EUROPEAN JOURNAL OF BIOMEDICAL AND PHARMACEUTICAL SCIENCES
http://www.ejbps.com

ISSN 2349-8870
Volume: 10
Issue: 2
338-356
Year: 2023

IN SILICO DRUG TARGETING AGAINST *EREMOTHECIUM GOSSYPII* USING PLANT-BASED BIOACTIVE COMPOUNDS

Chaitanya A. Borkar and Rupesh S. Badere*

Department of Botany, MJP Educational Campus, RTM Nagpur University, Nagpur- 440033.

***Corresponding Author: Rupesh S. Badere**
Department of Botany, MJP Educational Campus, RTM Nagpur University, Nagpur- 440033.

Article Received on 10/12/2022 **Article Revised on 31/12/2022** **Article Accepted on 20/01/2023**

ABSTRACT
We report the findings of drug targeting studies carried-out to identify the plant-derived molecule against the *Eremothecium gossypii* responsible for stigmatomycosis in cotton. Initially, we predicted the essential genes in *E. gossypii* by training the machine classifier employing the data from *S. cerevisiae* using seventeen genome features. Later, we selected the essential genes, whose products localize in the plasma membrane, for molecular docking of plant-derived drugs *viz.*, azadirachtin, cleistanthin A, cleistanthin B, daturine, embelin and nicotine. The studies showed maximum interaction between cleistanthin B and plasma membrane ATPase of the pathogen.

KEYWORDS: BLAST, essential genes, machine learning, molecular docking.

1. INTRODUCTION
Diseases and pest attack on crops cause a considerable amount of economic loss every year. *Eremothecium gossypii* (Syn. *Ashbya gossypii*) causes a disease in cotton variously termed as 'stigmatomycosis', 'internal boll rot', 'cotton staining'.[1] The disease causes staining and rot of cotton bolls in the infected plant.[2] African cotton-growing regions have suffered tremendous economic loss because of this fungus. The gravity of stigmatomycosis is clear from the fact that there was a time when the prevalence of this disease made it virtually impossible to grow cotton in some parts of the world. In present times, in the absence of any specific fungicide, the disease is controlled through the use of insecticides targeting the vector–Hemipteran insect.[3]

Insensitivity to the general methods of disease control, such as in the case of *E. gossypii*, necessitates the application of specific management techniques. Administration of chemicals specifically affecting the target pathogen is one such way. Drug targeting identifies the putative molecule(s), among thousands, having the activity against pathogen. Later, the bioactivity of the short-listed molecules can be experimentally confirmed. However, mostly such screening process demands unreasonable time and thus are impractical. *In silico* tools are handy in such situations, which are applied in various fields of biology, including virtual drug targeting.[4] Virtual drug targeting can decrease the time required to search for a molecule active in disease management of the crop.

Drug targeting involves a ligand (drug) and a receptor (an enzyme/protein of pathogen). Therefore, the precise selection of the ligand and the target protein/enzyme is required to achieve the effective hit rate. The preference for natural molecules is because of their broad biological activities, low impact on environment and safety of the non-target organisms.[5] This article focuses on identifying a natural bioactive molecule as a candidate drug against the *E. gossypii*. We selected six such plant-derived molecules for drug targeting in the pathogen. The candidate targets for the selected drugs in the pathogen were chosen based on the two criteria. First, the target should play a vital role in the survival of *E. gossypii* and second, it should be localized on the cell membrane for ease in the drug targeting. Considering these points, the essential genes in pathogen and their products localized on the cell membrane were identified using the machine learning approach.[6] Subsequently, molecular docking identified the most effective putative drug for the management of *E. gossypii* in cotton.

2. MATERIAL AND METHODS
2.1 Prediction of essential genes in *E. gossypii* using machine learning
The essentiality of over 95% of the genes in *Saccharomyces cerevisiae* has been experimentally proved[6]. Fortunately, 94% of the 4,776 annotated protein-coding genes of *E. gossypii* are homologous to *S. cerevisiae*.[7] Therefore, the known essential genes in *S. cerevisiae* were used to train the machine classifier to predict the essential genes in *E. gossypii*.

2.1.1: Assessing the fungal genome

The list of 3,500 ORFs of genes with known essentiality in *S. cerevisiae* was downloaded from http://www.gersteinlab.org/proj/predess/. The mRNA sequence data consisting of 4,776 genes of *E. gossypii* was downloaded from the NCBI database under the genome section.

2.1.2: Features related to essentiality

Seventeen genome features were used to train the machine learning classifier about the essentiality (Table 1). Out of these, 14 features (except network topology features) for *S. cerevisiae* were used from Michael Seringhaus's and Alberto Paccanaro's research work available at http://www.gersteinlab.org/proj/predess/. Rest of the 3 features and all the 17 features of *E. gossypii* were obtained using various *in silico* tools mentioned in Table 1. A nomogram was prepared using Logistic Regression classifier in Orange software to assess the relative importance of each feature to essentiality.

2.1.3: Training the machine classifier for essentiality

Three separate machine learning simulators viz., Orange[8], Rapidminer[9] and Weka[10] were used to predict the essential genes in *E. gossypii*. Within these simulators, various classifiers were assigned the job of prediction of essential genes (Table 2).

2.1.4: Cross validation of training dataset

The training dataset was tenfold cross-validated separately in all the three simulators to select the better performing classifier algorithm to predict the essentiality of the individual gene in *E. gossypii*. This dataset was cross-validated with varying number of features to cross-check if the cumulative effect of all the features increases the predictability of essential genes. The cross-validation was performed for the training dataset comprising of 8 (except localization and network topology features), 14 (except network topology features) and 17 features separately for a comparative analysis. Subsequently, the cross-validation results were analyzed by the comparative study of the confusion matrix and receiver operating characteristic (ROC) curve.

Table 1: Seventeen genome features used to train the machine classifier about essentiality of the gene.

SN	Feature	Category	Annotation	Description	Type	Source	Reference
1.	Sequence intrinsic	General	GC	GC content	Real	CODONW	6
2.		General	AA	Length of putative amino acids in a protein	Integer	CODONW	
3.		General	HYDRO	Hydrophobicity score	Real	CODONW	6, 29
4.	Sequence Derived	Codon adaptation	CAI	Codon adaptation index	Real	CODONW	6, 30
5.		Codon adaptation	NC	The effective number of codons	Real	CODONW	6, 31
6.		Close to stop codon	CLOSESTOP	Percentage of codons one third base away from a stop codon	Real	BIOEDIT + EXCEL	6, 32
7.		Rare amino acid	RAREAA	Percentage of rare amino acids in translated ORF	Real	BIOEDIT + EXCEL	
8.		Localization	HELIX	Number of predicted transmembrane helices	Integer	TMHMM	6
9.		Localization	MT	Predicted subcellular localization: Mitochondria	Binary	LOCTREE 3	
10.		Localization	CYTO	Predicted subcellular localization: Cytoplasm	Binary	LOCTREE 3	
11.		Localization	ER	Predicted subcellular localization: Endoplasmic reticulum	Binary	LOCTREE 3	
12.		Localization	NUCLEUS	Predicted subcellular localization: Nucleus	Binary	LOCTREE 3	
13.		Localization	VACUOLE	Predicted subcellular localization: Vacuole	Binary	LOCTREE 3	
14.		Localization	OTHER	Predicted subcellular localization: Any other compartment	Binary	LOCTREE 3	
15.	Experimental genome data	Protein-protein interaction	BC	Betweenness centrality (Network topology feature)	Real	STRING + CYTOSCAPE	33, 34
16.		Protein-protein interaction	CC	Closeness centrality (Network topology feature)	Real	STRING + CYTOSCAPE	
17.		Protein-protein interaction	DC	Degree centrality (Network topology feature)	Integer	STRING + CYTOSCAPE	

Table 2. The list of machine learning simulators and the classifier algorithms used to determine the essentiality of the gene.

SN	Simulator	Classifier algorithms	Reference
1	Orange	Naïve Bayes, Neural Network, Random Forest, Logistic Regression, Adaboost, SVM, Stochastic Gradient Descent, Tree and kNN	8
2	Rapidminer	Naïve Bayes, Decision tree, Naïve Bayes (Kernel), k-NN, Decision stump, and Random Forest	9
3	Weka	Naïve Bayes, Logistic regression, Adaboost, J48 tree, Random Forest Random Tree	10

2.1.5: Preparation of special training dataset

A special training dataset was used to improve the predictability of the training dataset (*S. cerevisiae*) and minimize the occurrence of false-positive results. The special training dataset was prepared using Orange simulator from the data of confusion matrices of the selected classifier. The output was exported to MS-Excel 2013, and the entries predicted as false-positive and false-negative by more than one algorithm were removed to get the special training dataset. This special training data was again fed to the Orange simulator to cross-validate the performance of the classifier. Lastly, the cross-validation results of the normal and special training datasets were compared with respect to the confusion matrix.

2.1.6: Prediction of essential genes in *E. gossypii*

All the three simulators were used to predict the essential genes in *E. gossypii*. The machine learning classifiers trained on *S. cerevisiae* special training dataset were applied to the dataset of *E. gossypii* comprising of features arranged in the same order.

