

**COMPARATIVE EVALUATION OF TOOLS FOR GENE REGULATORY
NETWORK PREDICTION AND NETWORK RECONSTRUCTION
USING GENOMIC DATA**

By

RESHMA BHASKER T.

(2013-09-103)

Thesis

Submitted in partial fulfilment of the

Requirement for the degree of

B. Sc. – M. Sc. (INTEGRATED) BIOTECHNOLOGY

Faculty of Agriculture

Kerala Agricultural University, Thrissur



B. Sc. – M. Sc. (INTEGRATED) BIOTECHNOLOGY

DEPARTMENT OF PLANT BIOTECHNOLOGY

COLLEGE OF AGRICULTURE

VELLAYANI, THIRUVANANTHAPURAM-695522

KERALA, INDIA

2018

DECLARATION

I, hereby declare that this thesis entitled “**COMPARATIVE EVALUATION OF TOOLS FOR GENE REGULATORY NETWORK PREDICTION AND NETWORK RECONSTRUCTION USING GENOMIC DATA**” is a bonafide record of research work done by me during the course of research and that the thesis has not previously formed the basis for the award of any degree, diploma, associate ship, fellowship or other similar title, of any other University or Society.



RESHMA BHASKER T.

(2013-09-103)

Place: Vellayani

Date: 07.12.2018

भा.कृ.अनु.प- केंद्रीय कन्द फसल अनुसंधान संस्थान 3

(भारतीय कृषि अनुसंधान परिषद, कृषि और किसान कल्याण मंत्रालय, भारत सरकार)
श्रीकार्यम, तिरुवनन्तपुरम-695 017, केरल, भारत

ICAR- CENTRAL TUBER CROPS RESEARCH INSTITUTE

(Indian Council of Agriculture Research, Ministry of Agriculture and Farmers Welfare, Govt. of India)
Sreekariyam, Thiruvananthapuram-695 017, Kerala, India



CERTIFICATE

Certified that this thesis entitled “**COMPARATIVE EVALUATION OF TOOLS FOR GENE REGULATORY NETWORK PREDICTION AND NETWORK RECONSTRUCTION USING GENOMIC DATA**” is a record of research work done independently by Ms. Reshma Bhasker T. (2013-09-103) under my guidance and supervision and this is not previously formed the basis for the award of any degree, diploma, fellowship or associateship to her.

Place: Sreekariyam

Date: 07-12-2018

Dr. J. Sreekumar

(Chairman, Advisory Committee)

Principal Scientist

Section of Extension and Social Sciences

ICAR-CTCRI, Sreekariyam,

Thiruvananthapuram, 690517

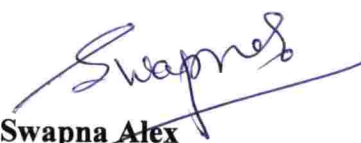
डॉ. जे. श्रीकुमार / Dr. J. SREEKUMAR
प्रधान वैज्ञानिक (कृषि सांख्यिकी)
Principal Scientist (Agricultural Statistics)
एक्सटेंशन और सामाजिक विज्ञान अनुभाग
Section of Extension and Social Sciences
भा.कृ.अनु.प-केंद्रीय कंद फसल अनुसंधान संस्थान
I C A R - Central Tuber Crops Research Institute
श्रीकार्यम / Sreekariyam
तिरुवनन्तपुरम / Thiruvananthapuram - 695 017

CERTIFICATE

We, the undersigned members of the advisory committee of Ms. Reshma Bhasker T. (2013-09-103), a candidate for the degree of B. Sc. – M. Sc. (Integrated) Biotechnology, agree that the thesis entitled “**Comparative evaluation of tools for gene regulatory network prediction and network reconstruction using genomic data**” may be submitted by Ms. Reshma Bhasker T. in partial fulfilment of the requirement for the degree.



