6

Bioinformatics Approaches to Improve and Enhance the Understanding of Plant–Microbe Interaction: A Review

Surojit Sen
Mariani College

Sunayana Rathi
Assam Agricultural University

CONTENTS

6.1 Introduction

Increasing the crop yield and productivity is the primary goal of all agricultural activities. In the present scenario of global warming, optimization of plant production system for better yield in areas of limited fertility is targeted. The growth and productivity of plants depend much on the interaction with the microbes present in their immediate environment. Plants share their habitat with complex microbiota that include bacteria, oomycetes, fungi, archaea, and viruses (Agler et al., 2016). The complexity is determined by the shared environment and the biotic and abiotic interactions involved at different levels. The outcome of host–parasite interaction depending on the resources available may be positive, neutral, or negative. Microbes can thus be considered as mutualistic, commensal, or pathogenic and require a well-balanced interaction for the sustainable productivity of plants. Hence, it is very important to understand the interactions involved and modes of control at the molecular level. A knowledge base developed in this direction will help treat and prevent infection and also reduce crop loss.

In host–pathogen relationship prediction, biochemistry-based approaches play a major role, which may be further supported by bioinformatics. Bioinformatics approach may play an important role by utilizing extremely large data sets generated in the post-genomic era. Further, the use of modern techniques such as machine learning and network analysis will provide a better insight into the interaction between host and pathogen and develop new strategies. Generally, two strategies are adopted for the management of host–pathogen interactions. The first one is to reduce or disable the virulence of the pathogen by targeting the machinery it uses. The other one is to target the host machinery so that the host immune system can be boosted to shield it from the pathogen attack. Therefore, it is highly significant to predict the key host–pathogen interactions in order to get the outcome as desired. Whichever may be the strategy, in both

DOI: 10.1201/9781003171416-6

the cases, a thorough understanding of the interaction network is needed and bioinformatics may help a lot to understand the mechanism involved and decipher the knowledge of plant–microbe interactions.

Bioinformatics approaches to study host–pathogen interaction can be broadly divided into two categories: biological and computational. The biological category is inspired by the traditional biological knowledge of structure and homology, while computational methods are data-driven and need high-throughput computational tools such as network analysis and machine learning.

6.2 Genomics

With the improvement of sequencing techniques, modern genomics has produced large amounts of publicly available DNA sequence data and subsequently, a huge amount of data have been produced in other fields of omics too. The development of computational science and internet has helped biologist to submit and archive these data in retrievable databases/repositories. Host–pathogen interaction data are not an exception to that. These data can be easily retrieved from these open-source portals and analyzed to gather knowledge. The metadata developed may help in understanding the host–pathogen interaction more precisely and develop new strategies in this direction.

Novel techniques in genomic field have transformed the identification and detection strategy of host–pathogen interactions. These techniques can also provide new insights to understand their underlying dynamics. The availability of genomic data has helped biologists to study genomic signatures of host–pathogen interactions by searching for the association of single gene to genome-wide scans. The genomic sequences available can be used for phylogenetic and comparative analysis of host and pathogen. By thorough genome scans, mutations causing resistance in host can be easily detected and further comparative population genetic studies of the host can help in presuming the impact of pathogen. It is also seen that comparative sequence analysis of resistant and susceptible host can identify the differences in size orientation and location of the genes involved. Genes involved in known pathosystem can be targeted to study in unknown systems of our interest. In this regard, whole-genome comparisons can help and this has become feasible only due to the gradual reduction in the costs of high-throughput sequencing recently. Since a host's response during infection by a particular pathogen most often involves multiple genes, whole-genome approaches have high potentiality of unfolding polygenic responses (Daub et al., 2013).

In order to understand the underlying genetics of the interactions between hosts and pathogens, genetic variation can be studied at different levels such as within species, across species, within population, or across population. Genotype–phenotype association studies can also be used to understand the genetic architecture more precisely. Hence, genome-wide association study (GWAS) is successfully used in unrevealing the host's responses to pathogen exposure. Although the whole-genome approach has been established as a benchmark for many host–pathogen studies, there are many limitations such as non-availability of reference genome of many non-model organisms and poor annotation of reference genome. This may lead to low rate of discovery of important regions of host genome responding to pathogen.

Combining selection scans with association studies can reveal the differences in infectious disease susceptibilities and identify specific protective genes and alleles. Once the resistant genes/QTLs are identified, they can be introgressed and pyramided by marker-assisted selection or through genetic engineering.

There are many toolkits and repositories of scripted pipelines available for genomic data analysis, such as https://github.com/pditommaso/awesome-pipeline, core R packages (https://cran.r-project.org/web/packages/GenomicTools/index.html), R Bioconductor, R Markdown (https://rmarkdown.rstudio.com), or Jupyter (https://jupyter.org); for graphical user interface-guided data integration analysis, "Galaxy" can be used. Databases such as NCBI, EMBL, DDBJ, and Stanford genomic resource (http://genome-www4.stanford.edu/) also provide many tools for visualization and analysis of genomic data (Table 6.1).