2.1.7: Confirmation of predictions

The data between all the algorithms as-well-as between all the simulators was sorted using MS-Excel 2013 to confirm the predicted results. First, a final average value, indicating the proportion of essentiality and non-essentiality, was calculated for each simulator. This average value was derived by calculating the arithmetic mean of the value of predictions of the entire algorithm. Thus, the genes were assigned essentiality based on these average values. Later, the genes assigned by each simulator as 'essential' were placed together in a tabular format. The genes predicted as 'essential' by more than one simulator were selected for further studies.

2.1.8: Selection of essential genes coding for the products localized in plasma membrane

Among the predicted essential genes, the genes whose product localizes in the plasma membrane were selected from the dataset of *E. gossypii* prepared for prediction. The localization of the product of selected genes on the plasma membrane was confirmed by ascertaining the localization and function of the homologous protein in *S. cerevisiae*.

2.1.9: BLAST search with host *Gossypium hirsutum*

The homologs of the proteins encoded by the selected genes (i.e., the gene whose product localizes in the plasma membrane) were searched in the genome of host i.e., *G. hirsutum* using BLAST server of NCBI database to avoid the unintended interaction of the ligand with the host protein(s). The 'program selection' option was set to "Somewhat similar sequences (blastn)" during the search.

2.2 Molecular docking

2.2.1: Protein modelling

The molecular docking studies require the 3D conformation data of the receptor and ligand. The experimentally verified 3D conformer data was not available online in RCSB Protein Data Bank (PDB). Therefore, the sequence of protein was downloaded from the UniProt database and submitted to the protein modelling server 'The Protein Model Portal'.[11]

2.2.2: Preparation of the ligand

Six plant-based compounds were selected for docking studies *viz.*, azadirachtin, cleistanthin A, cleistanthin B, daturine, embelin and nicotine (Table 3). The 3D conformation of the selected ligand (except azadirachtin) was downloaded from the PubChem Database in '.sdf' format. The .sdf format was changed to .pdb format using the Openbabel software.[12] The PubChem database had the data of 2D conformation of azadirachtin, which was downloaded. Later, using the 2D conformer image, the 3D structure of azadirachtin was derived using Avogadro software.[13]

2.2.3: Docking performance

Molecular docking was performed using Autodock Vina software[14] to study the protein-ligand interactions. The results of docking were visualized using the PyMOL Molecular Graphic System, Version 2.0 Schrodinger, LLC.

2.2.4: Force-field application to ligands

The predicted docking score were re-analyzed by applying force-field (MMFF94) to all the ligands using Avagadro software.[13] The 'auto-optimization' tool was used to apply the force-field MMFF94, and the results were analyzed by performing two-way ANOVA in R software[15] using 'protein model' and 'ligand' as factors.

Table 3. The list of plant-based compounds used for molecular docking.

SN	Common Name (Plant source)	IUPAC name	Molecular formula
1	Azadirachtin (*Azadirachta indica*)	Dimethyl (1*S*,4*S*,5*R*,6*S*,7*S*,8*R*,11*S*,1 2*R*,14*S*,15*R*)-12acetyloxy-4,7dihydroxy-6-[(1*S*,2*S*,6*S*,8*S*,9*R*,11*S*)-2hydroxy-11-methyl-5,7,10trioxatetracyclo[6.3.1.$0^{2,6}$.$0^{9,11}$]dodec-3-en-9-yl]6-methyl-14-[(*E*)-2methylbut-2-enoyl]oxy-3,9dioxatetracyclo[6.6.1.$0^{1,5}$. $0^{11,15}$]pentadecane-4,11dicarboxylate	$C_{35}H_{44}O_{16}$
2	Cleistanthin A (*Cleistanthus collinus*)	9-(1,3-benzodioxol-5yl)-4-[(3*R*,4*R*,5*R*)-3hydroxy-4,5dimethoxyoxan-2yl]oxy-6,7-dimethoxy-3*H*benzo[f][2]benzofuran-1-one	$C_{28}H_{28}O_{11}$
3	Cleistanthin B (*Cleistanthus collinus*)	9-(1,3-benzodioxol-5yl)-6,7-dimethoxy-4[(2*S*,3*R*,4*S*,5*S*,6*R*) 3,4,5trihydroxy-6-(hydroxymethyl)oxan-2yl]oxy 3*H*benzo[f][2]benzofuran-1-one	$C_{27}H_{26}O_{12}$
4	Daturine (*Datura* sp.)	[(1*R*,5*S*)-8-methyl-8azabicyclo[3.2.1]octan3-yl] (2*S*)-3-hydroxy-2phenylpropanoate	$C_{17}H_{23}NO_3$
5	Embelin (*Embelia ribes*)	2,5-dihydroxy-3undecylcyclohexa-2,5diene-1,4-dione	$C_{17}H_{26}O_4$
6	Nicotine (*Nicotiana tabacum*)	3-[(2*S*)-1methylpyrrolidin-2yl]pyridine	$C_{10}H_{14}N_2$

2.2.5: 2D interaction diagram

Protein-Ligand 2D interaction diagram was drawn with Ligplot Plus software using the docked file created by PyMOL (.pdb file).[16]

2.2.6: Predicting the binding site

The binding sites in protein were predicted from the Pocasa 1.1 Web Server.[17]

The *in silico* tools used in the present investigation are summarized in the Table 4.

Table 4. The list of *in silico* tools used in the present investigation.

Tool			Purpose
Name	**Type**	**Licence**	
AutoDock Vina	Software	Open source	Molecular docking
Avogadro	Software	Open source	Generation of 3D structure and force field application to the Ligand
BioEdit	Software	Open source	Counting of rare and near to stop-codon amino acids for close stop ratio and rare amino acid ratio
BLAST (NCBI)	Webserver	Open source	Homology search in host (*G. hirsutum*)
CodonW	Software	Open source	Calculation of GC content, amino acid length, hydrophobicity, Nc, CAI
Cytoscape	Software	Open source	Calculation of BC, CC, DC (Network topology features)
Gerstainlab.org	Website	Open source	Supplementary material of *S. cerevisiae* training dataset (14 features) (Seringhaus *et al.*, 2006)
Ligplot Plus	Software	Academic licence	2D interaction diagram of protein-ligand complex
LOCTREE3	Web server + Database	Open source	Subcellular localization features
MS Excel 2013	Software	Paid licence	Sorting of datasets
NCBI	Database	Open source	Genome (mRNA sequences) of *E. gossypii*
OPEN BABEL	Software	Open source	Conversion between various formats of 3D structures
Orange	Software	Open source	Prediction of essentiality of genes in *E. gossypii* (Machine learning simulators)
Phyre2	Webserver	Open source	Prediction of protein 3D structure
POCASA 1.1	Webserver	Open source	Predictions of top-5 binding pockets
Pubchem	Database	Open source	Ligand (drug) 3D structure
PyMOL	Software	Paid licence	Visualization of 3D structure
Rapidminer	Software	Paid licence	Prediction of essentiality of genes in *E. gossypii* (Machine learning simulators)
RaptorX	Webserver	Open source	Prediction of protein 3D structure
STRING	Database	Open source	Protein-protein interaction data (for Network topology features)
The protein model portal	Webserver	Open source	Submission of protein sequences for modelling via various protein modellers (webservers)
TMHMM	Web server	Open source	Provided transmembrane helices feature
UniProt	Database	Open source	Proteome (protein sequences) of *E. gossypii*.
Weka	Software	Open source	Prediction of essentiality of genes in *E. gossypii* (Machine learning simulators)

3. RESULTS

3.1. Prediction of essential genes in *E. gossypii* using machine learning

Training the classifier

Among all the ORFs and mRNA sequences downloaded, the sequences containing the data of all the 17 genome features selected for the study were sorted out for further investigation. With this criterion, 3,497 genes of *S. cerevisiae* (training dataset) and 3,968 genes of *E. gossypii* (testing dataset) were sorted out. The nomogram obtained illustrates the relative contributions of each predictive feature to the target class *i.e.*, essentiality. The vacuolar, mitochondrial and cytoplasmic localization showed a negative correlation with essentiality while the nuclear, endoplasmic reticular and other compartment localization showed a positive correlation to the essentiality (Figure 1). For the network topology features (betweenness centrality, closeness centrality, degree centrality), the features having higher values showed a positive correlation with essentiality and *vice versa*. The codon adaption index (CAI) and length of amino acids (AA) features having higher values also showed a positive correlation to the essentiality. For the rest of the features, higher values showed negative correlation while lower values showed positive correlation with the essentiality. Subsequently, the training dataset was cross-validated using three simulators. First, using Orange simulator a comparative study of cross-validation.