Dr. J. Sreekumar
(Chairperson, Advisory Committee)
Principal Scientist (Agrl. Statistics),
Section of Extension and Social
Sciences, ICAR-CTCRI. Sreekariyam,
Thiruvananthapuram- 695 017



Dr. Swapna Alex
(Member, Advisory Committee)
Professor & Head
Department of Plant Biotechnology
College of Agriculture, Vellayani
Thiruvananthapuram – 695 522



Dr. M. N. Sheela
(Member, Advisory Committee)
Principal Scientist & Head
Division of Crop improvement
ICAR-CTCRI. Sreekariyam
Thiruvananthapuram - 695 017



Dr. K. B. Soni
(Member, Advisory Committee)
Professor & Course Director
B. Sc. – M. Sc. (Integrated) Biotechnology
Department of Plant Biotechnology
College of Agriculture, Vellayani
Thiruvananthapuram – 695 522



Dr. M. K. Rajesh
(External Examiner)
Principal Scientist
(Biotechnology)
ICAR-CPCRI
Kasaragod – 671 124, Kerala

ACKNOWLEDGEMENT

I would like to offer my genuine gratitude to the God almighty for blessing me with this opportunity to explore the realms of science. I am deeply grateful to my beloved advisor Dr. J. Sreekumar for his guidance, patience and motivation. I am indebted to him for the continuous encouragement and support he has provided over the span of this one year.

My special thanks to the present Director of ICAR-CTCRI, Dr. Archana Mukherjee for allowing me to successfully accomplish my project and also for the support provided throughout the course of research. I wish to convey my deep sense of gratitude to Dr. Anil Kumar, DEAN, College of Agriculture, Vellayani and Dr. Swapna Alex, Professor and Head, Department of Plant Biotechnology, College of Agriculture for their insightful comments and encouragement. I also would like to put on record my sincere thanks to Dr. Sheela Immanuel, Head, Section of Extension and Social Sciences, ICAR-CTCRI for extending all facilities required for completing my work. My special thanks to Dr. M. N. Sheela, Principal Scientist and Head, Division of Crop Improvement, for being part of my advisory committee and providing necessary suggestions and guidance's at appropriate times.

My heartfelt thanks to Dr. K. B. Soni, our Course Director for her continuous support and encouragement throughout the course of study and specially for her valuable comments and guidance that helped a lot in generating research interest in the field.

I would like to express my heartfelt gratitude to Mr. Ambu Vijayan, who was always there to guide me to proceed in the right track by providing valuable information regarding the technical aspects of Bioinformatics research. My special thanks to my co-researchers in Bioinformatics and Statistics Lab, ICAR-CTCRI, Miss Gayathri, Miss Sruthi, Miss Haritha, Mis Shilpa, Mr. Akshay, Miss Rekha and Miss Priya for their consistent support and motivation that paved way for the completion of this project.

My special mention to Mr. Athul, Miss Sahla and Miss Aswathy who were both my lab mates and classmates for their whole-hearted cooperation and for the moral support that they have provided during my difficult times. My special thanks to my beloved friends Mr. Achuth P. Jayaraj, Miss Bhagyalakshmi and Miss. Alina A. Nazir without whom I wouldn't be able to make it. I also thank my all other class mates for their selfless care and invaluable motivation.

I must express my profound gratitude to my parents and my sisters for providing unfailing support and continuous encouragement throughout my life.

I also would like to thank all teaching and non-teaching staffs, students, researchers etc. of both KAU and ICAR-CTCRI, who have contributed for the completion of this work directly or indirectly, for their timely help and continuous mentoring.

RESHMA BHASKER T.

TABLE OF CONTENTS

Sl. No.	Chapters	Page No.
1.	INTRODUCTION	1-3
2.	REVIEW OF LITERATURE	4-20
3.	MATERIALS AND METHODS	21-31
4.	RESULTS	32-48
5.	DISCUSSION	49-52
6.	SUMMARY	53-55
7.	REFERENCES	56-71
8.	APPENDICES	72-81
9.	ABSTRACT	83-84

LIST OF TABLES

Sl. No.	Title	Page No.
1.	Parameters for determining interaction statistics	28
2.	Screening of Immunity related genes from immune protein domains	32
3.	Characteristic parameters of Hidden Markov Model	34
4.	Genes predicted to confer TMV resistance in cassava	37
5.	Network Statistics for N=50	41
6.	Network Statistics for N=100	42
7.	Network Statistics for N=150	43
8.	Parameters for interaction statistics	44
9.	Parameters for comparing the performance of the methods	45
10.	Comparison of tools used for GRN prediction	47

LIST OF FIGURES

Sl. No.	Title	Between Pages
1.	Comparison of cassava productivity in India	5-6
2.	Phytozome Database	32-33
3.	Pfam Database	32-33
4.	Workflow for the construction of regulatory network of immunity related genes in cassava	30-31
5.	Domain annotation and alignments in HMMER v3.1b2	34
6.	Model of HMMERsearch result	33-34
7.	Interactions of the predicted genes obtained from STRING v10.5	37-38
8.	Model of the generated HMM file	36
9.	Distribution of sequences according to biological processes	37-38
10.	Gene Ontology Mapping of the identified genes	37-38