6.3 Gene expression Data

Expressed Sequence Tags: ESTs are obtained from cDNA libraries by partial random sequencing. They are single-read mRNA sequences of approximately 300–500 nucleotides in length. ESTs represent

TABLE 6.1
Bioinformatics Databases/Repositories of Host–Pathogen Interactions

Name	URL	Description
VFDB	http://www.mgc.ac.cn/VFs/main.htm	Virulence factor database of bacterial pathogens.
PATRIC	http://www.patricbrc.org/	Provides integrated data and analysis tools for bacterial infectious diseases.
ViPR	https://www.viprbrc.org/brc/home.spg?decorator=vipr	Virus pathogen database and analysis resource.
Expasy	https://www.expasy.org/	Swiss bioinformatics resource portal.
ViralZone	https://viralzone.expasy.org/	Resource for viral data.
V-pipe	https://www.expasy.org/resources/v-pipe	Bioinformatics pipeline for assessing viral genetic diversity.
HPIDB 3.0	https://hpidb.igbb.msstate.edu/	Host–pathogen interaction database.
GPS-Prot	http://gpsprot.org/	Data Visualization for Protein-Protein Interactions.
PHI-base	http://www.phi-base.org/	Host–pathogen interaction database.
DIP	https://dip.doe-mbi.ucla.edu/dip/Main.cgi	Database of Interacting Proteins.
BioGRID	https://thebiogrid.org/	Database of Protein, Genetic, and Chemical Interactions.
IntAct	https://www.ebi.ac.uk/intact/	Molecular Interaction Database.
PID	http://www.ndexbio.org/#/user/301a91c6-a37b-11e4-bda0-000c29202374	Pathway interaction database.
PHIDIAS	http://www.phidias.us/	Pathogen-Host Interaction Data Integration and Analysis.
PHISTO	https://www.phisto.org/	Pathogen-Host Interaction Search Tool.
HoPaCI-db	http://mips.helmholtz-muenchen.de/HoPaCI	Host–Pathogen Interaction database.
mentha	http://mentha.uniroma2.it/index.php	Interactome database.
MINT	https://mint.bio.uniroma2.it/	Molecular Interaction database.
SIGNOR 2.0	https://signor.uniroma2.it/	Signaling network database
MatrixDB	http://matrixdb.univ-lyon1.fr/	Extracellular matrix proteins, proteoglycans, and polysaccharides interaction database.
IMEx	http://www.imexconsortium.org/	Molecular interaction data.

expressed genes of organs or tissues at a specific developmental stage. An enormous number of ESTs have been produced from thousands of species in the past few years and are available freely in databases such as dbEST of NCBI, DDBJ, and EMBL. In case of non-model organisms, where whole-genome sequencing data are not available, EST data sets are utilized as an alternative for providing valuable resources to develop gene-associated markers such as SSR and SNP.

Microarrays: Microarray is a laboratory technique used to detect the expression profile of thousands of genes at the same instance. Microarray tool can be used to analyze RNA expression profile of both pathogens and hosts by the help of microarray chips to ensure gene expression and identify regulatory mechanisms involved in the pathogenic state. It can assist in hypothesizing functions of uncharacterized resistant genes of host and also in identifying virulence genes that promote colonization or those that cause damage to host tissue. It is also used to identify the genetic polymorphism of specific loci associated with a particular trait. Hence, by this technique, genes involved in pathogenicity can be identified in the study of host–pathogen interactions. This can be achieved by measuring and comparing the gene expression of host cells before and after infection. The gene expression pattern analysis can provide insight into the gene regulatory network for host during all stages of infection. Many microarray studies have been performed in the past decades, leading to accumulation of enormous amount of expression data. The need to store and analyze these data has led to the creation of many new expression databases. Some of these databases/tools of gene expression data are listed in Table 6.2.

RNA-Seq: In contrast to microarrays, genes with low abundance, sequence variation, and even novel transcripts can be easily identified by RNA-Seq. Moreover, since the expression analysis for non-model

TABLE 6.2

Widely Used Databases and Tools for Gene Expression Data

Databases/Tools	Description
GEO (Gene Expression Omnibus) (https://www.ncbi.nlm.nih.gov/geo/)	Database for gene expression profiling and RNA methylation profiling derived from microarray and RNA-Seq experiments.
ArrayExpress (http://www.ebi.ac.uk/arrayexpress/)	Repository of functional genomics data.
AMAD software package	Provides basic microarray data storage and retrieval capabilities.
MGOS database	Contains data obtained from *O. sativa* and *M. grisea.*
OryzaExpress (http://bioinf.mind.meiji.ac.jp/OryzaExpress/)	Gene expression database for rice.
NASCArrays (http://arabidopsis.info/affy)	Repository of microarray data of *Arabidopsis thaliana* with data mining tools.
PathoPlant (http://www.pathoplant.de)	Microarray expression data of co-regulated genes involved in plant defense responses.
PLEXdb (http://www.plexdb.org)	Plant and plant–pathogen microarrays.
OmicsDB::Pathogens (http://pathogens.omicsdb.org)	A database for exploring functional networks of plant pathogens.
PlaD (http://systbio.cau.edu.cn/plad/index.php or http://zzdlab.com/plad/index.php)	Transcriptomics database for plant defense responses to pathogens.
PHI-base (www4.rothamsted.bbsrc.ac.uk/phibase/)	Database for pathogen–host interactions.