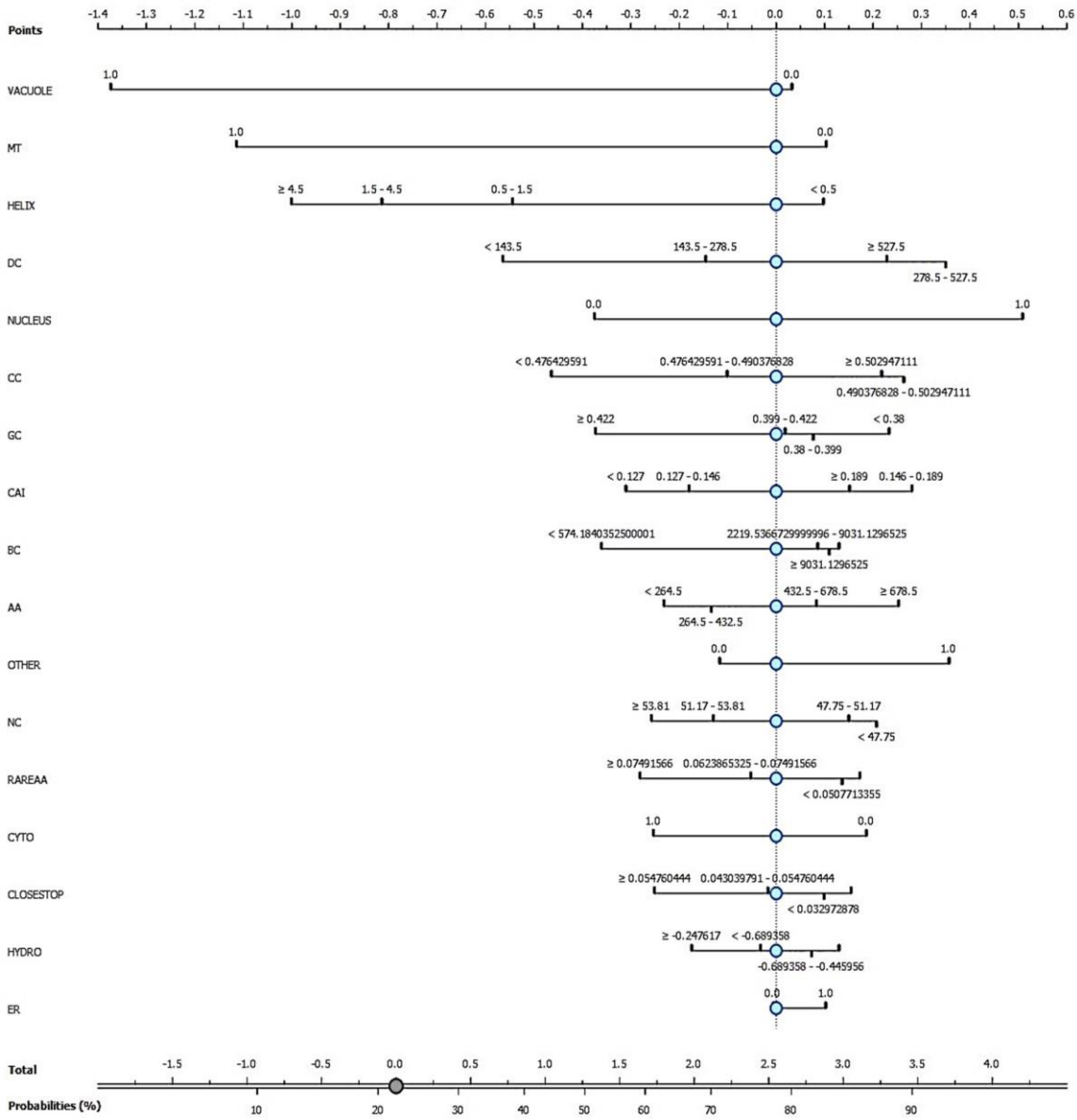


Fig. 1. A nomogram showing relative importance of each feature to the essentiality.

was performed for three different training datasets. These three datasets contained 8 features (except localization and network topology feature), 14 features (except network topology features) and 17 features, respectively. With Orange simulator, the cross-validation results of all the three training datasets (8 features, 14 features and 17 features) showed that most of the genes fall in the false-negative and true-negative category. The percentage of true-positive and false-positive categories increased as the number of features used for analysis increased. In contrast, as the number of features for analysis increased, the percentage of false-negative and true-negative entries decreased from 91.90 to 73.50% and 98.10 to 93.70%, respectively (Figure 2). The training dataset containing 17 features performed better in terms of true-positive result and therefore was selected to proceed further with. However, 17 feature dataset also accompanied higher false-negative as-well-as higher false-positive results. Since we were interested in predicting essential genes, the occurrence of false-negative entries could be tolerated, but not the false-positive ones. Therefore, a special training dataset was prepared for the study.

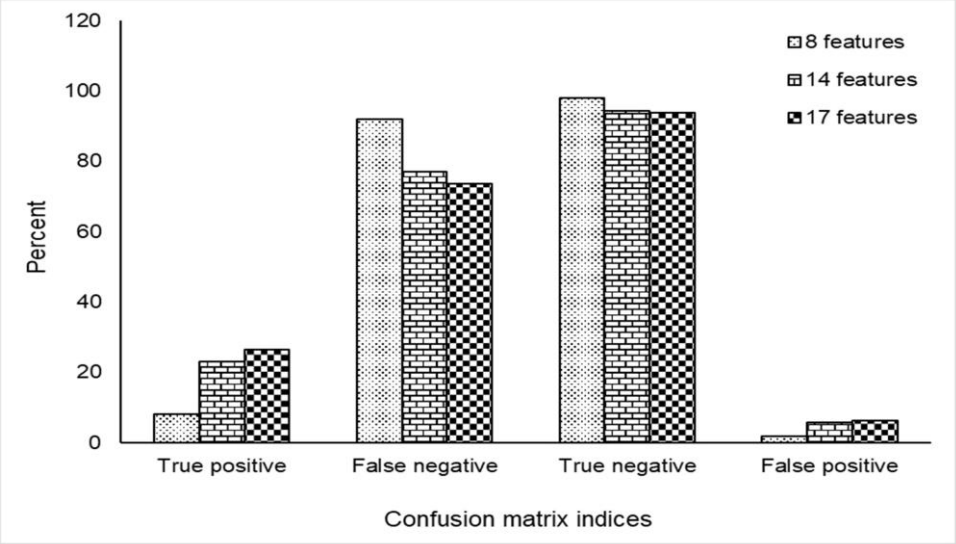


Fig. 2: A comparative analysis of training datasets using Orange simulator.

Further in Orange simulator, only four classifiers viz., Neural Network, Logistic Regression, Random Forest and Naïve Bayes showed the AUC (area under the ROC curve) value above 0.6 and thus were considered reliable for prediction (Figure 3). The Rapidminer simulator showed the better performance of Naïve Bayes (kernel), Random Forest, Random tree and Decision Stump classifier algorithms. The ROC curve of these classifiers forms a steep curve and stand out from the rest of the three classifiers *viz*., k-NN (2), Naïve Bayes and Decision Tree (Figure 4). On the other hand, algorithms like Naïve Bayes, Logistic regression, Adaboost, J48 tree, Random Forest and Random Tree were selected for the cross-validation on the basis of the results obtained from the Weka simulator. The confusion matrix obtained from Weka have most of the entries in the true-negative and false-negative category. However, the true-positives and false-positives entries varied from 5.7 to 35.4% and 1.2 to 20.3%, respectively. The Random Tree classifier performed the best while the Adaboost classifier was weak in performance (Figure 5).

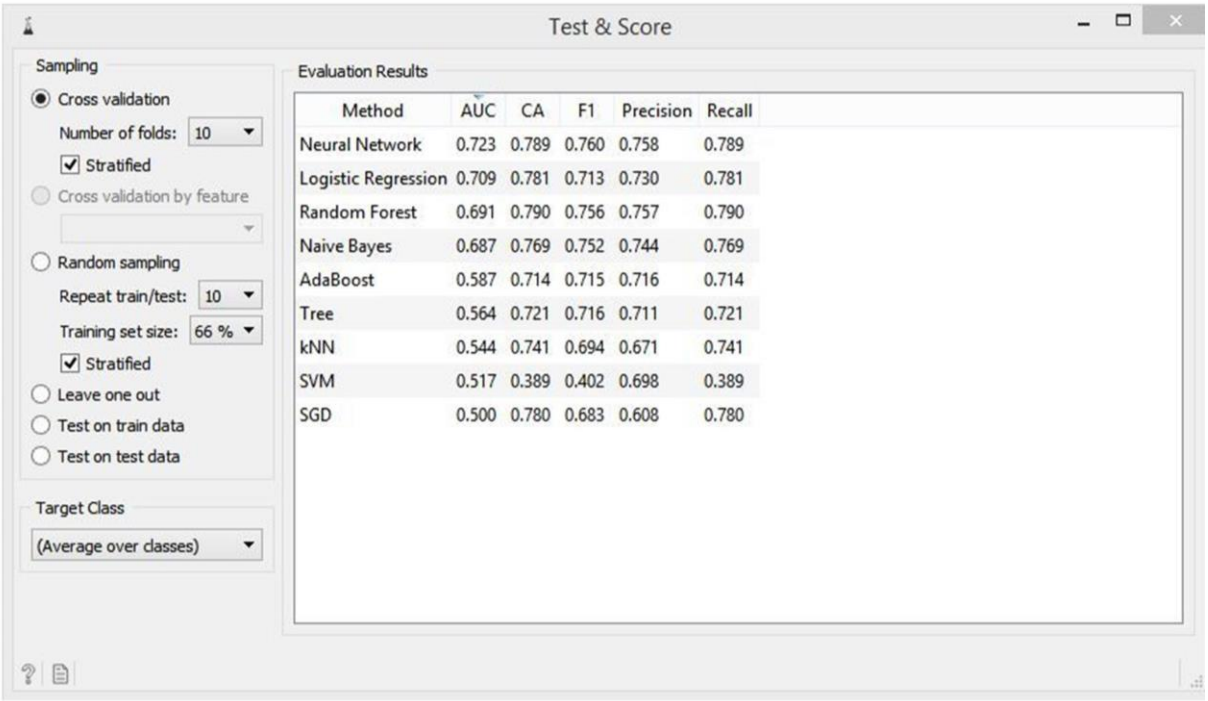


Fig. 3: Evaluation results obtained using Orange simulator for the training dataset of *S. cerevisiae* (17 features).

