferring tolerance to biotic and abiotic stresses, followed by exploiting these genes in the development of resilient cultivars, is one of the durable strategies for bringing sustainability in crop yields [2, 3].

After the genome sequencing of *Arabidopsis thaliana* genome, a project was launched by the National Science Foundation (NSF) for determining the function of 25,000 Arabidopsis genes [4]. Rice was the first genome-sequenced crop (International Rice Genome Sequencing Project 2005) followed by sequencing of a number of genomes of major crops. All these sequencing projects released a large amount of data. For arranging and analyzing these data, a number of bioinformatic tools have been developed, which helped a lot in drawing important biological conclusions, predicting gene functions, etc. Furthermore, development of unconventional mapping populations and online resources of molecular markers [4] facilitate researchers to identify quantitative trait loci (QTLs). A number of databases have been developed to tackle the newly generated genomic data. These databases provided a foundation to build hypothesis, to design experiments, and to infer knowledge about a particular organism. Moreover, the datasets and "omics" resources of numerous species facilitated the assessment of "omics" properties among species, which further allows studying of conserved genes and evolutionary relationships. Bioinformatics is a crucial tool to access datasets of "omics" and to gather a substantial biological knowledge [5].

From the sequence analysis to the identification of genes, clustering of associated sequences and study of evolutionary relationships using phylogenetics are major tasks of bioinformatics. It also includes the identification and functional annotation of all genes, proteins, and active sites of protein structure in the cell [6]. At present, with the advancement in NGS tools, a voluminous sequencing data is emerging. For deducing meaningful information from these data, it is important for the science of bioinformatics to coevolve with the genomic tools. In this regard, the main three components including mathematics, computer science, and biology upon which the whole citadel of bioinformatics is based, should evolve in parallel to the sequencing tools. It would pave the way for deducing useful information (phylogenies, syntenic relationship, predicting genes, and their function) from the data in a shortest possible time [6, 7].

2. Databases

Databases are collection of organized data that can be retrieved from a website easily for addressing different queries. For managing and handling a database, different hardware and software programs in a computer are needed. The data are organized in structured records that can cater the easy retrieval of information. Broadly, biological databases are classified into

sequence databases, relevant to protein and nucleic acid sequences, and structure databases, only relevant to proteins. The first database was developed after a short period of sequencing the insulin protein in 1956. The "Protein Data Bank" was the first ever biological database developed in 1971. Biological databases have flourished enormously due to availability of huge amount of data being generated every day [8]. The individual laboratories maintained the preliminary databases of protein sequences; later, the creation of a combined formal database called SWISS-PROT protein sequence database was introduced in 1986. Now a plethora of data resources are available for study and research purposes and CDROMs (on request from), which are constantly being updated with the availability of new data [9].

Biological databases generally offer software tools to analyze the data available on it and to compare new data with already available data. With the help of these computational methods, the laborious and costly "wet lab" work can be avoided. In future, prospects are dealing with some hindrances such as limited awareness of data, complications in data retrieval, availability of limited data analysis tools, and inadequate literature reference accessibility [10]. A number of biological databases are available that can be divided into three categories on the basis of their contents: (1) primary databases—contain raw nucleotide sequences (GenBank, EMBL, and DDBJ), (2) secondary databases—contain highly annotated data (SWISS-PROT and Protein Information Resource), and (3) specialized databases—deal with particular organism and unique data (FlyBase, WormBase, and TAIR). A major problem in interlinking these databases is the lack of format compatibility. This problem is overcome by using a specified language known as Common Object Request Broker Architecture (CORBA) [11].

At National Center for Biotechnology Information (NCBI), text-based search and retrieval of information can be undertaken by deploying Entrez. It deals with all databases, for example, PubMed, Nucleotide and Protein Sequences, Complete Genomes, etc. In sequence retrieval system (SRS), the Boolean operators are used for undertaking complex searching. It is also used for sequence retrieval, abstract searching, references, etc.

2.1. Dedicated databases for plant genomics

A number of databases deal with datasets focused on particular genes and transcription factors (TFs) related to plant issues and cellular processes. First, a genome-wide finding of repertories of TFs encoded by genes in Arabidopsis genome was described [12]. Accessibility of complete genome sequences in the last few years has enabled us to assemble catalogs of TFs based on their function and association of regulatory systems in different plant species. Numerous databases deliver datasets about genes putatively involved in encoding TFs. These databases are based on predictions made by computational methods (sequence similarity search and hidden Markov model (HMM) conserved DNA-binding domains search). In recent years, GRASSIUS was established to compile resources and tools for undertaking comparative genomics of regulatory sequences in grass species [13]. The Grass Transcription Factor Database (GrassTFDB, another database) of GRASSIUS contains combined sequence information on RiceTFDB, MaizeTFDB, CaneTFDB, and SorghumTFDB. These can be searched through a website. Information of the predicted genes coding TF (carried out by doing

annotations across the three genome sequences of legumes) is available on the LegumeTFDB [14]—an extended database of the SoybeanTFDB.

The enhancement of the PGSB PlantsDB database framework has been accomplished with new tools, and sufficient new data have been added into the system particularly for the large complex genomes of wheat, barley, and rye. New resources such as GenomeZipper and CrowsNest for the comparative analysis of data RNASeq Expression Browser have been established. The transPLANT project makes available a platform to compile heterogeneous data about plant genome, for example, integrated searches over multiple databases (**Table 1**).

Database	URL	Species
RARTF	http://rarge.gsc.riken.jp/rartf/	Arabidopsis
AGRIS, AtTFDB	http://arabidopsis.med.ohio-state.edu/AtTFDB/	Arabidopsis
DATF	http://datf.cbi.pku.edu.cn/	Arabidopsis
DRTF	http://drtf.cbi.pku.edu.cn/	Rice
DPTF	http://dptf.cbi.pku.edu.cn/	Poplar
TOBFAC	http://compsysbio.achs.virginia.edu/tobfac/	Tobacco
SoybeanTFDB	http://soybeantfdb.psc.riken.jp/	Soybean
PlantTFDB	http://planttfdb.cbi.pku.edu.cn/ 22	Plant species
PlnTFDB	http://plntfdb.bio.uni-potsdam.de/v3.0/ 20	Plant species
GRASSIUS, GrassTFDB	http://grassius.org/grasstfdb.html	Maize, rice, sorghum, and sugarcane
LegumeTFDB	http://legumetfdb.psc.riken.jp/	Soybean, *Lotus japonicas*, and *Medicago truncatula*
DBD	http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Home	700 species
PlnTFDB	http://plntfdb.bio.uni-potsdam.de/v3.0/	Plant species
STIFDB	http://caps.ncbs.res.in/stifdb2/	Arabidopsis and rice
PlantTFDB	http://planttfdb.cbi.pku.edu.cn/	83 species
PGSB	http://pgsb.helmholtz-muenchen.de/plant/index.jsp	Barley, wheat, and rye

Table 1. Databases can be exploited for undertaking transcription factor studies in plants.

3. Analysis of the "omic" data

3.1. Sequence retrieval

First step is the identification and retrieval of sequences from different databases (NCBI, TAIR, Gramene, Rap-db, TIGR, Phytozome, PlantGDB, UniProt and SwissProt) developed for

handling protein, DNA, RNA, and Expressed Sequence Tag(EST) sequences (**Table 2**). Sequence retrieval is not only carried out through query words but it can also be done using BLAST and or their specific accession numbers. To find out similar sequences from databases, BLAST variations according to sequence retrieval could be performed.