organisms can be performed by RNA-Seq, the expensive step of producing species-specific arrays can be avoided. Because of these advantages, recently, RNA-Seq technology has become popular for studying genome-wide expression profile. Entire RNA molecules are sequenced to measure the expression levels of all transcripts in order to harness knowledge of known as well as novel unidentified defense genes of host and effector genes of pathogen. Using the RNA-Seq method, the total transcriptional activity of both the host and pathogen can be studied before and after infection. Data can be analyzed for identifying the differentially expressed genes during infection. Plant–pathogen mixed RNA-Seq databases are available, which can be accessed and analyzed using bioinformatics tools.

More recently, high-throughput RNA sequencing has been developed, which paved the way for capturing all classes of coding and noncoding transcripts in both the pathogen and the host. This technique, called dual RNA-Seq technique, not only allows understanding the physiological changes in pathogen and host during infection, but also reveals hidden molecular phenotypes of virulence associated with small noncoding RNAs that were not visible in standard assays. The assay pipeline involves the following steps: RNA extraction → rRNA depletion → deep sequencing → parallel read mapping with host and pathogen genome → cross-mapping → aligned reads → normalization.

The normalized reads are then subjected to downstream analyses such as quantification, differential expression, pathway analysis, and network inference.

Differential expression analysis is generally done using popular tools such as edgeR, DESeq2, and limma/voom, available through Bioconductor of R statistical programming language. Various algorithms are also available for this purpose of downstream analysis, among which pipelines such as Tuxedo suite are standard.

The list of genes (both pathogen and host) produced as a result of differential expression analysis can further be interpreted in terms of gene function to hypothesize new tests. Databases such as Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) provide software suites for this purpose. Further, the metabolic network can be reconstructed by specialized knowledge bases such as BioCyc. Tools are also available for the reconstruction of molecular signature and gene set enrichment analysis from RNA-Seq data. The link between the identified genes can be inferred by network analysis called network inference (NI), and global regulatory networks can be constructed from the expression data.

MicroRNAs: MicroRNAs (miRNAs) are small noncoding RNAs that are endogenously found in organisms for the regulation of gene expression. They are derived from single-stranded RNA precursors that can form stem-loop structures. Generally, they are 18–24 nucleotides long and depending on the extent of base-pairing with the target mRNAs, miRNAs can silence gene expression. It is found that host miRNAs target pathogen virulence genes, while pathogen's miRNAs target plant resistance genes. Thus, miRNA can mediate trans-kingdom gene regulation and can play a great role in host–pathogen interactions.

Experimentally, miRNAs are discovered by cloning (Long and Chen, 2009), microarray screening (Barad et al., 2004), *in situ* hybridization (Yao et al., 2012), or next-generation sequencing of small RNAs (Landgraf et al., 2007), while computationally, miRNAs and their targets are identified by *in silico* genomic or EST sequence analysis. miRBase (http://www.mirbase.org/) is a searchable database for published miRNA sequences and annotation. Several computational prediction downloadable programs are available, such as miRPlant (http://sourceforge.net/projects/mirplant/), miRNA EMBL (http://www.russelllab.org/miRNAs/), MIREAP (https://sourceforge.net/projects/mireap), miRA (https://github.com/mhuttner/miRA), C-mii (http://www.biotec.or.th/isl/c-mii), and Web servers such as microHARVESTER, miRU, DIANA Tools, miRanda, and EIMMo are also used. For target prediction, tools such as psRNATarget (http://plantgrn.noble.org/psRNATarget/) (Dai and Zhao, 2011), TAPIR (Bonnet et al., 2010) (http://bioinformatics.psb.ugent.be/webtools/tapir/), TargetScan (http://www.targetscan.org/vert_ 72/), and miRTour Web server are frequently used. miRDB is an online database for miRNA target prediction and functional annotation. Apart from these, various R packages are available at https://bioconductor.org for miRNA prediction.

The predicted miRNA can further be validated using quantitative reverse transcriptase PCR (qRT-PCR).

6.4 Protein–Protein Interaction (PPI) Prediction

Eukaryotic cells have thousands of gene products in their proteome, which undergoes complex interactions throughout life, forming functional pathways to provide signals from outside the cell and a proper cellular response to the signals. Proteins are the workhorses of host–pathogen interaction network too.