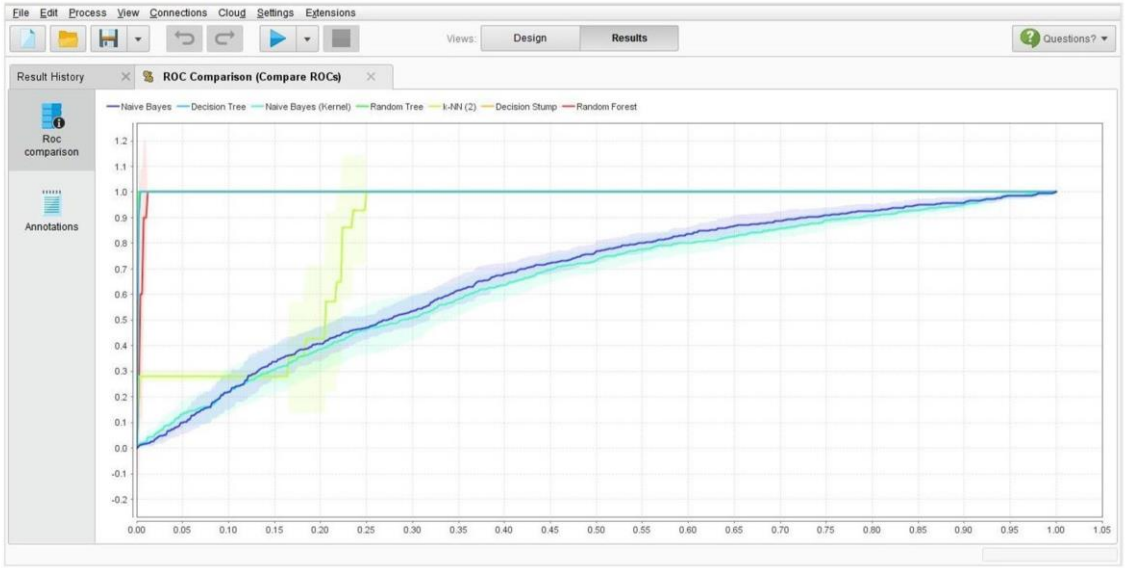


Fig. 4: ROC curve for cross-validation of *S. cerevisiae* training dataset (17 features) using Rapidminer.

A special training set was prepared out of 17 genome features of *S. cerevisiae* primarily to minimize the false-positive predictions. A comparison between the confusion matrices of the normal and special training dataset reveals that the value of the special dataset is more than the normal dataset for true-positive and true-negative predictions. Similarly, the value for false-negative and false-positive is lesser in the special training dataset compared to the normal training dataset. This observation is true for all the classifiers used viz., Neural Network, Naïve Bayes, Random Forest and Logistic Regression. Since, the efficiency of the special dataset is more than the normal dataset; special training dataset of *S. cerevisiae* was used as input for the selected machine learning classifier algorithm (Table 5).

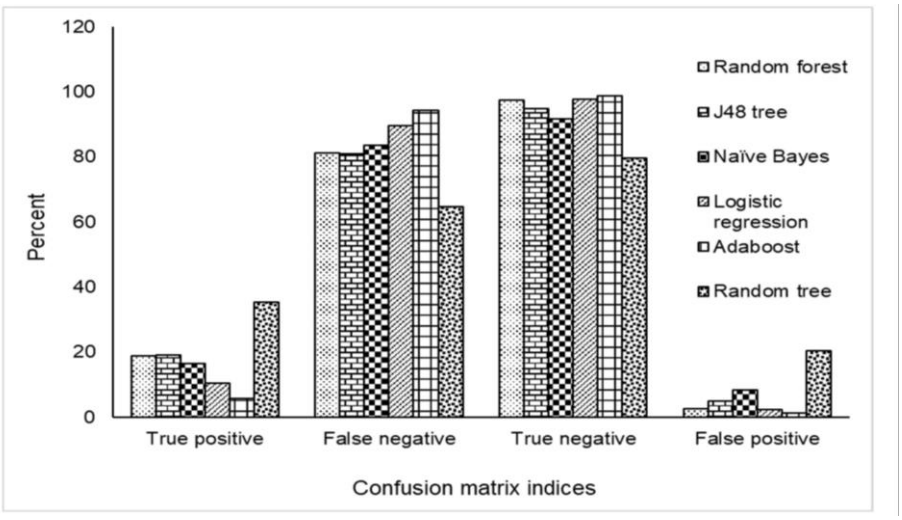


Fig. 5: Confusion matrix indices obtained for the training dataset of S. cerevisiae (17 features) using various classifiers in Weka.

Table 5: A comparison between the normal and special datasets with respect to the prediction performance through confusion matrix.

Classifier	Confusion matrix indices							
	True positive (%)		False negative (%)		True negative (%)		False positive (%)	
	Normal	Special	Normal	Special	Normal	Special	Normal	Special
Neural Network	26.5	71.3	73.5	28.7	93.7	98.9	6.3	1.1
Naïve Bayes	31.5	70.1	68.5	29.9	89.7	95.5	10.3	4.5
Random Forest	23.9	40.2	76.1	59.8	94.5	99.5	5.5	0.5
Logistic Regression	8.2	27.6	91.8	72.4	97.8	99.0	2.2	1.0
Average	22.53	52.30	77.48	47.70	93.93	98.23	6.08	1.78

Selection of target in *E. gossypii*

Given a gene and its associated 17 input variables, each classification algorithm generated a separate probability estimate of essentiality. In each simulator, the unweighted average of these estimates of all the selected algorithms were used to predict the essential genes. In Orange simulator, only Logistic Regression and Random Forest predicted the essential genes in *E. gossypii*. The other two algorithms *viz*., Naïve Bayes and Neural network predicted all the genes as non-essential. Therefore, unweighted average of only Logistic regression and Random forest algorithms was considered. Thus, 461 and 3,507 genes were predicted as essential and non-essential, respectively. The Naïve Bayes (kernel) predicted 144 genes as essential and 3,523 as non-essential genes from the provided dataset using the Rapidminer simulator. However, the other algorithms predicted almost all the genes as 'non-essential'. Similarly, all but Adaboost and J48 predicted all the genes as 'non-essential' in Weka simulator. Therefore, the unweighted average of Adaboost and J48 tree only was considered for the analysis of essentiality. The Adaboost and J48 tree predicted 446 genes as essential while the remaining 3,522 genes were predicted as non-essential.

Based on our criteria of designating a gene as 'essential' only if it is identified as essential by more than one simulator, 230 genes were chosen as 'essential' in *E. gossypii* (Table 6). Out of these 230 genes, the products of three of them localized in the plasma membrane. These genes encoded the plasma membrane ATPase (NM_211643.1), CDP-diacylglycerol-serine-O-phosphatidyl transferase (NM_210567.1) and leucine-rich repeat protein (NM_209864.1). Later, the localization of the products of these genes in plasma membrane was confirmed and functionally annotated by searching their homologs in the database. The homology of all the three genes was the most with the homologs of *S. cerevisiae*. Further, the homology also confirmed the plasma membrane localization and functional annotation of the products of the gene with accession numbers NM_211643.1 and NM_210567.1. However, the localization and functional annotation of the product of the gene with accession number NM_209864.1 was doubtful, as the homolog of this protein locates in the nucleus in *S. cerevisiae.*

Prevention of unintended interaction

One of the attributes of the gene product for drug targeting is that it should not be present in the host. This ensures that the ligand does not interfere with the normal functions in the host. Therefore, the homology search between the selected three genes of *E. gossypii* and genome of *G. hirsutum* was performed using Blastn. The highest 'query cover' during the search was 9, 2 and 1% for the gene with accession number NM_211643.1, NM_210567.1 and NM_209864.1, respectively (Table 7). Since none of the genes showed the appreciable homology with the host genes, all the three gene products were selected for molecular docking.

Table 6. The accession number (GenBank) of the genes predicted to be essential in *E. gossypii*.