S. No.	Database	URL	Description
1.	TRANSFAC	http://transfac.gbf.de/TRANSFAC/	Transcription factor database
2.	TFD	http://www.tfdg.com/Pages/tfddata.html	Transcription factor database
3.	TRRD	http://www.mgs.bionet.nsc.ru/mgs/dbases/trrd4/	Transcription regulatory region database
4.	PlantCARE	http://sphinx.rug.ac.be:8080/PlantCARE/	Plant cis-acting regulatory elements database
5.	PLACE	http://www.dna.affrc.go.jp/htdocs/PLACE/	Plant cis-acting regulatory elements database
6.	RegulonDB	http://www.cifn.unam.mx/Computational_Genomics/regulondb/	Database on transcriptional regulation in *Escherichia coli*
7.	SCPD	http://cgsigma.cshl.org/jian	Promoter database of yeast
8.	EPD	http://www.epd.isb-sib.ch/	Eukaryotic promoter database
9.	PRATT	http://web.expasy.org/pratt	It is an online server tool used to identify pattern of amino acids
10.	Phobius	http://phobius.sbc.su.se/	Identification of signal peptides
11.	SignalP 4.0	http://www.cbs.dtu.dk/services/SignalP/	Identification of signal peptides
12.	TargetP	http://www.cbs.dtu.dk/services/TargetP/	Subcellular localization of sequences
13.	LOCTREE3	https://www.rostlab.org/services/loctree3/	Subcellular localization of sequences
14.	Wolf PSORT	http://www.omictools.com/**wolf**-**psort**-tool	Subcellular localization of sequences
15.	Plant-mPLoc	www.csbio.sjtu.edu.cn/bioinf/**plant**-multi/	Subcellular localization of sequences
16.	Cello v2.5	https://bioinformatictools.wordpress.com/tag/**cello**/	Subcellular localization of sequences
17.	PSI-Pred	http://bioinf.cs.ucl.ac.uk/psipred/	Prediction of transmembrane regions of the gene
18.	CIMMiner	http://discover.nci.nih.gov/cimminer/	To explore the expression of a gene or protein on heat map
19.	DNASTAR	http://www.dnastar.com/	Making of sequence assembly
20.	PromPredict	http://nucleix.mbu.iisc.ernet.in/prompredict/prompredict.html	Promoter analysis
21.	FoldIndex	http://bioportal.weizmann.ac.il/	It is used to predict folding state of proteins

Table 2. Databases which are helpful in studying regulatory elements and promoter sequences of a gene.

3.2. Multiple sequence alignment

Multiple sequence alignment (MSA) deals with aligning three or more biological sequences, which may be DNA, RNA, and/or protein. Primarily, its purpose is to study similarity among sequences that can help to assess the evolutionary linkage and their common ancestry. It can be undertaken by many sequence analysis softwares including but not limited to ClustalW online software [15], ProbCons, and MAFFT [16]. Some other MSA tools are DNAMAN, T-Coffee, M-Coffee, R-Coffee, Expresso, PSI-Coffee, PSAlign, PRRN, MUSCLE, POA, MEME, etc.

A number of algorithms are available to generate MSA of proteins and DNA sequences. The basic approach in producing multiple alignments is to optimize the sum of pairs (SP) score. This approach is practical, and reproduces high-quality MSA dataset [17]. Mathematical approach (also called as probabilistic and stochastic methods) exploits the probability in developing MSA. Hidden Markov model is a masterpiece example of this approach. In this approach, MSA data are modeled as probabilistic models. All possible combination of gaps, mismatches, and matches are assigned with probabilities, and the algorithm finds the most likely MSA [18]. Other approaches are genetic algorithms and simulated annealing, which break a series of possible MSA into segments followed by their rearrangement. It can use an existing MSA and refines it by a series of rearrangements [19].

3.3. Domain and motif study

Domain always refers to a conserved part of protein sequence and structure, which can evolve, function, and exist independently. Whereas motif is a well-maintained sequence of protein or DNA that remains the same to execute certain function [20]. For characterization of a gene, it is always advisable to study its functional domains and motifs. The novel sequences identified can be subjected for analyzing their domains and motifs to predict their functions. For motif analysis, MEME tool can be used, while for domain analysis PFAM, InterProScan, and SMART tools can be used.

Large protein molecules comprise of structural and functional domains. Structural domains regions are either compact, globular modules, or separated clearly from the flanking regions including membrane regions or long coiled-coil helices that are separating the other domains [21]. These domains can be seen in proteins as semi-independent three-dimensional (3D), and have the ability to fold independently [22]. These domains constitute the “units of evolution” [23] and have typical functions [24]. Structural Classification of Proteins (SCOP) database has been used extensively for assigning domains in proteins [25]. Most databases and methods (e.g., Class Architecture Topology Homology database) are not fully automated, which combine several other methods for assigning domains to the proteins [26]. Protein Informatics System for Modeling (PrISM) is the only completely automated method that can be used to assign sequence-continuous domains to proteins of known 3D structures [27]. If the structure (3D) of the protein is not known, then a number of alternative methods and databases are available. For example, one of the most prominent databases is putative protein domains (ProDom) [28].

3.4. Structural analysis

Modeller is used to generate 3D structure [29]. LOMETS server is used to find the best template for comparative modeling. DOPE (discrete optimized protein energy) helps to find best model by calculating each structure's value that is evaluated through PROSAII [30] and PROCHECK [31]. To calculate electrostatic surface and solvation properties of complex compounds, APBS [32] is used. For structure alignment, PDBsum tool [33] is deployed. Structure of gene can also be displayed on GSDS2.0 (Gene Structure Display Server) [34]. YASARA software is used to draw 3D structure, c-terminal, n-terminal, and domains of proteins [35]. Chromosomal position of genes can be located by NCBI map viewer tool, Mapchar 2.1, and cucumber genome database map viewer tool.

3.5. Analysis of regulatory elements

The regulatory elements encode a protein that binds to promoter or operator region of a gene for up- and/or downregulating its expression. For instance, catabolite activator protein (CAP) is a regulatory element present in prokaryotes, which regulates the lac operon [36].

Regulation of gene expression takes place at transcription level by specific sequences known as transcription factors—inhibit or initiate the transcription. These factors can be repressors, activators, or both. It is worth mentioning that repressors inhibit the binding of RNA polymerase with the transcription complex (promoters)—thus blocking the transcription. However, activators are activated by the enabling binding of RNA polymerase with the transcription complex.

These elements can be found *in silico* by deploying PlantCARE [37], and PLACE program. PLACE is repository of motifs occur cis-acting regulatory DNA elements of plants. This database also gives information about the variations in motifs found in different genes or plant species. Relevant literature and comprehensive description of different motifs can be retrieved from this database. Several research groups have identified a number of genes including WRKY genes, Ascorbate Peroxidase, PSY, etc. using different bioinformatics tools [38–40].

3.6. Mutation identification

Mutation alters the nucleotide sequences of a gene that may change the gene expression. These mutations can be identified using conventional as well as NGS tools [41, 42]. Sequencing of cytosine methylome (methylC-seq), transcriptome (RNA-seq), and small RNA transcriptome (small RNA-seq) in Arabidopsis was undertaken by deploying NGS tools. Genome-scale methylation patterns and a direct relationship between the location of sRNAs and DNA methylation were identified [43]. Protein-protein interactions occur in majority cellular processes. The interactome, representing complete set of all protein-protein connections, is vital for studying the molecular networks [44]. Correlated mutation analysis can be harnessed to predict interface residues. Protein-protein interaction can be studied by detecting correlated mutations at interface [45].

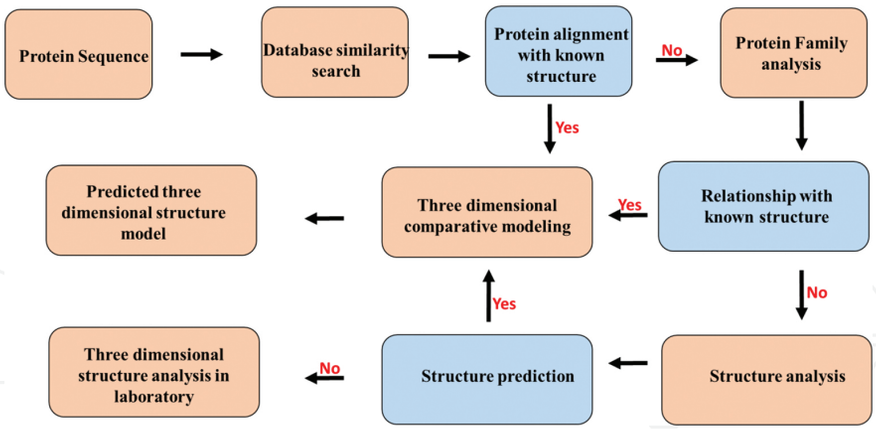

Figure 1. Flow chart diagram for protein structure prediction **(Source: Ref. [117]).**

S. No.	Software/server	Link	Description
1.	MODELLER	http://salilab.org/modeller/	Comparative modeling of protein 3D structures
2.	3DJigsaw	http://bmm.cancerresearchuk.org/~3djigsaw/	Predict structure and function of protein
3.	ESyPred3D	http://www.unamur.be/sciences/biologie/urbm/bioinfo/easypred/	Homology modeling with increased alignment performance
4.	SWISS-MODEL	http://swissmodel.expasy.org/	Automated protein homology modeling server
5.	YASARA	http://www.yasara.org/	Molecular modeling tool
6.	RaptorX	http://raptorx.uchicago.edu/	Protein structure prediction
7.	HHPred	http://toolkit.tuebingen.mpg.de/hhpred	Homology detection and structure prediction server
8.	Phyre^2	http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index	3D structure prediction
9.	ROSETTA	http://boinc.bakerlab.org/resetta/	3D structure prediction
10.	I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	Predict structure and function of protein
11.	Bhageerah	http://www.scfbio-iitd.res.in/bhageerath/index.jsp	Energy-based protein structure prediction server

Table 3. Bioinformatics tools which are helpful in predicting protein structure.