Protein–protein interaction is the most prominent way how a pathogen interacts with its host. Proteins are a sequence of amino acids bonded by peptide to form a string called primary structure. The sequence of amino acids in the primary structure determines the structure as well as the function of the protein. Local folding of the primary structure caused by interaction between the side chains of amino acids results in the secondary structure (alpha-helix and beta-sheet). Alpha-helices are responsible for structure and membrane spanning domains, while beta-sheets provide the docking site for enzymatic reactions. In tertiary structure, further folding of beta-sheets and alpha-helices occurs to form a complex three-dimensional entity. This structural entity is anchored by ionic interactions, disulfide bridges, hydrophobic interactions, and van der Waals forces. Even after folding, a number of posttranslational modifications such as cleavage, phosphorylation, glycosylation, or ubiquitination take place. After all these complex modifications, protein attains a final shape and is ready for interactions. Hence, the determination of three-dimensional structure of proteins is of utmost importance in the study of protein–protein interaction between pathogen and host. But unfortunately, the determination of 3D structure of a protein is difficult and time-consuming. Traditionally, X-ray crystallography/NMR is used for 3D structure determination, but unfortunately, many proteins get distorted during crystallization. Till date, only a small fraction of 3D structures of host/pathogen proteome has been determined.

The more the primary sequence similarity, the more the chance of an interaction among the proteins (interologs). Protocols have been developed to map known sequences of interaction interface onto pairs of sequences (homologous or orthologous) in different organisms. At least 80% sequence similarity is required for this purpose, and hence, correct determination of PPI decreases as the evolutionary distance increases. Such homologous proteins with primary sequence similarity are searched for in the pathogen/or host, which may also interact with known annotated proteins (interologs). Interolog mapping has a high false-positive hit rate; hence, in order to improve the quality of the mapping results, further filtering based on cellular localization, biological functions, and temporal expression profile is required to significantly identify potential host–pathogen protein–protein interactions.

Homology-Based Prediction: Similar sequences usually have similar functions; homology-based prediction methods work based on this assumption. This has been found true in case of a large degree of similarity or evolutionary conservation of proteins under investigation, which are quite abundant. It is an anticipation that protein–protein interaction would be conserved across related species. In this prediction method, the genomic data of hosts and pathogens are analyzed to identify proteins homologous with the known interacting protein. These are then compared for the determination of likelihood of any protein–protein interactions occurring between the host and the pathogen. In this approach, interaction templates of host and pathogen genomic sequences are considered to find out the probable sets of PPIs. In order to filter out non-homologous sets, a homology detection algorithm is applied to these PPIs. Then, further filtration is done to the newly obtained sets through stage-specific and tissue-specific expression data of pathogen and host. Filtering is also done with the help of predicted localized data. Homology-based approaches are widely used for the prediction of host–pathogen PPI as this approach is considered to be simple and having a well-supported biological background. Simple data such as template PPIs and protein sequences are required for the purpose of prediction, and hence, they can be adapted and applied to other multiple host–pathogen systems.

A major drawback of homology-based prediction is the detection of high rate of false positives (Mariano and Wuchty, 2017). Further, protein pairs predicted by these *in silico* approaches may have differential temporal and spatial expressions and hence may rarely have the chance to interact. Successful homology-based prediction approaches, therefore, require filters that account for these criterions. The use of random forest classifiers can help in this regard (Figure 6.1).

Structure-Based Prediction: It is generally believed that when a pair of protein structures is similar to known interacting pair of proteins, it is more likely that they will be interacting in a similar pattern. The structural information of proteins can be used extensively for the prediction of host–pathogen interactions computationally by comparing with already known interactions with other proteins. In this approach, the host and pathogen genomes are first scanned for structural similarity with already known protein complexes to find out probable interactions using the structural similarity. The result is finally filtered by expression data of tissue-specific host proteins and stage-specific pathogen proteins. The set of proteins hence identified have a high interaction probability.

There are number of tools available for predicting PPI based on 3D structures of the interacting proteins, such as docking and MD (molecular dynamics) simulation. But unfortunately, till date only a small fraction of 3D structures of host/pathogen proteome has been determined. Hence, there is little focus on the prediction of host–pathogen PPIs through this technique. However, bioinformatics tools also provide an alternative to solve the 3D structures of the proteins whose X-ray structures are not available.

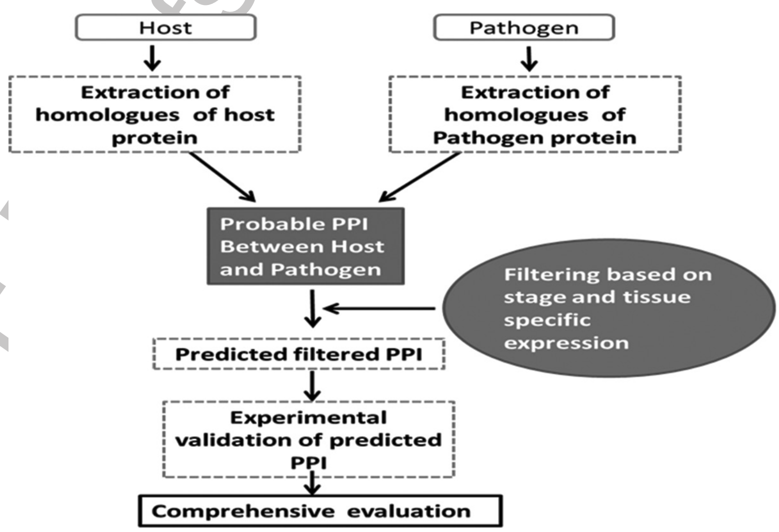

FIGURE 6.1 Predictions of host–pathogen interactions using homology-based approach.