NM_001355341.1	NM_209046.1	NM_210300.2	NM_211345.1
NM_001355358.1	NM_209068.1	NM_210303.1	NM_211393.1
NM_207722.1	NM_209119.2	NM_210338.1	NM_211411.1
NM_207731.1	NM_209123.1	NM_210360.2	NM_211413.2
NM_207782.1	NM_209169.1	NM_210380.1	NM_211428.1
NM_207805.2	NM_209194.1	NM_210402.1	NM_211431.2
NM_207825.2	NM_209195.1	NM_210416.1	NM_211484.1
NM_207840.1	NM_209200.1	NM_210425.1	NM_211495.1
NM_207849.1	NM_209240.1	NM_210427.1	NM_211535.2
NM_207854.2	NM_209248.2	NM_210444.1	NM_211536.1
NM_207880.1	NM_209267.1	NM_210454.1	NM_211539.1
NM_207885.1	NM_209286.2	NM_210463.2	NM_211553.1
NM_207888.1	NM_209296.1	NM_210465.2	NM_211555.1
NM_207907.2	NM_209315.1	NM_210553.2	NM_211569.1
NM_207972.1	NM_209351.1	NM_210554.1	NM_211572.1
NM_207975.1	NM_209352.2	NM_210567.1	NM_211623.3
NM_207994.1	NM_209361.2	NM_210609.2	NM_211625.1
NM_207998.1	NM_209363.1	NM_210621.1	NM_211641.2
NM_208007.1	NM_209367.2	NM_210683.2	NM_211643.1
NM_208025.1	NM_209426.1	NM_210709.2	NM_211645.1
NM_208026.2	NM_209513.2	NM_210730.1	NM_211668.1
NM_208027.1	NM_209522.2	NM_210744.1	NM_211704.1
NM_208046.1	NM_209545.1	NM_210762.2	NM_211712.1
NM_208056.1	NM_209553.1	NM_210765.1	NM_211731.1
NM_208060.1	NM_209557.1	NM_210812.1	NM_211733.1
NM_208065.1	NM_209584.1	NM_210861.1	NM_211798.1

NM_208095.1	NM_209586.1	NM_210869.1	NM_211800.2
NM_208185.1	NM_209589.1	NM_210890.2	NM_211830.1
NM_208188.1	NM_209621.2	NM_210896.1	NM_211839.1
NM_208222.1	NM_209652.2	NM_210904.1	NM_211862.1
NM_208246.1	NM_209703.2	NM_210948.1	NM_211868.1
NM_208251.2	NM_209709.1	NM_210983.1	NM_211894.1
NM_208292.2	NM_209736.1	NM_210988.2	NM_211911.1
NM_208298.1	NM_209752.2	NM_210997.2	NM_211922.1
NM_208357.2	NM_209793.2	NM_211039.1	NM_211935.1
NM_208360.2	NM_209801.2	NM_211044.2	NM_211944.1
NM_208410.2	NM_209816.1	NM_211073.1	NM_211971.2
NM_208419.2	NM_209823.2	NM_211081.1	NM_211976.1
NM_208507.2	NM_209833.2	NM_211156.2	NM_212004.1
NM_208579.2	NM_209864.1	NM_211182.1	NM_212031.2
NM_208606.1	NM_209912.2	NM_211185.2	NM_212061.2
NM_208613.1	NM_210002.1	NM_211187.1	NM_212104.2
NM_208646.2	NM_210019.1	NM_211191.1	NM_212127.2
NM_208654.1	NM_210020.1	NM_211220.1	NM_212168.1
NM_208682.1	NM_210021.1	NM_211230.2	NM_212169.2
NM_208691.2	NM_210082.1	NM_211233.1	NM_212181.2
NM_208695.1	NM_210100.2	NM_211234.2	NM_212211.2
NM_208744.2	NM_210163.1	NM_211250.1	NM_212246.1
NM_208806.1	NM_210176.1	NM_211269.1	NM_212263.1
NM_208843.2	NM_210186.2	NM_211272.1	NM_212265.1
NM_208850.1	NM_210192.2	NM_211276.1	NM_212270.1
NM_208854.1	NM_210208.2	NM_211292.2	NM_212291.2
NM_208886.2	NM_210212.1	NM_211304.1	NM_212321.1
NM_208902.1	NM_210213.1	NM_211306.1	NM_212333.1
NM_208936.1	NM_210240.1	NM_211310.1	NM_212374.2
NM_208994.1	NM_210262.1	NM_211326.2	NM_212387.2
NM_209001.1	NM_210265.1	NM_211336.2	NM_212393.1
NM_209011.1	NM_210291.1		

Table 7: The result of Blastn search carried out for the selected essential gene.

Accession number of the gene of *E. gossypii*	Blastn homologue	Max score	Total score	Query cover (%)	E value	Identity (%)
NM_211643.1	*Gossypium hirsutum* cultivar TM-1 chromosome 19, ASM98774v1	91.5	447	9	2e-15	70.98
NM_210567.1	*Gossypium hirsutum* cultivar TM-1 chromosome 13, ASM98774v1	38.3	38.3	2	10.0	95.65
NM_209864.1	*Gossypium hirsutum* cultivar TM-1 chromosome 19, ASM98774v1	43.7	43.7	1	0.66	90.91

3.2. Molecular Docking

'The Protein Model Portal' server was assigned the function of modelling the 3D structure of the protein of interest (target). Out of all the assigned modelling portals, only Phyre2 and RaptorX returned the 3D structures of all the three proteins selected for docking (Figures 6-7). The identities of the two predicted models for each protein were more or less similar (Table 8). Therefore, both the models were selected for molecular docking except the model of the protein with the accession number NP_984511.1. Since the identity for NP_984511.1 was low, only Phyre2 model was used for docking studies. Similarly, the 3D conformation of the selected ligands was also retrieved from the online resources (Figure 8).

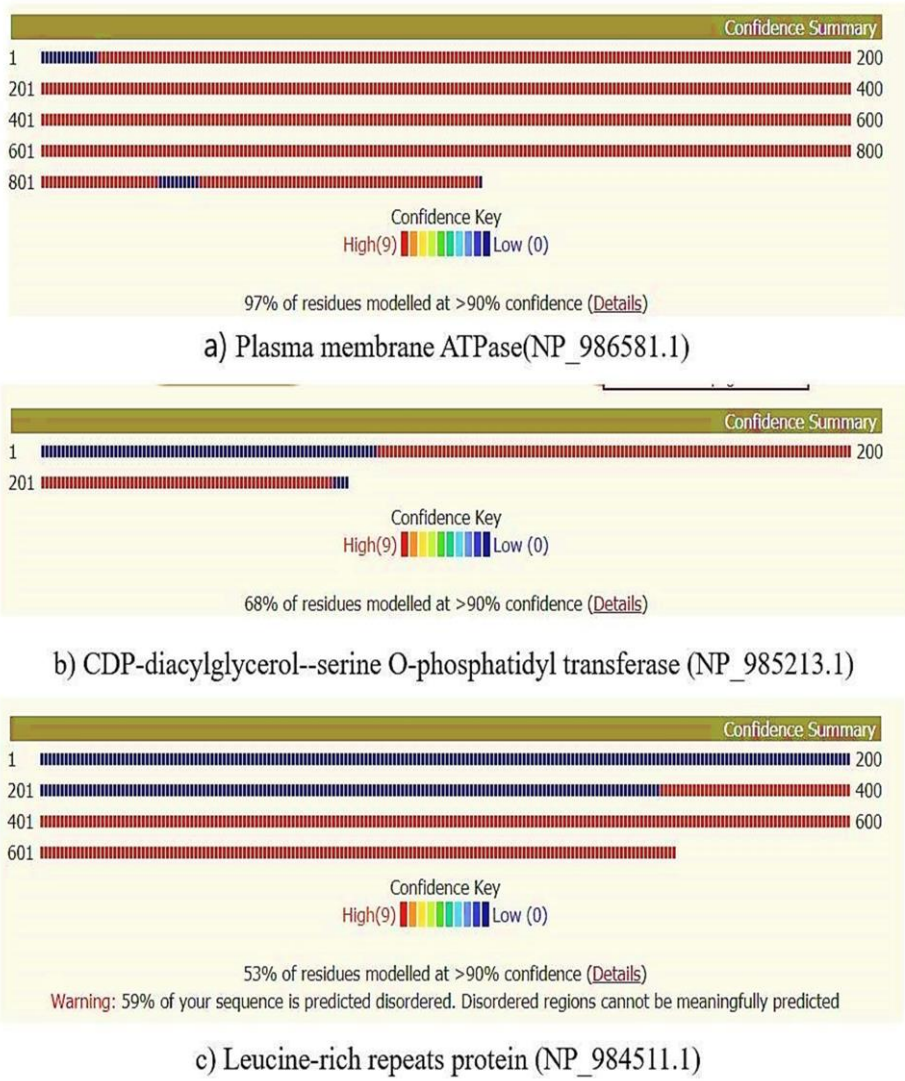


Fig. 6 Confidence summary of the Phyre2 protein model for the proteins selected for molecular docking.

Autodock Vina software was assigned the function of blind docking between receptor and ligand. There were five protein models (two each for NP_986581.1 & NP_985213.1 and one for NP_984511.1) and six ligands for testing. Thus, 30 receptor-ligand pairs were formed for docking. For each receptor-ligand pair, the Vina generated 'nine' conformations of ligand in complex with the receptor position. These conformations were ranked on the basis of binding energy, which indicates the strength of interaction and the affinity with which a compound binds to a pocket of the target protein. For each pair, only the best-docked position (lowest binding energy) was short-listed (Table 9). Later, the docking was also performed after force-field application and the docking scores calculated (Table 9). The two-way ANOVA performed using protein model and ligand as the factors revealed no significant difference in the docking score before and after force-field application.

A 2D receptor-ligand interaction diagram was generated using Ligplot Plus software. The diagram showed the polar as-well-as hydrophobic interactions of the ligand with the surrounding amino acids of the receptor (Figures 9-10).