3.7. Protein structure prediction

It is the prediction of protein from amino acids. Protein structure can be predicted by undertaking similarity searches, MSAs, secondary structure prediction, identification of domains,

solvent accessibility predict, itself protein fold recognition, making 3D models, and model validation [46]. For example, small heat shock proteins (smHSPs, largely present in plants) are ubiquitous in nature, and their size is ranged from 17 to 30 kDa. These proteins are encoded by six nuclear gene families. Every gene family encodes a protein that is present in different part of the cell including cytosol, mitochondria, chloroplast, and endoplasmic reticulum. These proteins protect plants from high temperature stress [47].

3.7.1. Protein structure prediction steps

Following is the flow sheet diagram that elaborates the process of protein 3D structure prediction using bioinformatics tools (**Figure 1**). Various online and offline resources that can be used for the prediction of protein structure are described in **Table 3**.

3.8. Phylogenetic analysis

Phylogenetic analysis is the study of evolutionary relationships among different organisms. Phylogenetic analysis corresponds to the evolutionary interactions that can be presented in branching form. Phylogenetics refers as cladistics is a set of respective descendants such that it evolves from a respective single ancestor (**Figure 2**). Cladistics is a specific methodology of theorizing almost every evolutionary interactions [48]. In order to construct a phylogenetic tree, different methods are used that are based on the nature of the data and algorithms used. Each method is based on certain assumptions. Thus, the method used to draw evolutionary relationship on one kind of dataset may not be equally good for the other kind of dataset. It is therefore suggested that a number of distance-based methods [unweighted pair group method arithmetic mean (UPGMA) and neighbor joining (NJ)] and character-based (CB) methods [maximum parsimony (MP), maximum likelihood (ML)] should be run.

3.8.1. Distance-based method

The distance-based method also called as phonetic method depends upon the extent of dissimilarity (the distance) to derive a tree from the two aligned sequences. This method can rebuild the accurate tree if whole genetic divergence proceedings are precisely verified in the sequence. Tree construction is based on the resultant genetic distances from sequenced data, distances from immunological studies, and Euclidean distance applied in various ways [49].

3.8.1.1. Unweighted pair group method arithmetic mean (UPGMA)

It is the simplest procedure for studying the phylogenetic relationship among different organisms which uses the clustering approach and uncorrected data to make a tree. It joins tree branches based on the criterion of greatest similarity among pairs and averages of joined pairs. UPGMA generates a correct topology with true branch lengths only when the natural mutation is proportional to time (a molecular clock) or approximately equal to raw sequence dissimilarity [50]. However, these conditions are rarely met in practice. Distance matrix is recalculated, and this procedure is continued until the operational taxonomic units [OTU (=

neighbors)] are grouped in one cluster. However, this method does not reflect the evolutionary descents.

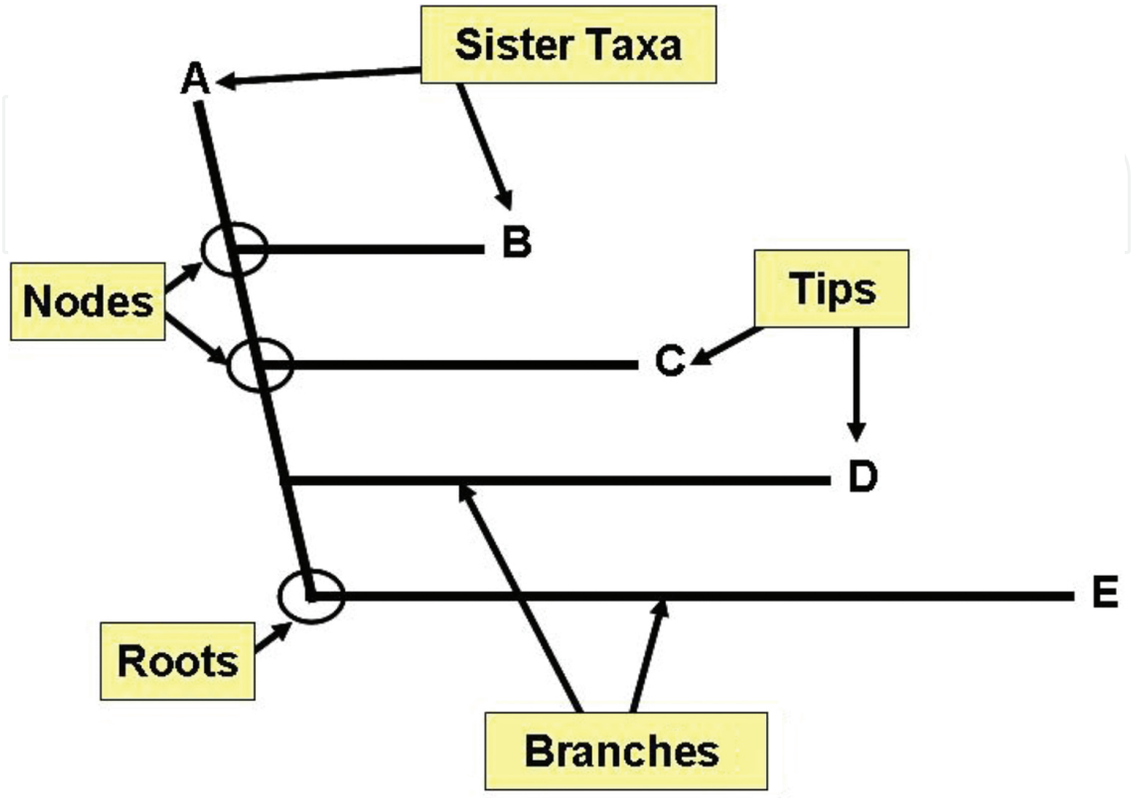

Figure 2. A descriptive diagram of phylogenetic analysis based on biological data.

3.8.1.2. Neighbor joining method (NJ)

This method is usually pragmatic with distance tree making, irrespective of optimization measure. This method works on the principle to discover pairs of OUTs(Operational Taxonomic Units) that curtails the total branch length at respective stage of clustering of OTUs beginning with a star-like tree. Branch length and distance matrix are recalculated until one terminal is found. This method can be used to obtain the branch lengths in addition to the topology of a parsimonious tree speedily [51]. This method is relatively efficient than that of the UPGMA. This method can analyze a large dataset. Construction of one possible tree and also the biased tree are the major drawbacks of this method.

3.8.2. Character-based methods (CB)

These methods are also called cladistic methods that use directly the aligned characters, for instance, DNA or protein sequences, through tree inference. The algorithm based on character takes an aligned set of characters, for example, DNA sequences, and builds a tree relating the changes in discrete characters, desirable to create the observed set of characters. These methods assume that a set of sequences descended from a common ancestor that may change by

mutation and/or selection process without involving any kind of hybridization or horizontal gene transfers. Character-based algorithms are comprised of two groups: maximum likelihood and maximum parsimony [50].

3.8.2.1. Maximum likelihood (ML)

Different statistical tools are exploited to assess hypothesis of evolutionary history. It constructs all possible trees of evolutionary history from a given data. Multiple alignment is done in this method. Probability of all possible topologies for each data partition is estimated to identify a tree with the highest probability at all partitions based on the maximum similar phylogeny. In this method, whole sequence information is used to evaluate all the possible trees. This method cannot handle a large amount of data.