Protein structure prediction is usually done by homology modeling. This method attempts to create a new structure based on a set of known structures with sequence similarity. In addition, molecular dynamics approach uses Newtonian mechanics to simulate the atom-scale interactions. There are reports that successfully used these approaches to provide insights into host–pathogen interactions or plants' defense mechanisms (Sarma et al., 2012; Dehury et al., 2013; Dehury et al., 2015).

For interspecies PPI prediction, the 3D structural homology is identified by scanning host and pathogen genomes. Proteins having similarity to known protein complexes are assessed for the detection of putative interactions using structural information. The left out interactions are filtered based on the biological context for several pathogens. This strategy was first adopted by Davis et al. (2007).

Although prediction based on structure is a powerful tool, it seems that pathogens, in order to achieve binding stability, evolve their protein interfaces without sequence or structural similarity to native interacting proteins. Pathogen protein interfaces sometimes overlap with and even compete with or mimic the endogenous host protein interfaces (Figure 6.2).

Domain-Based Approaches: A protein domain is the conserved region of a protein's three-dimensional structure that is responsible for a specific biological function. Domains are created while the proteins get folded in nature, and they evolve independently (Hleap and Blouin, 2016). Domains are the regions of contact during PPIs, and hence, domain–domain interactions (DDIs) drive PPIs considerably. A number of studies have been carried out based on known intra-species DDIs for the prediction of HP-PPIs. Protein–domain association studies are highly predictive when machine learning algorithms such as support vector machine (SVM) and random forest (RF) are used (Barman et al., 2014). Based on the information of primary sequence of interacting domains, the large-scale detection of hypothetical interactions between proteins has been possible. In a study of human protein interaction network, Dyer and his associates combined DDIs with protein sequence k-mers and topological properties of protein using support vector machine algorithm to predict host–pathogen interactions (Dyer et al., 2011). Domain-based prediction method is helpful in the identification of common functionality features, which allow pathogens to interact with multiple hosts.

Motif- and Integration-Based Approaches: Motifs are small recognizable regions of protein having unique biological functions. Proteins interact through a reusable set of binding motifs with their partners. Motifs complement each other among the partners despite slight difference between individual proteins. Motifs always show a conserved pattern, and hence during analysis, once these patterns are recognized and validated, the remaining part of data can be discarded for reducing the computational cost. Biologists have exploited this in various studies involving protein–protein interaction networks and identification of transcription factor binding sites (Das and Dai, 2007), prediction of secondary structure

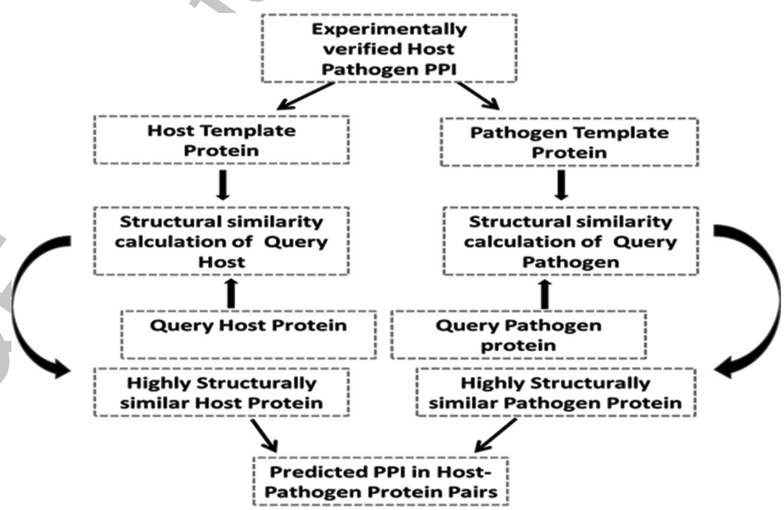

FIGURE 6.2 Predictions of host–pathogen interactions using structure-based approach.

(Bi et al., 2008) as well as enzymatic activity studies and identification of functional residues (Hulo et al., 2008). In all of these studies, a known data set is analyzed to detect statistical over-representation of patterns. These patterns are then applied to new proteins to draw inferences.

Motif identification can be carried out by various tools available in the open-source portals. Tools such as miniMotif and PSI-BLAST look for an over-represented set of amino acids for the identification of these sites. These *de novo* motif algorithms compare a set of proteins with a known function with a set of proteins without the desired function. However, these *de novo* algorithms are not fit for instances where the function is already well studied. In such cases, databases such as PROSITE, eukaryotic linear motif (ELM), and PFAM can be used to curate a large collection of linear sequence motifs. These functional motifs can be used to predict protein-protein interactions of pathogen and host.