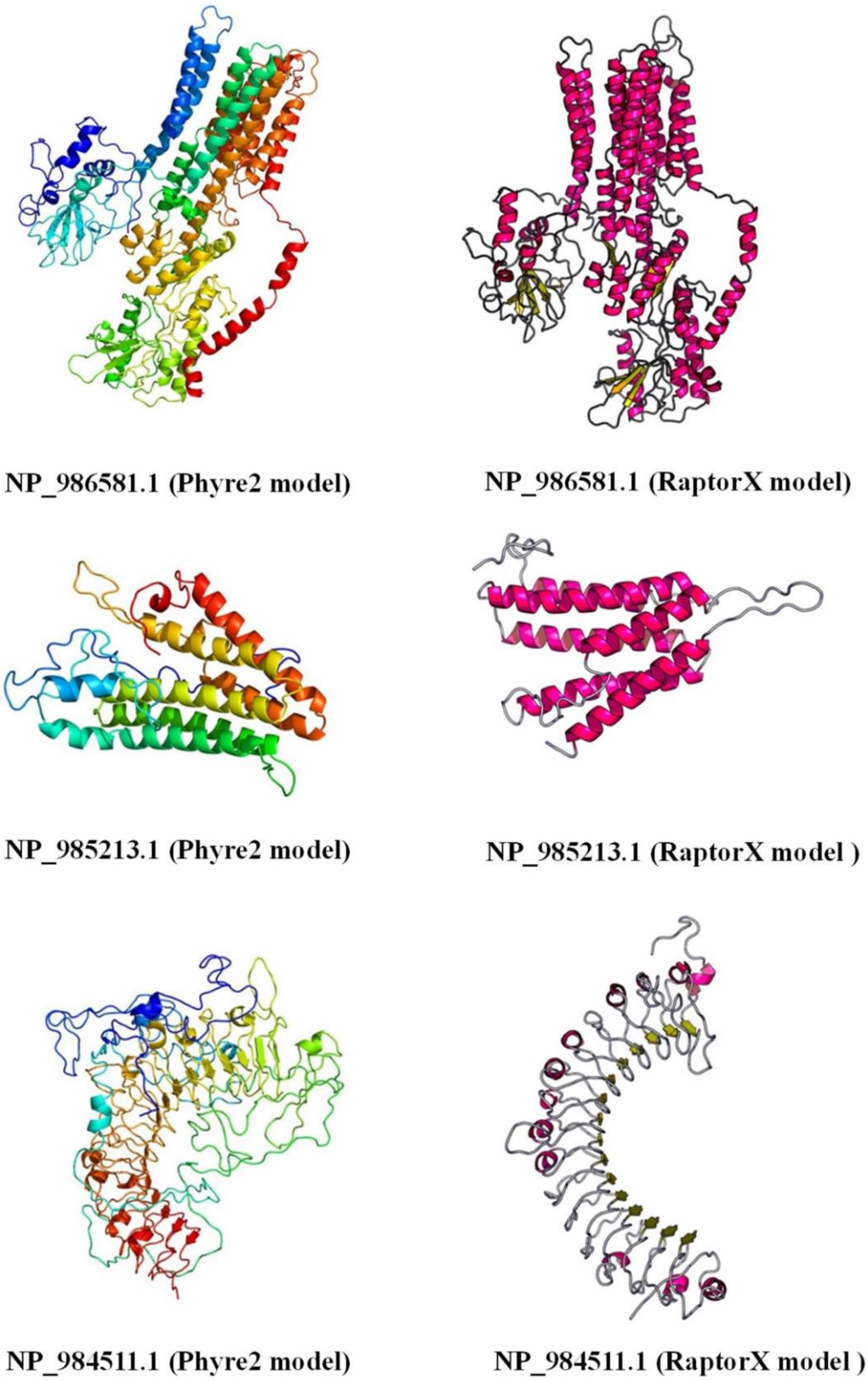


Fig. 7. The predicted 3D conformations of the selected proteins.

In all the receptor-ligand complexes the number of hydrophobic interactions are more than the polar interactions displaying the high efficiency of the selected ligands (Figure 11). Cleistanthin A and cleistanthin B had higher interaction with the plasma membrane ATPase (Phyre2 and RaptorX) and CDP-diacylglycerol-serine-O-phosphotidyl transferase (Phyre2). Similarly, embelin also had higher interaction with CDP-

diacylglycerol-serine-O-phosphotidyl transferase (RaptorX) and leucine-rich protein (Phyre2). In addition, cleistanthin B also displayed the higher polar interactions with all the models. However, we could not get the output for azadirachtin, probably because the 3D structure of azadirachtin was not experimentally determined but deduced by Avogadaro.

Lastly, the binding sites were predicted using the online server Pocasa. This server generated the data of the top-5 binding pocket for every protein model. The visualization of the binding pocket and the best-docked position of each ligand indicates that cleistanthin B is bound in all 5 models in one of the top-5 binding pockets (Figure 12). Similarly, the azadirachtin, cleistanthin A, daturine and embelin bound in 4 models in top-5 position while the nicotine bound in 2 models in top-5 binding pockets.

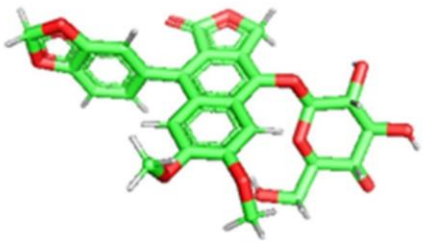
Table 8: Protein model identity of the three proteins localising in the plasma membrane of *E. gossypii*.

Accession number of the protein of *E. gossypii*	3D modelling server	Identity (%)
NP_986581.1 (Plasma membrane ATPase)	Phyre2	97
	RaptorX	100
NP_985213.1 (CDP-diacylglycerol-serine-O-phosphatidyl transferase)	Phyre2	68
	RaptorX	72
NP_984511.1 (Leucine-rich repeats protein)	Phyre2	53
	RaptorX	52

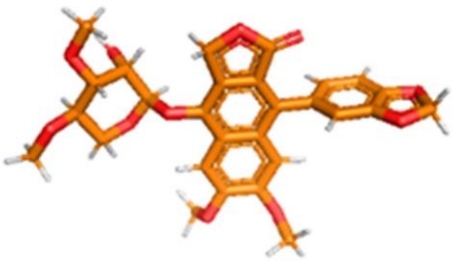
4. DISCUSSION

The recent years have witnessed a surge in using an *in silico* approach in drug targeting and discovery. Workers prefer the virtual screening for putative drugs against the target because of the environmental and human health issues.[18,19,20] However, the application of *in silico* tools in plant pathology is still in its infancy.[21]

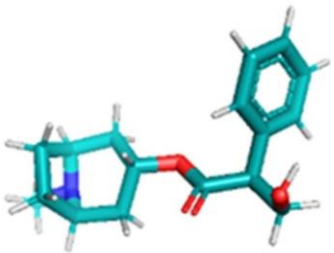
Our study shows that the features such as vacuolar localization, mitochondrial localization, transmembrane helices, betweenness centrality (below value 2219.5) show a strong negative correlation to the essentiality. The features such as nuclear localization, ER localization and other localization, on the contrary, show a strong positive correlation with the essentiality of the gene.



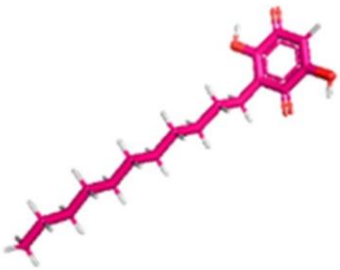
Cleistanthin A



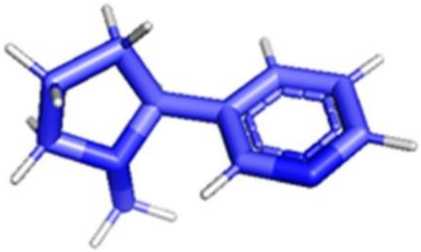
Cleistanthin B



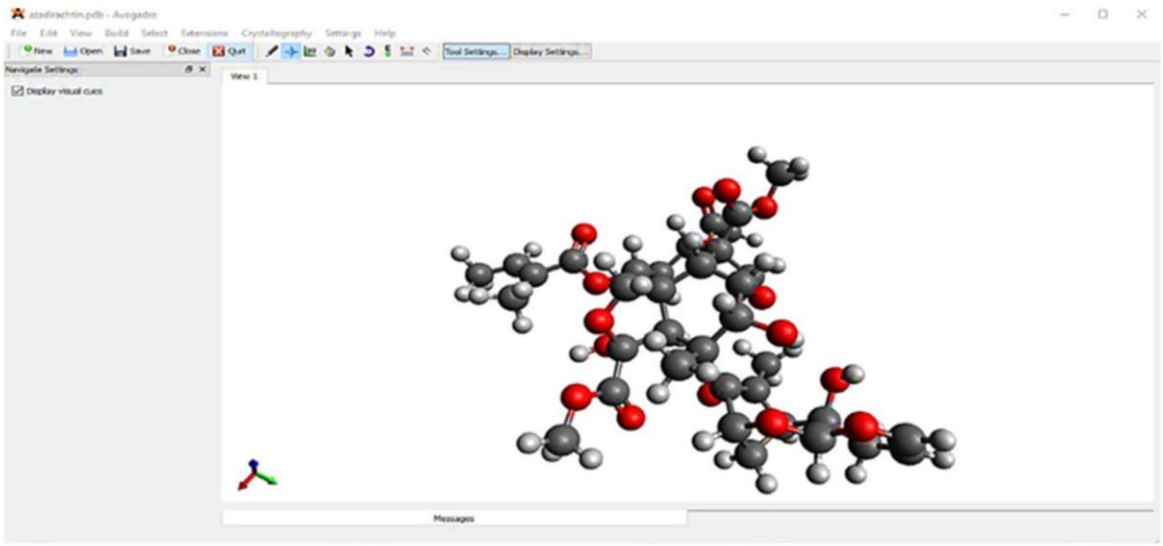
Daturine



Embelin



Nicotine



Azadirachtin (Generated in AVOGADRO)

Fig. 8. The 3D conformation of the ligands selected for the study.