3.8.2.2. Maximum parsimony (MP)

This method uses the philosophy of "the simpler hypothesis is better than the complicated ones" [52]. By this criterion, the MP tree is one with few character-state transformations for all the sequences from a common ancestor. It works by selecting trees that minimize the total tree length. For each site in the alignment, all possible trees are evaluated that is not the characteristic of other methods. This method is less dependent on suppositions about the evolution of sequences than the other strategies to construct a tree. This procedure is handicapped when the data are heterogeneous.

3.8.3. Evaluation of trees

Phylogenetic trees can be statistically evaluated for reliability of branches/clades created using (1) skewness test, (2) bootstrapping analysis, and (3) likelihood ratio tests where all have currently computerized algorithms. Skewness test never has approximation with dependability of specific topology; it is subtle to very small amounts of respective signal contemporary in otherwise random information set. Bootstrapping analysis is a resampling or rechecking tree evaluation methodology that works with distance, likelihood, and parsimony method. The outcome of bootstrap examination is a number related with specific branch in phylogenetic tree giving up the amount of bootstrap duplicates that ropes the monophyly of particular clade. Likelihood ratio tests support the likelihood ratio (tests) that is easily applicable to ML (maximum likelihood) examination. Value of likelihood is calculated for implication against normal circulation of fault in optimal models [50].

3.8.4. Software mostly used for phylogenetic analysis

Phylogeny inference package (PHYLIP) [53] contains 30 programs that cover the main flows of phylogenetic analysis. It is a freely available software and is accessible to almost all kinds of computer platforms (Mac, UNIX, DOC, etc.). In addition, phylogenetic analysis using parsimony (PAUP) software is widely used to infer and interpret the evolutionary tree. Now the old version has been upgraded (PAUP*) after the inclusion of maximum likelihood and distance methods. Other than those described above, some phylogenetic programs have

unique proficiencies but mostly inadequate in their respective actions, and movability. These include molecular phylogenetics (MOLPHY) [54], TREE-PUZZLE [55], FastDNAml [56], and MACCLADE [57].

3.9. Molecular dynamics simulations for plant molecules

Molecular dynamics simulations are the principal methods for elaborating the physical foundation of structure, function, and interaction of biological macromolecules (e.g., proteins and nucleic acids). Earlier, proteins were considered as comparatively rigid structures that now have been changed by a dynamic model in which the internal movements and conformational changes are key players in determining their functions. Computer simulations are carried out in comprehending the characteristics and arrangements of different molecules related to physical structure and interactions, otherwise not possible to observe by other means. There are two major classes of simulation techniques, i.e., molecular dynamics and Monte Carlo. These simulations have been used extensively in characterizing plant compounds (natural distillates) followed by finding optical counter parts with identical efficiency [58].

3.10. Proteomics and transcriptomics

Study of proteins along with mRNA transcripts is referred as proteomics and transcriptomics, respectively [59]. Due to intrinsic complexity, experimental workflows and variety of data types, storage, and open depository of proteomics data based on mass spectrometry (MS) are still insufficiently established. Many public sources with particular purposes for MS proteomics research have been established to fulfill this need. These databases are Global Proteome Machine Database (GPMDB), PRIDE, PeptideAtlas, ProteomicsDB, Mass Spectrometry Interactive Virtual Environment (MassIVE), PeptideAtlas SRM Experiment Library (PASSEL), etc. Moreover, for the purpose of enhanced integration and harmonized sharing of public warehouses, the ProteomeXchange consortium has been developed recently to capitalize on its advantage for the scientific community [60].

For transcriptomics studies, there are numerous databases comprising microarray data: NASCArrays, ArrayExpress, Genevestigator, Stanford Microarray Database, and the Gene Expression Omnibus, which are freely available [61]. An example of the transcriptome database is Chickpea Transcriptome Database (CTDB), which has information about the tools used for transcriptome sequence, conserved domain(s), molecular markers, transcription factor families, and complete gene expression information [62].

3.11. Protein-protein interactions

The protein-protein interactions (PPIs) control the expensive scope of biological procedures that include interactions between cells, metabolic as well as developmental pathways. This noncovalent bonding brings a range of interactions and associations between proteins. PPIs can be classified in several ways depending upon their contrasting structural and functional characteristics [63]. There are several *in vivo* and *in vitro* methods for finding PPIs but our focus

is on computational approaches. Computer modeling assisted with mathematical methods facilitates the study of different processes [64]. *In silico* methods combining the computational modeling are being used to study protein interactions. The *in silico* analysis integrates multiple data types including gene coexpression, colocalization, functional category, and the occurrence of orthologs or interologs to derive a global network in a species [65]. A list of webservers can be used to predict protein-protein interaction (**Table 4**).

S. No.	Web server	Description	URL
1.	Coev2Net	Coev2Net is a general framework to predict, assess, and boost confidence in individual interactions inferred from a high-throughput experiment	http://groups.csail.mit.edu/cb/coev2net/
2.	InterPreTS	InterPreTS uses tertiary structure to predict interactions	http://gabrmn.uab.es/interpret/
3.	PrePPI	PrePPI predicts protein interactions using both structural and nonstructural information	http://technology.sbkb.org/portal/page/350/
4.	iWARP	iWARP is a threading-based method to predict protein interaction from protein sequences	http://groups.csail.mit.edu/cb/iwrap/
5.	PoiNet	PoiNet provides PPI filtering and network topology from different databases	http://poinet.bioinformatics.tw/
6.	PreSPI	PreSPI predicts protein interactions using a combination of domains	http://code.google.com/p/prespi/
7.	PIPE2	PIPE2 queries the protein interactions between two proteins based on specificity and sensitivity	http://cgmlab.carleton.ca/PIPE2
8.	HomoMINT	HomoMINT predicts interaction in human based on ortholog information in model organisms	http://mint.bio.uniroma2.it/HomoMINT
9.	SPPS	SPPS searches protein partners of a source protein in other species	http://mdl.shsmu.edu.cn/SPPS/
10.	InPrePPI	InPrePPI predicts protein interactions in prokaryotes based on genomic context	http://inpreppi.biosino.org/InPrePPI/index.jsp
11.	STRING	STRING database includes protein interactions containing both physical and functional associations	http://string.embl.de
12.	MirrorTree	The MirrorTree allows graphical and interactive study of the coevolution of two protein families and assess their interactions in a taxonomic context	http://csbg.cnb.csic.es/mtserver/
13.	TSEMA	TSEMA predicts the interaction between two families of proteins based on Monte Carlo approach	http://tsema.bioinfo.cnio.es/
14.	COG	COG shows phylogenetic classification of proteins encoded in genomes	http://www.ncbi.nlm.nih.gov/COG/

Table 4. A number of important computational tools use to study protein-protein interactions.

3.11.1. Arabidopsis protein interaction analysis

More than 10 freely accessible protein interaction databases are available for *A. thaliana*. An intelligent bioinformatics web device, ANAP (Arabidopsis Network Analysis Pipeline) has been created for incorporating Arabidopsis protein collaboration databases. A total of 11 Arabidopsis protein collaboration databases having 201,699 protein association sets, 15,208 identifiers, 89 connection discovery routines, 73 species that interface with Arabidopsis, and 6161 references were incorporated in ANAP [66].

3.11.2. Computational identification of protein-protein interactions in rice

Complexity of plant molecules always hinders progress toward exploring the protein-protein interaction networks on large scale. A total of 5049 proteins with 76,585 interactions were predicted in rice using Predicted Rice Interactome Network (PRIN). The prolonged molecular network in PRIN has greatly improved the ability to analyze the function and organization of genes and gene networks [67].

3.11.3. iPlants: the world's plant online

This database has been designed to develop a comprehensive working list of scientific names of all plant species. Through this database, authenticated names of plant species (agreed by the scientific community) with their alternative synonyms can be found. This type of list will empower untrained botanists to get useful information about different plant species. *i*Plants will also resolve the existing confusions found in the published taxonomies. A total of 422,000 known plant species and 1,500,000–1,700,000 scientific names are used to refer these plant species are present in this database.

This database will help in exploiting plant biodiversity information in different breeding as well as gene cloning programs [68].

3.11.4. Reactome

Reactome database provides access without any restriction about the peer-reviewed pathways [69]. This database is equipped with bioinformatics tools, which can be used to examine, visualize, interpret, and analyze knowledge about pathway. The information in this database is generated by the experts (curators and software developers) and cross-referenced to other databases, for example, NCBI, Ensembl, UniProt, UCSC Genome Browser, HapMap, KEGG, ChEBI, PubMed, and GO. In this database, orthologous reaction for over 20 nonhuman species including rice, Arabidopsis, and *Escherichia coli* can be found. This database can be accessed in the form of online text book [70]. Biological pathways and reaction can be viewed in a number of formats, comprising of PDF, SBML, and BioPax [71]. Recent version "v55" of Reactome was released in December 2015.