Motif–Domain and Motif–Motif Interaction-Based Approaches: These approaches can also be used as foundations for host–pathogen PPI prediction and have gained importance recently. Motifs of one protein sometimes interact with domains or even motifs of another protein. This has been studied in HIV–human interactions by Evans and his coworkers (2009). They generated an HIV-1 and human interactome with the help of annotated ELMs in HIV-1 proteins that interacted with counterpart human protein domains. Integration of primary and secondary sequence information enhances *in silico* host–pathogen PPI predictions. However, other auxiliary data can also be used to reduce the impact of false positives. Currently, assimilation of features such as domain information, sequence features, ELM data, GO features, graph topological properties, and gene co-expression data are used to train the classifiers. This strategy was successfully used by Coelho and his coworkers to predict the human oral microbial interactome by incorporating domain information, protein sequence features, and GO annotations (Coelho et al., 2014).

Surface Electrostatics and Epitope Prediction: Epitopes are the antigenic determinants of pathogens and can be recognized by the host immune system. Interacting protein surfaces show electrostatic as well as non-covalent interactions. Antibodies, which bind to the epitopes, also show a number of electrostatic and non-covalent interactions. These interactions take place either through backbone carbon or through side chain carbon. This allows a number of host proteins to recognize pathogen antigens by shared physicochemical properties. Several computational protein interaction identification tools have already integrated these electrostatic attributes, thereby enabling epitope prediction on the basis of surface energetics.

Analysis of Dynamic Character of PPI: Three-dimensional structures obtained from X-ray crystallographic method are cumbersome, time-consuming, and expensive, and also many proteins' structure gets distorted during crystallization. So, recent studies have employed NMR to complement the static crystallographic data with dynamic functional data. This allows the proteins to be studied in their natural state, the way they actually fold in nature. Solution-state NMR can determine the interaction between pathogen protein and the domains of host protein by emphasizing how surface charge distribution, intrinsic disorder, and mimicry of host protein help in specific binding. NMR united with molecular dynamics simulations can enhance the prediction process if a preexisting structure is available. A blend of cryo-electron microscopy, MDs, and solid-state NMR can further help in building a model on how interaction takes place in natural environment.

6.5 Machine Learning-Based Predictions

Prediction methods based on machine learning are widely used in host–pathogen interactions. Figure 6.3 illustrates some of the machine learning methods that are being widely used to study the host–pathogen interactions.

Supervised learning has been used for the successful prediction of PPIs in the host–pathogen domain by considering more than 35 features of host and pathogen. The features considered can be sequence similarity, gene expression profiles, similarity in post-translational modifications, GO, tissue distribution, and various other features of host and pathogen interactome. After the initial analysis, top three or top six features of utmost importance are selected so that the data can be classified into interacting and

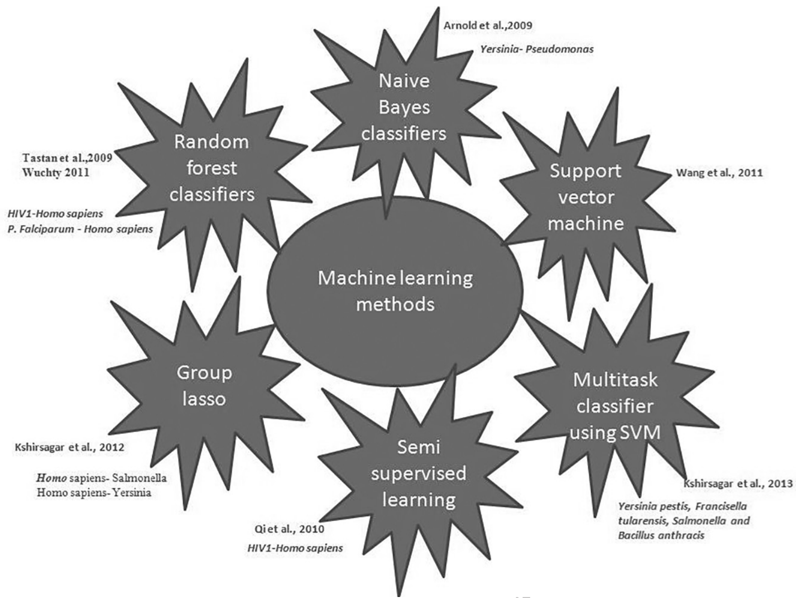

FIGURE 6.3 Various machine learning methods.

non-interacting classes. In most of the cases, supervised learnings exploit RF classifiers for these kinds of classifications.

Naive Bayes algorithm is used for the classification of training samples based on similarity. The similarity here is measured with the help of Smith–Waterman local alignment algorithm. Input features such as amino acid composition, amino acid frequencies, and amino acid properties are used, and finally, most important features are strategized using feature selection strategy. In some cases, features derived from the secondary structure are also used. PSIPRED software is used for structure prediction, and from the predicted structure, features are selected for input vector.