11.	Distribution of sequences according to molecular function	37-38
12.	KEGG Pathway Mapping of identified sample identified gene	37-38
13.	Distribution of sequences in the cellular component	37-38
14.	Workflow for Weighted Gene Coexpression Network Analysis	39
15.	Networks with a size of N=50 constructed using different algorithms	40-41
16.	Networks with a size of N=100 constructed using different algorithms	41-42
17.	Networks with a size of N=150 constructed using different algorithms	42-43
18.	Networks constructed with ARACNE having different network sizes, N=500, 1000, 1500	43-44
19.	Synthetic Transcriptional Regulatory Network Constructed using SynTRen	45-46
20.	Accuracy Plot of different methods	45-46
21.	Statistical comparison of performance of different tool	45-46

22.	Protein-Protein Interaction obtained from STRING v10.5	45-46
23.	Neighbourhood pattern of the identified genes obtained from STRING database	45-46
24.	The reconstructed pathway visualized using Cytoscape v3.6.1	47-48

LIST OF APPENDICES

Sl. No.	Title	Appendix No.
1.	List of immunity related genes identified from cassava	I
2.	Simulated dataset generated from SynTReN	II

LIST OF ABBREVIATIONS

amiRNAs	artificial microRNAs
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
BANJO	BAyesian Network inference with Java Objects
BiHEA	Hybrid Evolutionary Approach for Microarray Biclustering
Cas9	Caspase-9
CBSD	Cassava Brown Streak Disease
CBSV	<i>Cassava Brown Streak Virus</i>
CKSUM	Training alignment checksum
CMD	Cassava Mosaic Disease
CMG	<i>Cassava Mosaic Geminivirus</i>
CRISPR	Clustered Regularly Interspaced Palindromic Repeats
CWB	Cassava witches broom
EFFN	Effective Number of Sequences
FANOVA	Functional Analysis of Variance
FAOSTAT	Food and Agriculture Organization Corporate Statistical Database
GA	Genetic Algorithm
GeNESiS	Gene Network Evolution Simulation Software
GEPASI	General Pathway Simulator
GNW	GeneNetWeaver
GRN	Gene Regulatory network
GRNCOP2	Gene Regulatory Network Inference by Combinatorial Optimization 2

GRRANN	Gene Regulatory network-based Regularized Artificial Neural Network
IITA	International Institute of Tropical Agriculture
IRP	Immune Related Protein
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MICMIC	Methylation Regulation Network Inference by Conditional Mutual Information based PC-algorithm
NSEQ	Number of Sequences
OS	Operating System
RNA	Ribo-Nucleic Acid
SCENIC	Single Cell rEgulatory Network Inference and Clustering
SDC	Sulfur-containing Defence Compounds
SEBINI	Software Environment for Biological Network Inference
sgnesR	Stochastic Gene Network Expression Simulator in R
SINCERITIES	SINGLE CELL Regularized Inference using Time-stamped Expression profileS
SIR	Sulfur-Induced Resistance
SIRENE	Supervised Inference of Regulatory Network
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
SynTRen	Synthetic Transcriptional Regulatory Networks
TRN	Transcriptional Regulatory Network
URL	Uniform Resource Locator
US	United States
WGCNA	Weighted Gene Expression Network Analysis

INTRODUCTION

1. INTRODUCTION

Cassava (*Manihot esculenta* Crantz, $2n=36$) is a perennial shrub, found to be originated in the Amazon Basin (Olsen *et al.*, 1999) with its centre of diversity in the Brazilian-Bolivian region (Nassar *et al.*, 2002). It belongs to Euphorbiaceae family and Fabid superfamily, in which several plants such as rosids, legumes and poplars that are distantly related, are included. (Prochnik *et al.*, 2012). Cassava gains immense importance as a staple food in countries like Africa with a global production of 277 million tonne. It seems to cover an area of 23 million ha cultivated land with a yield of 11,800 kg/ha (FAOSTAT, 2016). Genome-wide association studies for the genetic improvement of cassava has led to sequencing of 158 diverse cassava varieties and identification of 3,49,827 single-nucleotide polymorphisms (SNPs) and indels (Zhang *et al.*, 2018). Being one of the most important crops in tropical and sub-tropical countries constituting a major source of carbohydrates, its productivity is highly threatened by mainly two viral diseases, Cassava Mosaic Disease (CMD) and Cassava Brown Streak Disease (CBSD). Resistance to such pathogens in plants is achieved by the action of a well advanced innate immune system contributed by multiple layered defence protein network structures. These proteins are involved in the activation of several complex plant responses like protein interactions, signal transduction pathways and gene expression changes. The study of their interaction networks would help in the better understanding of plant immune responses that help in adopting better crop improvement and management strategies.