Table 9: The docking scores obtained before and after the application force-field.

Protein Model \ Ligand	Azadirachtin		Cleistanthin A		Cleistanthin B		Daturine		Embelin		Nicotine	
	No FF	FF	No FF	FF	No FF	FF	No FF	FF	No FF	FF	No FF	FF
NP_986581.1 (Phyre2)	-8.6	-8.6	-8.7	-8.1	-8.9	-8.9	-7.8	-7.9	-6.0	-5.7	-5.9	-6.5
NP_986581.1 (RaptorX)	-8.1	-8.1	-8.8	-9.1	-8.7	-9.1	-7.8	-7.8	-7.3	-5.8	-6.2	-6.6
NP_985213.1 (Phyre2)	-7.5	-7.5	-9.1	-9.2	-8.6	-9.2	-7.5	-7.4	-5.8	-6.3	-6.1	-5.8
NP_985213.1 (RaptorX)	-7.8	-7.8	-8.3	-8.3	-89	-8.3	-7.5	-7.8	-8.5	-7.7	-6.2	-6.3
NP_984511.1 (Phyre2)	-7.9	-7.9	-8.2	-8.7	-9.8	-8.7	-7.8	-7.6	-7.3	-6.7	-6.1	-5.9

Note: No FF = Before the application of force field, FF = After the application of force field

For the rest of the features, some had slightly higher negative odd log value and *vice versa*. Thus, in our case, each feature has played the role in determining the essentiality of the gene. Nuclear localization is the most common feature related to the essentiality of the gene.[6,22] The trend of relating essentiality with higher values of network topology features is in accordance with the previous results.[23,24] In addition, the direct relation between the protein length and essentiality of the gene is reported.[6,25]

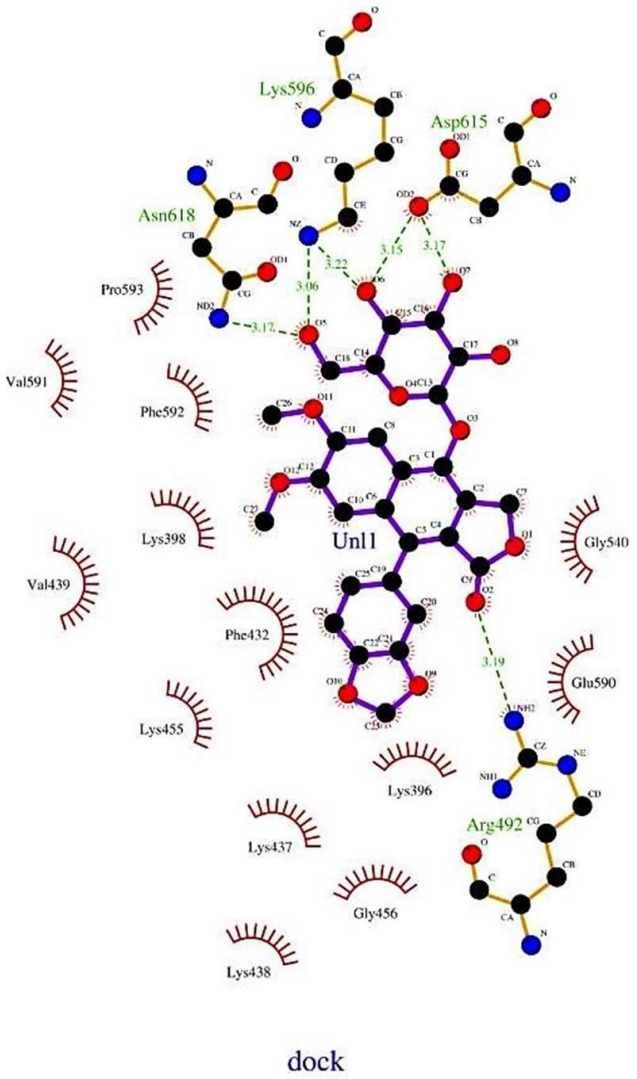


Fig. 9. The 2D protein-ligand interaction diagram for NP_986581.1 (Phyre2)—cleistanthin B complex.

Plasma membrane localized proteins are the prime targets for drug targeting experiments. Out of the three predicted essential genes products, in our study, the vital role of plasma membrane ATPase in survival and pathogenicity is experimentally demonstrated.[26,27] The docking results, without and with application of force field, showed that cleistanthin A and cleistanthin B performed better than the other ligands evaluated in the present study. This suggests that the docking score predictions are reliable. 2D-interaction diagram generated by Ligplot Plus software has given the information about the polar and hydrophobic interactions between ligand and the surrounding amino acids of the protein. Here also the score of cleistanthin A and cleistanthin B was higher compared to the other ligands. The highest interaction in terms of polar as-well-as hydrophobic interactions was between cleistanthin B and plasma membrane ATPase (NP_986581.1). Various toxicity studies have also shown that the cleistanthin A and B inhibits the H^+ ATPase activity in animals.[28] Ligplot Plus could not generate interaction diagram for complexes having azadirachtin presumably because we have used the structure of azadirachtin deduced by Avogadro. The experimentally determined 3D structure of azadirachtin was not at our disposal. Therefore, we excluded azadirachtin from the comparative analysis of the total interactions. It was no matter of concern because the docking score of azadirachtin was less than cleistanthin A and cleistanthin B. The performance of cleistanthin B was slightly better than the cleistanthin A in terms of polar as-well-as total interactions between ligand and the surrounding amino acids. Also, the analysis of top-5 binding pockets predicted from the Pocasa server shows that the cleistanthin B could bind to one of the top-5 pockets in all the 5 protein models while cleistanthin A could bind to one of the top-5 pockets in only 4 protein models. All these docking results suggest that the performance of the cleistanthin B is better than the other selected ligands.

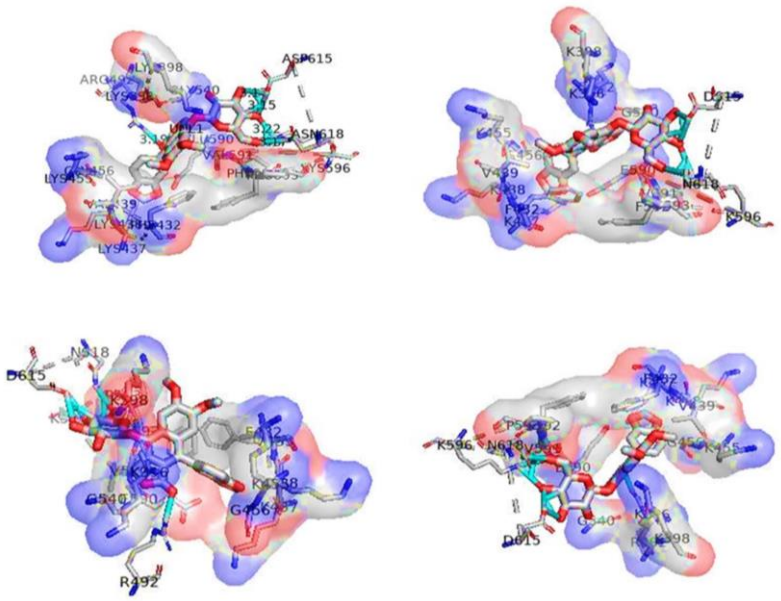


Fig. 10. The NP_986581.1 (Phyre2)—cleistanthin B complex in docked position showing the polar (sky-blue colour) and hydrophobic (surface amino acids) interactions.

4. CONCLUSION

From the present study, it is concluded that cleistanthin B interacts substantially with plasma membrane ATPase (GenBank accession number: NP_986581.1) encoded by the gene with accession number NM_211643.1 in *E. gossypii*. Therefore, it has the potential to acts as a 'green drug' against *E. gossypii* in *G. hirsutum*. However, a detailed wet-lab study is required to prove the utility of cleistanthin B.