Machine learning approaches such as SVM algorithms are trained by carefully picking positive and negative training sets of protein interactions. All approaches of prediction by supervised machine learning need appropriate training for both positive and negative sets with sequence or higher-order information to robustly classify interacting proteins between host and pathogen. Inclusion of non-interacting set in the training data influences the accuracy of identification of interacting pairs from the non-interacting ones. In order to create a negative training set, highly dissimilar protein sequences from other organisms are selected. Dissimilarity here is compared with the interacting proteins of the pathogen in question. SVM with these training sets increases the prediction accuracy.

Multi-task classification frameworks can be used for establishing relationship between host and multiple pathogens. Based on the similarity of infection initiated by different pathogens, this machine learning technique is used for the classification of PPI into interacting and non-interacting classes. The classification is based upon the hypothesis that similar pathogen targets the same critical biological process of the host.

In semi-supervised multi-task method of prediction, the data set of host is processed through both supervised and semi-supervised learning. The supervised classifier works on labeled PPI data and trains the semi-supervised classifier with partially labeled PPIs. The supervised classifier shares network layers with the semi-supervised classifier. This entire framework is used to improve the prediction of interacting pairs.

Group lasso technique, on the other hand, can be used to improve the supervised learning-based prediction. In this technique, the missing data set values are replaced by the values generated from cross-species information. This has been successful in increasing the prediction accuracy by more than 70%.

9780367774424_C006.indd 109 05/04/22 11:40 PM

6.6 Systems Biology Approach

Systems biology is a holistic approach to understanding the complex biological systems using mathematical modeling and analysis of high-throughput data. It focuses on single- or multi-level computational analysis and modeling of experimental data resulting from new hypotheses. It can be approached in two ways: bottom-up and top-down. In the bottom-up approach, sub-models are built and later integrated to find out the integration of cell components, which is followed by building of *in silico* models comprising all pathways of cell–pathogen or host–pathogen interaction.

In the top-down approach, a genome-wide analysis is performed with the help of data obtained from omics technologies (such as genomics, transcriptomics, proteomics, and metabolomics).

Identification of key molecules and their interaction is carried out in the following three steps:

1. **Identification of Key Molecules (Biomarkers)**: First, biomarkers such as DNA/RNA sequences, proteins, mutations (SNPs), transcripts of coding region (microarray or RNA-Seq data of differentially expressed genes), noncoding transcripts (miRNAs and piRNAs), or metabolites are identified. Sometimes machine learning approach is used for the prioritization of key biomarkers.
2. **Network Modeling of Regulatory Interactions:** The next step is the systems-level understanding of the molecular mechanisms of all the involved biological processes by means of mathematical modeling. The network is generally represented by nodes which denote the molecules such as proteins, DNA, RNA, or metabolites and edges representing interactions between the nodes. Based on the prior knowledge of omics data, interaction networks such as gene regulatory networks or even genome-wide networks can be generated and inferred.
3. **Identification of Disease Modules:** Then the group of molecules and interactions among them, which are linked to a phenotype of interest, is identified. The next step is the integrated analysis of interaction networks for the discovery of disease-associated modules. This integrated study can reveal disease modules with partially overlapping molecular mechanism. Proteins and their degree of overlap correlate biological similarity or disease symptoms. This can be used successfully for discovering the affected mechanism.

Systems biology and computational modeling can be employed to metabolic engineering to anticipate the effect of genetic engineering on metabolism. Recently, constraint-based modeling (examining the function of metabolic networks by relying on physicochemical constraints) has gained popularity and has been proven successful for large-scale microbial networks (Price et al., 2003). Once the gene network

TABLE 6.3

List of Tools Related to Systems Biology

Name	Description
Cytoscape	Data integration, network visualization, and analysis.
MEGA	Phylogenetic analysis and creation of dendrograms.
GenMAPP	Visualization and analysis of genomic data in the context of pathways
BioTapestry	Interactive tool for building, visualizing, and simulating genetic regulatory networks.
PathVisio	Tool for displaying and editing of biological pathways.
PathView	Pathway-based data integration and visualization.
InCroMap	Integration of omics data and experimental data and their joint visualization in pathways.
CellDesigner	Diagram editor for gene regulatory networks.
Complex Pathway Simulator (COPASI)	Simulation and analysis of biochemical networks.
SBMLToolbox	Analysis of SBML model in MATLAB.

or metabolic network are identified or disease-associated modules are modeled, gene editing techniques can be used to control plant–pathogen interactions to obtain customized plants with enhanced yield. Highly efficient gene editing tools such as zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and clustered regularly interspaced short palindromic repeats (CRISPRs) can help in achieving such a goal (Table 6.3).

6.7 Conclusions

Bioinformatics has played and will continue to play a significant role in exploiting the data available for exploring the host–pathogen interaction and enhancing our knowledge in this field. All the processes of prediction mentioned here has their own advantages and disadvantages. Hence, the use of these tools and selecting the right one needs deeper exploration. Further, the non-availability of reference genome of non-model organism, proper annotation, and curation are the major challenges. Since the results and inference drawn depend much on the quality of the input data, these challenges are to be addressed properly. Although *in silico* techniques help to speed up the prediction process, time-to-time validation is required to conclusively decide on the causes and consequences of host–pathogen interactions and to combat them.