3.12. Metabolomics

Study of all or utmost metabolites in an organism are denoted as metabolomics. It is a complex research field that involves interdisciplinary interaction of different sciences. One of the numerous methods is soft independent modeling of class analogy (SIMCA). Besides this, an effective protocol for data mining in metabolomics has also been developed [72]. In recent years, numerous databases containing data about compound names and structures, mass spectra, metabolic pathways, metabolite profile, and statistical/mathematical models are established. These databases are extremely useful for metabolomics research [73].

The MeRy-B (http://bit.ly/meryb) is dedicated to plants, and it provides information related to metabolites detected using NMR(Nuclear magnetic resonance), together with related analytical and experimental metadata. MeRy-B is equipped with a list of many plant metabolites along with the data of their experimental conditions, the features studied, and concentration of metabolites of 19 different species including the model plant species such as Arabidopsis [74].

4. Implications of bioinformatics in plant omics

Bioinformatics is an essential part of omics providing techniques to analyze large biological data sets and interpreting them into applications of "omic". Tools dealing with "omics" generate massive data that assist system biology to combine multivariate information into systems and models. The omics tools including high-throughput genome-scale genotyping platforms such as whole-genome resequencing, proteomics, and metabolomics offer better prospects for gene identification and exploration of molecular mechanisms. This information can be used to develop ideal genotypes suitable for varying climatic conditions [75].

4.1. Plant genome sequencing

With the advancements in high-throughput techniques, whole-genome characterization of a wide range of organisms has been possible. Nevertheless, the storage and management of this massive genomic data is a major challenge. Revolution in sequencing technologies has made it possible to sequence large and complex genomes at extremely low cost and in much less time period. Presently, the most popular methods of genome sequencing are shotgun sequencing and NGS. The NGS is very popular tool for the identification of housekeeping genes in crop plants. Many tools such as Genome Analyzer, the Applied Biosystems SOLiD System, Roche/454 FLX, and the Illumina/Solexa are commercially available for NGS [76]. NGS can be utilized for whole-genome sequencing, isolation of transcription factor binding sites, and expression of noncoding RNA and targeted resequencing [77]. Various software packages are available to assemble sequences, for example, Phred/Phrap/Consed [78], GAP4 [79], and chromaseq [80]. Another software called AMOS was developed by TIGR, which is useful for comparative genome assemblage [81].

4.2. Plant whole-genome resequencing

The most effective approach in functional genomics is the whole-genome resequencing. For reducing cost, target region can be sequenced. Microarray is also a common way of target region sequencing, which is based on hybridization to arrays comprising of synthetic oligonucleotides that match the target DNA sequence [82]. Recent NGS technology has made it possible to discover differences between individuals and populations especially of the crop species whose genomes have already been sequenced and assembled. Similar projects in Arabidopsis [83] and rice [84] generated a huge data of natural variations occurring within different accessions.

4.3. Plant comparative genomics and databases

Using comparative genomic approaches, functions to different genes (especially representing the less studied species) have been assigned. The developments in RNA interference and other technologies like mutagenesis have allowed phenotypic screens for genes—known as phenomics [85]. The field of phenomics is heavily dependent upon the interaction of plant genome with the prevailing environments. This science is largely dependent upon intensive collaboration between three disciplines including plant science, computer science, and engineering. Currently, there are yearly plant-focused image-processing tasks [86] that have positively stimulated the community and invigorated computer scientists to focus on developing joint plant datasets. Though there is limited accessibility to high-throughput phenotyping platforms. A current list of accessible image datasets can be accessed at the website [85].

4.4. Important information source of plant species

The most prevalent and unified information collection source is TAIR that maintains data of molecular biology, genetic and genomic of Arabidopsis [87]. Similarly, Salk Institute Genomic Analysis Laboratory (SIGnAL) deals with the omics research of Arabidopsis.

Gramene is an integrated source of information for grasses. It exploits the rice-genome-sequencing information as a foundation source for comparing the information of other members of grass family [88]. At this website, information about DNA and mRNA sequences, genome assembly and annotations, genes, genetic maps and physical maps, QTLs, and many more are available. These interesting features make this website more attractive for researchers, and it is being updated regularly with new attributes like genetic diversity data, comparison of genomes of *Oryza sativa* with its wild relatives or with the other taxa for undertaking evolutionary studies, etc. [89].

The portal site SoyBase [90] provides information about whole-genome sequence data. The portal site for Solanaceae genome is the Sol genomics network. It also provides information about the tomato-genome-sequencing project [91]. The MaizeGDB is a public database for *Zea mays* [92]. GreenPhylDB is a broad platform intended for facilitation of comparative functional genomics in *O. sativa* and *A. thaliana* genomes [93]. PLAZA 3.0 has been established to develop comparative genomics data of plants accessible via user-friendly web interface. Structural and functional annotation, phylogenetic trees protein domains, gene families, and detailed data

about genome organization can simply be inquired and envisioned [94]. A comparative genomics database named PIECE was established to accommodate information pertaining to gene structure comparisons and evolution. This database covers all the annotated genes mined from 25 plant species [95].

4.5. Use of bioinformatics for comparative genomics in plants

Availability of whole-genome sequences and bioinformatics have accelerated the process for identifying specific gene families in different plant species. These tools were also used to study the duplications as well as deletions in different plant genomes [96]. These results are helpful in phylogenetic studies [97], study of synteny and collinearity relationship, and inference of shared ancestry of genes [98]. The plant genome duplication database (PGDD) provides important data for studying the syntenic relationships of intragenome or cross-genome identified in the genome-sequenced species [99]. Analysis of orthologous clusters at genome level is a significant element in elucidating comparative genomics. Recognizing overlap between orthologous clusters can permit us to clarify the utility and evolution of proteins among multiple species. OrthoVenn is a web platform that is freely accessible and can be used for making comparisons and annotations of orthologous gene clusters. It can be accessed at [100]. Information regarding orthologs of plants and green algae can be searched at PlantOrDB [101].

4.6. Gene prediction and genome annotation

The characterization of introns and exons in a sequenced genome is referred as gene prediction. These predictions can be undertaken computationally or combination of manual as well as computational annotations. Numerous computer programs to find protein-coding genes are accessible through OMIC TOOLs website [102], which has been extensively used for genome annotations and genes prediction.

For structural annotations of a genome, a number of software packages were described [102, 103]. Additionally, tools (SynBrowse and VISTA) of genome comparison can be used to improve precision of gene identification. Repeat-Masker [104] was designed to find interspersed repeats and low complexity sequences in whole sequenced genome. Through this program, the repetitive sequences can also be masked. Similarly, a number of software programs (Repeat Finder, RECON, etc.) are available that can be used to find repeats in a sequenced genome.

4.7. Genome mapping and bioinformatics

Selecting suitable mapping tool and sequences search may claim adjustments in specificity and sensitivity of the search statistics. The process of finding candidate genes conferring traits can be accelerated for those crops where genetic and physical maps and annotated genome assemblies are available. A wide range of tools have been developed recently for illustrating maps and imagining genomes primarily to facilitate genome assembly.

NCBI is a source to assess all types of information regarding genomes. Access to various biological databases is possible using "Entrez." For aligned genetic, physical, and sequence information of eukaryotes including plants, a genome browser "Map viewer" has been developed. To display aligned map from various species entered in Map Viewer, a special plant query page can be accessed. Customized plant basic local alignment search tool (Plant BLAST) facilitates the process of exploring sequence similarity from the collection of mapped plants sequence data, and the resulting alignment can be visualized in genomic text using "Map viewer" [105], R/qtl [106], JoinMap [107], OneMap [108], MSTMap [109], Lep-MAP [110], and HighMap [111], which can be used to develop genetic linkage maps [112].