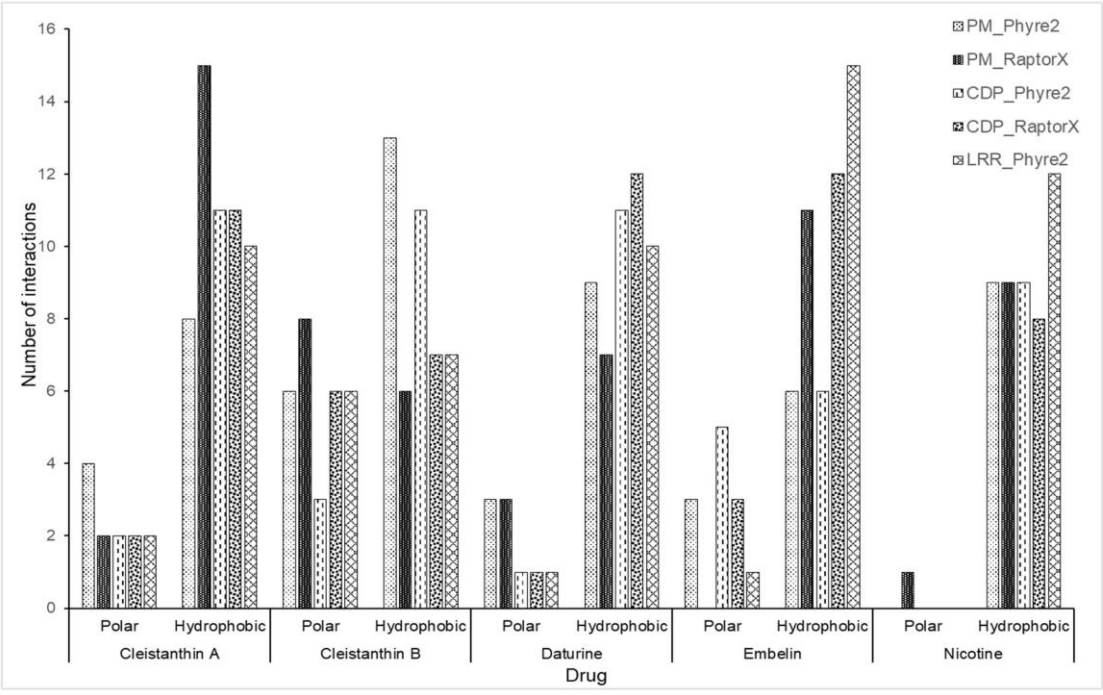


Fig. 11. Polar and hydrophobic interactions for the protein-ligand complexes.

PM_Phyre2: Plasma membrane ATPase protein with Phyre2 3D modelling server; PM_RaptorX: Plasma membrane ATPase protein with RaptorX 3D modelling server; CDP_Phyre2: CDP-diacylglycerol-serine-O-phosphotidyl transferse protein with Phyre2 3D modelling server; CDP_RaptorX: CDP-diacylglycerol-serine-O-phosphotidyl transferse protein with RaptorX 3D modelling server; LRR_Phyre2: Leucine-rich repeat protein with Phyre2 3D modelling server

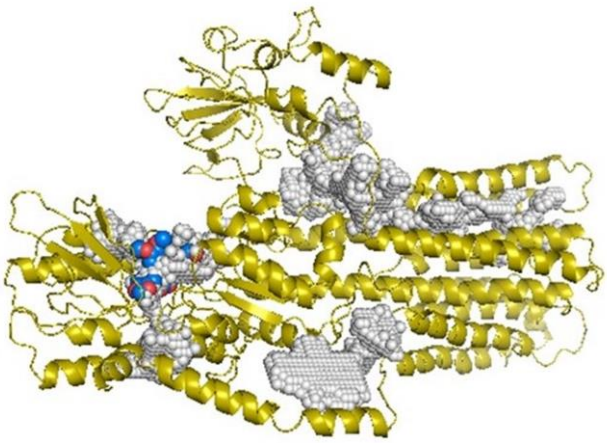


Fig. 12. The prediction of five best binding pockets for plasma membrane ATPase—cleistanthin B complex.

5. ACKNOWLEDGEMENT

We thank Dr. Roman Laskowski, European Bioinformatics Institute, UK for providing the *on gratis* access to the Ligplot Plus software.

REFERENCES

1. Pridham TG, Raper KB. *Ashbya gossypii*—Its significance in nature and in the laboratory. Mycologia, 1950; 42(5): 603-623.
2. Kurtzman CP, Fell JW, Boekhout T. The Yeasts: A Taxonomic Study, 5th edition. Burlington; Elsevier, 2011.
3. Michailides TJ. Pest Management Guidelines: Pistachio. University of California; ANR Publication, 2017.
4. Terstappen GC, Reggiani A. *In silico* research in drug discovery. Trends in Pharmacological Sciences, 2001; 22(1): 23-26. doi: 10.1016/s0165-6147(00)01584-4
5. Santra HK, Banerjee D. Natural products as fungicide and their role in crop protection. In: Singh J, Yadav A, (eds.). Natural bioactive products in sustainable agriculture Singapore; Springer, 2020; 131-219.
6. Seringhaus M, Paccanaro A, Borneman A, Snyder M, Gerstein M. Predicting essential genes in fungal genomes. Genome Research, 2006; 16(9): 1126-1135. doi: 10.1101/gr.5144106.
7. Dietrich FS, Voegeli S, Kuo S, Philippsen P. Genomes of Ashbya fungi isolated from insects reveal four mating-type loci, numerous translocations, lack of transposons, and distinct gene duplications. G3: Genes, Genomes, Genetics, 2013; 3(8): 1225-1239. doi: 10.1534/g3.112.002881
8. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M *et al.* Orange: Data mining toolbox in Python. The Journal of Machine Learning Research, 2013; 14(1): 2349-2353.
9. Mierswa I, Klinkenberg R. RapidMiner Studio 9.1 https://rapidminer.com/, 2018.
10. Frank E, Hal, MA, Witten IH. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", 4th edition. Burlington; Morgan Kaufmann, 2016.
11. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L *et al.* The Protein Model Portal—a comprehensive resource for protein structure and model information. Database (Oxford), 2013; doi: 10.1093/database/bat031.
12. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. Journal of Cheminformatics, 2011; 3: 1-14. doi: https://doi.org/10.1186/1758-2946-3-33.
13. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. Journal of Cheminformatics, 2012; 4(1): 1-17. doi: https://doi.org/10.1186/1758-2946-4-17.
14. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry, 2010; 31(2): 455-461. doi: 10.1002/jcc.21334
15. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna:Austria 2021. URL https://www.R-project.org/.
16. Laskowski RA, Swindells MB. LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery. Journal of Chemical Information and Modelling, 2011; 51: 2778-2786. doi: https://doi.org/10.1021/ci200227u
17. Yu J, Zhou Y, Tanaka I, Yao M. Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. Bioinformatics, 2010; 26(1): 46-52. doi: https://doi.org/10.1093/bioinformatics/btp599
18. Khanna V, Kumar A, Shanker R. Identification of novel drug targets and lead compounds in Anthrax

and Pneumonia causing pathogens using an *in silico* approach. Chemical Informatics, 2015; 1(1): 3. doi: 10.21767/2470-6973.100003.
19. Ogungbe IV, Setzer WN. The potential of secondary metabolites from plants as drugs or leads against protozoan neglected diseases—Part III: In-silico molecular docking investigations. Molecules, 2016; 21(10): 1389. doi: 10.3390/molecules21101389.
20. Alam A, Tamkeen N, Imam N, Farooqui A, Ahmed MM, Tazyeen S. *et al.* Pharmacokinetic and molecular docking studies of plant-derived natural compounds to exploring potential anti-alzheimer activity. In: In Silico Approach for Sustainable Agriculture, Singapore: Springer, 2018; 217-238.
21. Shanmugam G, Jeon J. Computer-aided drug discovery in plant pathology. Plant Pathology Journal, 2017; 33(6): 529-542. doi: 10.5423/PPJ.RW.04.2017.0084.
22. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinformatics, 2009; 10(1): 1-18. doi: https://doi.org/10.1186/1471-2105-10-290
23. Gustafson AM, Snitkin, ES, Parker SC, DeLisi C, Kasif S. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. BMC Genomics, 2006; 7(1): 1-16. doi: https://doi.org/10.1186/1471-2164-7-265.
24. Hwang YC, Lin CC, Chang JY, Mori H, Juan HF, Huang HC. Predicting essential genes based on network and sequence analysis. Molecular Bio Systems, 2009; 5(12): 1672-1678. doi: https://doi.org/10.1039/B900611G.
25. Deng J, Deng L, Su S, Zhang M, Lin X, Wei L. *et al.* Investigating the predictability of essential genes across distantly related organisms using an integrative approach. Nucleic Acids Research, 2011; 39(3): 795-807. doi: 10.1093/nar/gkq784.
26. Serrano R, Kielland-Brandt MC, Fink GR. Yeast plasma membrane ATPase is essential for growth and has homology with (Na^++K^+), K^+– and Ca^{2+}–ATPases. Nature, 1986; 319(6055): 689-693. doi: 10.1038/319689a0
27. Li J, Yang S, Li D, Peng L, Fan G, Pan S. The plasma membrane H+-ATPase is critical for cell growth and pathogenicity in *Penicillium digitatum*. Applied Microbiology and Biotechnology 2022; 106: 5123-5136. doi: 10.1007/s00253-022-12036-4.
28. Chrispal A. *Cleistanthus collinus* poisoning. Journal of Emergencies, Trauma, and Shock, 2012; 5(2): 160-166. doi: 10.4103/0974-2700.96486.
29. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology, 1982; 157(1): 105-132. doi: 10.1016/0022-2836(82)90515-0.
30. Sharp PM, Li WH The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research, 1987; 15(3): 1281-1295. doi: 10.1093/nar/15.3.1281.
31. Wright F. The 'effective number of codons' used in a gene. Gene, 1990; 87(1): 23-29. doi: 10.1016/0378-1119(90)90491-9.
32. Hall TA BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, 1999; 41: 95-98.
33. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research, 2003; 13(11): 2498-2504. doi: 10.1101/gr.1239303.
34. Cheng J, Wu W, Zhang Y, Li X, Jiang X, Wei G. *et al.* A new computational strategy for predicting essential genes. BMC Genomics, 2013; 14(1): 1-13. doi: https://doi.org/10.1186/1471-2164-14-910.