REFERENCES

Agler et al. (2016). Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biol*. **14**(1): e1002352.

Arnold et al. (2009). Sequence based prediction of type III secreted proteins. *PLoS Pathogen*. **5**(4):e1000376.

Barad et al. (2004). MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. *Genome Res*. **14**: 2486–2494.

Barman et al. (2014). Prediction of interactions between viral and host proteins using supervised machine learning methods. *PLoS One*. **9**: e112034.

Bi et al. (2008). A comparative study on computational two-block motif detection: algorithms and applications. *Mol. Pharm*. **5**(1): 3–16.

Bonnet et al. (2010). TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics*. **26**: 1566–1568.

Carvalho Leite et al. (2017). Computational prediction of host-pathogen interactions through omics data analysis and machine learning. In: Rojas, I., Ortuño, F. (Eds.), *Bioinformatics and Biomedica Engineering*. Springer International Publishing, 5th International Work-Conference, IWBBIO 2017, Granada, April 26–28, 2017, Proceedings, Part II.

Coelho et al. (2014). Computational prediction of the human– microbial oral interactome. *BMC Syst. Biol*. **8**: 24.

Dai X and Zhao PX. (2011). psRNATarget: a plant small RNA target analysis server. *Nucl Acids Res*. **39**: W155–159.

Das M K and Dai H K. (2007). A survey of DNA motif finding algorithms. *BMC Bioinformatics*. **8** (Suppl 7): S21.

Daub et al. (2013). Evidence for polygenic adaptation to pathogens in the human genome. *Mol Biol Evol*. **30**: 1544–1558.

Davis et al. (2007). Host-pathogen protein interactions predicted by comparative modelling. *Protein Sci*. **16**: 2585–2596.

Dehury et al. (2013). Structural analysis and molecular dynamics simulations of novel δ-endotoxin Cry1Id from Bacillus thuringiensis to pave the way for development of novel fusion proteins against insect pests of crops. *J Mol Model*. **19**: 5301–5316.

Dehury et al. (2015). Molecular recognition of avirulence protein (avrxa5) by eukaryotic transcription factor xa5 of rice (Oryza sativa L.): insights from molecular dynamics simulations. *J Mol Graph Model*. **57**: 49–61.

Dix et al. (2016). Use of systems biology to decipher host-pathogen interaction networks and predict biomarkers. *Clin Microbiol Infect*. **22**(7): 600–606.

Dyer et al. (2011). Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect Genet Evol*. **11**: 917–923.

Esvelt KM and Wang HH. (2013). Genome-scale engineering for systems and synthetic biology. *Mol Syst Biol*. **9**: 641.

Evans et al. (2009). Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med Genom*. **2**: 27.

Hleap J and Blouin C. (2016). The semantics of the modular architecture of protein structures. *Curr Prot Pept Sci*. **17**: 62–71.

Hulo N et al. (2008). The 20 years of PROSITE. *Nucl Acids Res*. **36**: D245.

Kshirsagar et al. (2012). Techniques to cope with missing data in host–pathogen protein interaction prediction. *Bioinformatics*. **28**(18): i466–i472.

Kshirsagar et al. (2013). Multitask learning for host–pathogen protein interactions. *Bioinformatics*. **29**(13): i217–i226.

Landgraf et al. (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*. **129**: 1401–1414.

Long J-EE and Chen H-XX. (2009). Identification and characteristics of cattle microRNAs by homology searching and small RNA cloning. *Biochem Genet*. **47**: 329–343.

Mariano R and Wuchty S. (2017). Structure-based prediction of host pathogen protein interactions. *Curr Opin Struct Biol*. **44**: 119–124.

Mathur et al. (2020). MicroRNA-mediated trans-kingdom gene regulation in fungi and their host plants. *Genomics*. **112**(5): 3021–3035.

Price et al. (2003). Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol*. **21**: 162–169.

Qi et al. (2010). Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics*. **26**: i645–i652.

Sarma et al. (2012). A comparative proteomic approach to analyse structure, function and evolution of rice chitinases: a step towards increasing plant fungal resistance. *J Mol Model*. **18**: 4761–4780.

Tastan et al. (2009). Prediction of interactions between HIV-1 and human proteins by information integration. *Pac Symp Biocomput*. **51**: 6–527.

Wang et al. (2011). High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics*. **27**(6): 777–784.

Westermann et al. (2017). Resolving host-pathogen interactions by dual RNA-seq. *PLoS Pathog*. **13**(2): e1006033.

Wuchty S. (2011). Computational prediction of host–parasite protein interactions between P. falciparum and H. sapiens. *PLoS One*. **6**: e26960.

Yao et al. (2012). In situ detection of mature miRNAs in plants using LNA modified DNA probes. *Methods Mol Biol*. **883**: 143–154.

Zhang Y. (2005). miRU: an automated plant miRNA target prediction server. *Nucl Acids Res*. **33**: W701–W704.