Numerous databases offer data for exploring markers in multiple crop species. DNA markers including Single nucleotide polymorphism (SNP), Simple sequence repeat (SSR), and conserved ortholog set (COS) markers can be predicted using PlantMarkers [113]. A famous site for Triticeae genome is GrainGenes that contains information about linkage maps and DNA markers of wheat, rye, barley, and oat [114]. Gramene, a database for comparative genomics, contains genetic maps of multiple plant species [89]. The Triticeae Mapped EST database (TriMEDB) gives information of mapped cDNA markers related to barley and wheat [115]. The CottonGen web-based database provides information and open access to genetic, genomic, and breeding data of cotton. CottonGen has improved tools for sharing, mining, retrieval, and visualization of data as compared with the CottonDB and Cotton Marker Database [116].

5. Conclusions

In this chapter, we described comprehensively the available resources and tools of bioinformatics pertaining to gene expression, databases, protein, and metabolite analyses and genome sequencing. Bioinformatics has been evolved rapidly over the last 15 years—emerged as a new key discipline of biology. A huge amount of genetic and genomic data have been generated using next-generation sequencing technologies that provide opportunities for generating huge genetic and genomic data. However, drawing useful genetic information is handicapped due to unavailability of skilled bioinformaticians. Still, there is room for some unsolved problems in bioinformatics like computerized data mining, vigorous inference of phenotype from genotype, trainings of students and recognized researchers in bioinformatics, etc. Bioinformatics is generating job opportunities for brilliant and skilled researchers in biology, statistics, and computer science. The remarkable evolution of bioinformatics has been confronted by a number of troublesome revolutions in science and technology. Even though, bioinformatics has developed possibly itself to a level above recognition. Today's bioinformatics is a luxury to biological scientists, generating huge data in all fields of biological sciences. In near future, bioinformatics will be an indispensable part of plant research, and novel tools and methods will be incorporated by every plant scientist. The next half century is the era of "data integration." Both basic and applied research will replenish the society for renewable energy, dropping world hunger and poverty, and protecting the environment.

Author details

Mehboob-ur- Rahman[1*], Tayyaba Shaheen[2], Mahmood-ur- Rahman[2], Muhammad Atif Iqbal[1] and Yusuf Zafar[3]

*Address all correspondence to: mehboob_pbd@yahoo.com and mehboob@nibge.org

1 National Institute for Biotechnology and Genetic Engineering, Faisalabad, Pakistan

2 Department of Bioinformatics and Biotechnology, Government College University, Faisalabad, Pakistan

3 Department of Technical Co-operation, IAEA, Vienna International Centre, Vienna, Austria

References

[1] Turner WR, Oppenheimer M, Wilcove DS. A force to fight global warming. Nature. 2009;462:278–279. DOI: 10.1038/462278a

[2] Azadi H, Ghanian M, Ghoochani OM, Taning CNT, Hajivand RY, Dogot T. Genetically modified crops: towards agricultural growth, agricultural development, or agricultural sustainability?. Food Reviews International. 2015;31:195–221. DOI: 10.1080/87559129.2014.994816

[3] Takeda S, Matsuoka M. Genetic approaches to crop improvement: responding to environmental and population changes. Nature Reviews Genetics. 2008;9:444–457. DOI: 10.1038/nrg2342

[4] Ruperao P, Edwards D. Identification of markers from next-generation sequence data plant genotyping. Methods in Molecular Biology. Bioinformatics. 2014;1245:29–47. DOI: 10.1007/978-1-4939-1966-6_3

[5] Mochida K, Shinozaki K. Genomics and bioinformatics resources for crop improvement. Plant Cell Physiology. 2010;51:497–523. DOI: 10.1093/pcp/pcq027

[6] Rao VS, Das SK, Rao VJ, Srinubabu G. Recent developments in life sciences research: role of Bioinformatics. African Journal of Biotechnology. 2008;7:495–503

[7] Alemu K. The role and application of bioinformatics in plant disease management. Advances in Life Science and Technology. 2015;28:28–33

[8] Cooray MPNS. Molecular biological databases: evolutionary history, data modeling, implementation and ethical background. Sri Lanka Journal of Bio-Medical Informatics. 2012;3:2–11. DOI: 10.4038/sljbmi.v3i1.2489

[9] Babu MM. Biological databases and protein sequence analysis [Internet]. 1997. Available from http://www.mrc-lmb.cam.ac.uk/genomes/madanm/pdfs/biodbseq.pdf [Accessed: 2016-04-26]

[10] Bry FK, Kröger P. A computational biology database digest: data, data analysis, and data management. Distributed Parallel Databases 2003;13:7–42. DOI: 10.1023/A:1021540705916

[11] Brown JR, editor. Handbook of Comparative Genomics: Basic and Applied Research. 1st ed. Boca Raton, FL: CRC Press; 2007. 400 p.

[12] Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, et al. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science. 2000;290:2105–2110. DOI: 10.1126/science.290.5499.2105

[13] Yilmaz A, Nishiyama MY, Jr., Fuentes BG, Souza GM, Janies D, Gray J, et al. GRASSIUS: a platform for comparative regulatory genomics across the grasses. Plant Physiology. 2009;149:171–180. DOI: 10.1104/pp.108.128579

[14] Mochida K, Yoshida T, Sakurai T, Yamaguchi-Shinozaki K, Shinozaki K, Lam-Son Phan Tran. LegumeTFDB: an integrative database of *Glycine max*, *Lotus japonicus* and *Medicago truncatula* transcription factors [Internet]. 2009. Available from: http://legumetfdb.psc.riken.jp [Accessed: 2016-04-28]

[15] Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Research. 2003;31:3497–3500. DOI: 10.1093/nar/gkg500

[16] Nuin PAS, Wang Z, Tillier ERM. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics. 2006;7:471–488. DOI: 10.1186/1471-2105-7-471

[17] Waterman MS, Smith TF, Beyer WA. Some biological sequence metrics. Advances in Mathematics. 1976;20:367–387. DOI: 10.1016/0001-8708(76)90202-4

[18] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32:1792–1797. DOI: 10.1093/nar/gkh340

[19] Goldberg DE, editor. Handbook of Genetic Algorithms in Search, Optimization and Machine Learning. 1st ed. Reading, MA: Addison-Wesley Longman; 1989. 372 p.

[20] Roey KV, Davey NE. Motif co-regulation and co-operativity are common mechanisms in transcriptional, post-transcriptional and post-translational regulation. Cell Communication Signal. 2015;13:45–60. DOI: 10.1186/s12964-015-0123-9

[21] Abrahams JP, Leslie AGW, Lutter R, Walker JE. Structure at 2.8 Å resolution of F_1-ATPase from bovine heart mitochondria. Nature. 1994;370:621–628. DOI: 10.1038/370621a0

[22] Jaenicke R. Folding and association of proteins. Progress in Biophysics and Molecular Biology.1987;49:117–237. DOI: 10.1016/0079-6107(87)90011-3

[23] Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Research. 1994;22:3600–3609.

[24] Teichmann SA, et al. The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. Journal of Molecular Biology. 2001;311:693–708. DOI: 10.1006/jmbi.2001.4912

[25] Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Research. 2002;30:264–267. DOI: 10.1093/nar/30.1.264

[26] Orengo CA, et al. The CATH protein family database: a resource for structural and functional annotation of genomes. Proteomics. 2002;2:11–21. DOI: 10.1002/1615-9861(200201)2

[27] Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. Journal of Molecular Biology. 2000;301:691–711. DOI: 10.1006/jmbi.2000.3975

[28] Corpet F, Gouzy F, Kahn D. The ProDom database of protein domain families. Nucleic Acids Research. 1998;26:323–326. DOI: 10.1093/nar/26.1.323

[29] Sali A. Modeller: program for comparative protein structure modeling by satisfaction of spatial restraints [Internet]. Available from: https://salilab.org/modeller/ [Accessed: 2016-04-21]

[30] Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins [Internet]. 2007. Available from: https://prosa.services.came.sbg.ac.at/prosa.php [Accessed: 2016-04-21]

[31] Roman AL, MacArthur MW, Smith DK, Jones DT, Hutchinson EG, Morris AL, Moss DS, Thornton JM. PROCHECK-Programs to check the Stereochemical Quality of Protein Structures [Internet]. 1998. Available from: http://www.ebi.ac.uk/thornton-srv/28 software/PROCHECK [Accessed:2016-07-11]

[32] Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proceedings of the National Academy of Sciences. 2001:98;10037–10041. DOI: 10.1073/pnas.181342398

[33] The European Bioinformatics Institute [Internet]. Available from: http://www.ebi.ac.uk/Tools/structure/ [Accessed: 2016-04-18]

[34] Hu B, Jin J, Guo AY, Zhang H, Luo J, Gao G. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics. 2015:31;1296–1297.

[35] Ratan A. ASSEMBLY algorithms for next generation sequence data [Internet]. 2009. Available from: http://www.yasara.org/products.htm#structure

[36] Maston GA, Evans K, Green MR. Transcriptional regulatory elements in the human genome. Annual Review of Genomics and Human Genetics. 2006;7:29–59. DOI: 10.1146/annurev.genom.7.080505.115623

[37] Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences [Internet]. 2002. Available from: http://hdl.handle.net/1854/LU-153936 [Accessed: 2004-01-14]

[38] Pandey S, Negi YK, Marla SS, Arora S. Comparative in silico analysis of ascorbate peroxidase protein sequences from different plant species. Journal of Bioengineering & Biomedical Science. 2011;1:103–107. DOI: 10.4172/2155-9538.1000103

[39] Han Y, Zheng QS, Wei YP, Chen J, Liu R, Wan HJ. *In silico* identification and analysis of phytoene synthase genes in plants. Genetics and Molecular Research. 2015;14:9412–9422. DOI: 10.4238/2015

[40] Rao G, Sui J, Zhang J. In silico genome-wide analysis of the WRKY gene family in Salix arbutifolia. Plant Omics Journal. 2015;8:353–360.

[41] He G, Elling AA, Deng XW. The epigenome and plant development. Annual Review of Plant Biology. 2011;62:411–435. DOI: 10.1146/annurev-arplant-042110-103806

[42] Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics. 2010;11:204–220. DOI: 10.1038/nrg2719

[43] Lister R, Gregory BD, Ecker JR. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Current Opinion in Plant Biology. 2009;12:107–118. DOI: 10.1016/j.pbi.2008.11.004

[44] Morsy M, Gouthu S, Orchard S, Thorneycroft D, Harper JF, Mittler R, Cushman JC. Charting plant interactomes: possibilities and challenges. Trends in Plant Sciences. 2008; 13:183–191. DOI: 10.1016/j.tplants.2008.01.006

[45] Guo F, Ding Y, Li Z, Tang J. Identification of protein-protein interactions by detecting correlated mutation at the interface. Journal of Chemical Information and Modeling. 2015;55:2042–2049. DOI: 10.1021/acs.jcim.5b00320

[46] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PLoS One. 2011;6:e28766. DOI: 10.1371/journal.pone.0028766

[47] Waters ER, Lee GJ, Vierling E. Evolution, structure and function of the small heat shock proteins in plants. Journal of Experimental Botany. 1996;47:325–338. DOI: 10.1093/jxb/47.3.325

[48] Brinkman FS, Leipe DD. Phylogenetic analysis. Methods of Biochemical Analysis. 2001;43:323–358.

[49] Wiley EO, Lieberman BS, editors. Handbook of Phylogenetics: Theory and Practice of Phylogenetic Systematics. 2nd ed. Chichester: Wiley; 2011. 432 p.

[50] Peng C. Distance based methods in phylogenetic tree construction. Neural Parallel and Scientific Computations. 2007;15:547–560.

[51] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987;4:406–425.

[52] Swofford DL, Olsen GJ, Waddel PJ, Hillis DM. Phylogenetic inference. In: Hillis DM, Moritz C, Mableeds BK, editors. Hand Book of Molecular Systematics. 2nd ed. Sunderland: Sinauer Associates;1996. pp. 407–446.

[53] Felsenstein, J. PHYLIP (Phylogeny Inference Package) [Internet]. 1993. Available from: http://evolution.genetics.washington.edu/phylip.html [Accessed: 2016-04-11]

[54] Adachi J, Hasegawa M. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood [Internet]. 1996. Available from: http://www.ism.ac.jp/ismlib/softother.e.html [Accessed: 2016-04-21]

[55] Schmidt HA, Strimmer K, Vingron M, Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 2002;18:502–504. DOI: 10.1093/bioinformatics/18.3.502

[56] Olsen GJ, Matsuda H, Hagstrom R, Overbeek R. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Computer Applications in the Biosciences. 1994;10:41–48. DOI: 10.1093/bioinformatics/10.1.41

[57] Maddison DR, Maddison WP. MacClade 4: analysis of phylogeny and character evolution [Internet]. 2005. Available from: http://macclade.org

[58] Kopec W, Telenius J, Khandelia H. Molecular dynamics simulations of the interactions of medicinal plant extracts and drugs with lipid bilayer membranes. FEBS Journal. 2013;280:2785–2805. DOI: 10.1111/febs.12286

[59] Turenne N. Role of a web-based software platform for systems biology. Journal of Computer Science and System Biology. 2011;4:35–41. DOI: 10.4172/jcsb.1000101e

[60] Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaíno JA. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics. 2015;15:930–949. DOI: 10.1002/pmic.201400302

[61] Bhardwaj T, Somvanshi P. Plant systems biology: insights and advancements. In: Barh D, Khan MS, Davies E, editors. Hand Book of Plant Omics: The Omics of Plant Science. New Delhi:Springer; 2015. pp. 791–819. DOI: 10.1007/978-81-322-2172-2_28

[62] Verma M, Kumar V, Patel RK, Garg R, Jain M. CTDB: an integrated chickpea transcriptome database for functional and applied genomics. PLoS One. 2015;10:0136880. DOI: 10.1371/journal.pone.0136880

[63] Nooren IMA, Thornton JM. Diversity of protein-protein interactions. The EMBO Journal. 2003;22:3486–3492. DOI 10.1093/emboj/cdg359

[64] You L. Toward computational systems biology. Cell Biochemistry and Biophysics. 2004;40:167–184. DOI: 10.1385/CBB:40:2:167

[65] Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Molecular System Biology. 2007;3:88–100. DOI 10.1038/msb4100129

[66] Wang C, Marshall A, Zhang D, Wilson ZA. ANAP: an integrated knowledge base for Arabidopsis protein interaction network analysis. Plant Physiology. 2012;158:1523–1533. DOI: 10.1104/pp.111.192203

[67] Zhu P, Gu H, Jiao Y, Huang D, Chen M. Computational identification of protein-protein interactions in rice based on the predicted rice interactome network. Genomics Proteomics Bioinformatics. 2011;9:128–137. DOI: 10.1016/S1672-0229(11)60016-8

[68] Alkin B. iPlants-the world's plant on line [Internet]. 2004. Available from: iPlants document library: www.iplants.intranets.com

[69] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. Nucleic Acids Research. 2005;33:428–432. DOI: 10.1093/nar/gki072

[70] Haw R, Stein L. Using the reactome database. In: Andreas D, et al editors. Hand book of Current protocols in bioinformatics. Chichester:Wiley; 2012. DOI: 10.1002/0471250953.bi0807s38

[71] Croft D. Building models using Reactome pathways as templates. Methods in Molecular Biology. 2013;1021:273–283. DOI: 10.1007/978-1-62703-450-0_14

[72] Fukusaki E, Kobayashi A. Plant metabolomics: potential for practical operation. Journal of Bioscience and Bioengineering. 2005;100:347–354. DOI: 10.1263/jbb.100.347

[73] Fukushima A, Kusano M. Recent progress in the development of metabolome databases for plant systems biology. Frontiers in Plant Science. 2013;4:73. DOI: 10.3389/fpls.2013.00073

[74] Deborde C, Jacob D, MeRy-B. A metabolomic database and knowledge base for exploring plant primary metabolism. Methods in Molecular Biology. 2014;1083:3–16. DOI: 10.1007/978-1-62703-661-0_1

[75] Iquebal MA, Jaiswal S, Mukhopadhyay CS, Sarkar C, Rai A, Kumar D. Applications of bioinformatics in plant and agriculture. In: Barh D, Khan MS, Davies E, editors. Hand

Book of Plant Omics: The Omics of Plant Science. Springer; 2015. pp. 755–790. DOI: 10.1007/978-81-322-2172-2_1

[76] Agrawal PK, Babu BK, Saini N. Omics of model plants. In: Barh D, Khan MS, Davies E, editors. Hand Book of Plant Omics: The Omics of Plant Science. New Delhi: Springer; 2015. pp 1–32. DOI: 10.1007/978-81-322-2172-2_1

[77] Tiwary BK. Next-Generation Sequencing and Assembly of Plant Genomes. In: Barh D, Khan MS, Davies E, editors. Hand Book of Plant Omics: The Omics of Plant Science. New Delhi:Springer; 2015. pp. 53–64. DOI: 10.1007/978-81-322-2172-2_1

[78] Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Research.1998;8:195–202.

[79] The GAP Group. GAP—groups, algorithms, and programming, version 4.8.3 [Internet]. 2016. Available from: http://www.gap-system.Org[Accessed:2016-07-11]

[80] Maddison DR, Maddison WP. Chromaseq: a Mesquite package for analyzing sequence chromatograms [Internet]. 2014. Available from: http://mesquiteproject.org/packages/Chromaseq[Accessed:2016-07-11]

[81] Rhee SY, Dickerson J, Dong Xu. Bioinformatics and its applications in plant biology. Annual Reviews of Plant Biology. 2006;57:335–360. DOI: 10.1146/ annurev.arplant.56.032604.144103

[82] Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M . Comparison of next-generation sequencing systems. Journal of Biomedicine and Biotechnology. 2012; DOI: 10.1155/2012/251364

[83] Weigel D, Mott R. The 1001 genomes project for *Arabidopsis thaliana*. Genome Biology. 2009;10:107–111. DOI: 10.1186/gb-2009-10-5-107

[84] Li JY, Wang J, Zeigler RS, The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience. 2014;3:8–10. DOI: 10.1186/2047-217X-3-8

[85] Fahlgren N, Gehan MA, Baxter I. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Current Opinion in Plant Biology. 2015;24;93–99. DOI: 10.1016/j.pbi.2015.02.006

[86] Rick HJ, de Zedde V. IPPN: International Plant Phenotyping Network [Internet]. 2015. Available from: http://www.plant-phenotyping.org [Accessed: 2016-04-21]

[87] Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Research. 2008;36:1009–1014. DOI: 10.1093/nar/gkm965

[88] Ware D, Jaiswal P, Ni, J et al. Gramene: a resource for comparative grass genomics. Nucleic Acids Research. 2002;30:103–105. DOI: 10.1093/nar/30.1.103

[89] Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, et al. Gramene: a growing plant comparative genomics resource. Nucleic Acids Research. 2008;36:947–953. DOI: 10.1093/nar/gkm968

[90] Grant, D, Nelson, RT, Cannon, SB, Shoemaker, RC. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Research. 2010;38:843–846. DOI: 10.1093/nar/gkp798

[91] Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, Strickler SR, Bombarely A, Fisher-York T, Pujar A, Foerster H, Yan A, Mueller LA. The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. Nucleic Acids Research. 2015;43:1036–1041.

[92] Lawrence CJ, Dong Q, Polacco ML, Seigfried TE, Brendel V. MaizeGDB, the community database for maize genetics and genomics. Nucleic Acids Research. 2004;32:393–397. DOI: 10.1093/nar/gkh011

[93] Conte MG, Gaillard S, Lanau N, Rouard M, Périn C. GreenPhylDB: a database for plant comparative genomics. Nucleic Acids Research. 2008;36:991–998. DOI: 10.1093/nar/gkm934

[94] Proost S, Van Bel MV, Vaneechoutte D, de Peer YV, Inze D, Mueller-Roeber B, Vandepoele K. PLAZA 3.0: an access point for plant comparative genomics Nucleic Acids Research. 2015;43:974–981. DOI: 10.1093/nar/gku986

[95] Wang Y, You FM, Lazo GR, Luo M-C, Thilmony R, Gordon S, Shahryar F, Kianian SF, Gu YQ. PIECE: a database for plant gene structure comparison and evolution. Nucleic Acids Research. 2012;41:1159–1166. DOI: 10.1093/nar/gks1109

[96] Sterck L, Rombauts S, Vandepoele K, Rouze P, Van de Peer Y. How many genes are there in plants (and why are they there)? Current Opinion in Plant Biology. 2007;10:199–203. DOI: 10.1016/j.pbi.2007.01.004

[97] Wall PK, Leebens-Mack J, Muller KF, Field D, Altman NS, dePamphilis CW. PlantTribes: a gene and gene family resource for comparative genomics in plants. Nucleic Acids Research. 2008;36:970–976. DOI: 10.1093/nar/gkm972

[98] Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. Science. 2008;320:486–488. DOI: 10.1126/science.1153917

[99] Lee TH, Tang H, Wang X, Paterson AH. PGDD: a database of gene and genome duplication in plants. Nucleic Acids Research. 2012;41:1152–1158. DOI: 10.1093/nar/gks1104

[100] Wang Y, Coleman-Derr D, Chen G, Gu YQ. OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Research. 2015;43:78–84. DOI: 10.1093/nar/gkv487

[101] Li F, Fan G, Lu C, Xiao G, Zou C, Kohel R, Ma Z, Shang H, Ma X, Wu J, Liang X, Huang G, G Percy R, Liu K, Yang W, et al. Genome sequence of cultivated upland cotton

(*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nature Biotechnology. 2015;33:524–530. DOI: 10.1038/nbt.3208

[102] Hudek AK, Cheung J, Boright AP, Scherer SW. Genescript: DNA sequence annotation pipeline. Bioinformatics. 2003;19:1177–1178. DOI: 10.1093/bioinformatics/btg134

[103] Allen JE, Pertea M, Salzberg SL. Computational gene prediction using multiple sources of evidence. Genome Research. 2004;14:142–148. DOI: 10.1101/gr.1562804

[104] Smit AFA, Hubley R, Green P. Repeat Masker Open-3.0 [Internet]. 1996–2010. Available from: http://www.repeatmasker.org [Accessed: 2016-04-21]

[105] Tatusova T, Smith White B, Ostell J. A collection of plant-specific genomic data and resources at NCBI. In: Edwards E, editor. Hand Book of Plant Bioinformatics, Methods and Protocols. New York:Humana Press; 2007. pp. 61–87.

[106] Broman KW, Wu H, Sen S, Churchill G A. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19:889–890. DOI: 10.1093/bioinformatics/btg112

[107] Jansen J, de Jong AG, van Ooijen JW. Constructing dense genetic linkage maps. Theoretical and Applied Genetics. 2001;102:1113–1122. DOI: 10.1007/s001220000489

[108] Margarido GRA, Souza AP, Garcia AAF. One map: software for genetic mapping in outcrossing species. Hereditas. 2007;144:78–79. DOI: 10.1111/j.2007.0018-0661.02000.x

[109] Wu Y, Bhat PR, Close TJ, Lonardi S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genetics. 2008;4:e1000212. DOI: 10.1371/journal.pgen.1000212

[110] Rastas P, Paulin L, Hanski I, Lehtonen R, Auvinen P. Lep-MAP: fast and accurate linkage map construction for large SNP datasets. Bioinformatics. 2013;29:3128–3134. DOI: 10.1093/bioinformatics/btt563

[111] Liu D, Ma C, Hong W, Huang L, Liu M, Liu H, et al. Construction and analysis of high-density linkage map using high-throughput sequencing data. PLoS One. 2014;9:e98855. DOI: 10.1371/journal.pone.0098855

[112] Fierst J. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Frontiers in Genetics. 2015;6:220. DOI: 10.3389/fgene.2015.00220

[113] Heesacker A, Kishore VK, Gao W, Tang S, Kolkman JM, Gingle A, et al. SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility. Theoretical and Applied Genetics. 2008;117:1021–1029. DOI: 10.1007/s00122-008-0841-0

[114] Carollo V, Matthews DE, Lazo GR, Blake TK, Hummel DD, Lui N. Grain genes. An improved resource for the small grains community. Plant Physiology. 2005;139:643–651. DOI: 10.1104/pp.105.064485

[115] Mochida K, Saisho D, Yoshida T, Sakurai T, Shinozaki K. TriMEDB: a database to integrate transcribed markers and facilitate genetic studies of the tribe Triticeae. BMC Plant Biology. 2008;8:72. DOI: 10.1186/1471-2229-8-72

[116] Yu J, Jung S, Cheng CH, Ficklin SP, Lee T, Zheng P, Jones D, Percy RG, Main D. CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Research. 2014;42:1229–1236. DOI: 10.1093/nar/gkt1064

[117] Mount DM. Hand Book of Bioinformatics: Sequence and Genome Analysis. 2nd ed. New York: Cold Spring Harbor Laboratory Press; 2001.

IntechOpen