The major challenge faced by biologists in the twenty first century to understand structure and functions of a living cell, is to get a clear picture of the complex intracellular web of interactions between numerous constituents of the cell such as DNA, RNA, proteins and small molecules. These convoluted interactions result in its particular biological activity. Hence there is a demand for the understanding of the Network Biology of a cell where the biological layers are represented as network models. A Gene Regulatory Network (GRN) can be defined

as a group of molecular regulators which both interact among themselves and also with other substances in the cell for accomplishing an objective of governing the gene expression levels of mRNA and proteins (Karlebach and Shamir, 2008). Genome wide expression analysis in combination with gene perturbation experiments provides powerful tool for the construction of plant GRNs (Krouk *et al.*, 2013).

Introduction of better techniques that could assist in gene expression analysis in a genome wide scale could be made possible by the introduction of techniques that could make the systematic characterisation of plant GRNs effortless. Possible incorporation of expression microarrays or RNA-seq experiments includes some of the approaches that could make the process better feasible. GRNs could provide sufficient information regarding the regulatory interactions that occur between regulators and their potential targets, gene-gene interactions, and potential protein-protein interactions could be obtained from GRNs (Simoes *et al.*, 2012).

GRN construction has been found necessary for the complete elucidation of disease ontology which could probably reduce the cost of drug development that could accelerate biomedical research and development. Time series transcriptomic data that are usually measured by genome-wide DNA microarrays has been traditionally used for GRN modelling till date. Several novel experimental and computational approaches like Boolean networks, Bayesian networks, mutual information-based approaches, correlation-based approaches etc. has made the characterisation of regulatory networks possible. Integration, interpretation and evaluation of data from genomic databases could also represent biological knowledge that are normally represented in the form of gene or protein networks which shows functional or co-expression relationship or any other structured representation (Leal *et al.*, 2013).

A difficult issue that we face in evaluation of GRN inference algorithms is in the precise measurement of direct regulatory relationship between genes and hence the gold standards scenarios where such interactions are known with high confidence, are rarely defined. A universally accepted strategy for the assessment of GRN inference methods is the use of the Area Under the Curve (AUC) as a global

metric of performance for an algorithm. A Receiver Operating Characteristic (ROC) curve may be possibly used for the evaluation of a weighted network against a gold standard and hence the best method for GRN reconstruction could be probably identified. As using expression data for GRN construction makes these methods less feasible and expensive, the possibility of incorporating gene sequence information for network construction could be evaluated for exploring large scale regulatory network that will better elucidate their functional properties.

The current study focuses on the generation of gene regulatory network using genomic data depicting the immune regulatory network of cassava, which would potentially identify the top candidate genes related to immune responses in a quicker and feasible manner, comparison of different computational methods for the prediction and analysis of regulatory network of genes, and development of an online visualization tool using these different methods.

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

2.1 CASSAVA

Cassava (*Manihot esculenta* Crantz), which is a major source of carbohydrate, that grows throughout the lowland tropics (Nassar *et al.*, 2008), is viewed as one of the most important crops and a major energy source (Cock *et al.*, 1982) in tropical and sub-tropical regions considering their socio-economic perspectives. Uniquely being the only cultivated species among the genus, cassava feeds around 800 million people around the world with main focus on countries where food shortages are common (Nhassico *et al.*, 2008). Its immense potential as functional foods and nutraceutical ingredients (Chandrasekhara *et al.*, 2016) paves way for a diverse range of application in the industrial sector. Small and medium-sized industries (SME) mainly in food and agriculture sector, plays a major role in the enrichment of income source in rural areas by increasing the value of additional activities of crops like cassava (Unteawati *et al.*, 2018). This typically diploid highly heterozygous species seems to play a major role in ensuring food security across the world due to its drought tolerance and appropriate adaptation to climate changes. High starch content in cassava, which are cultivated throughout tropical Africa, Asia and America, make it suitable both for biofuel applications (Jansson *et al.*, 2009) and human consumption. It includes about 94 reported species and about 6300 varieties, cultivated in more than 100 countries utilizing its root and leaves as source of food and feed.

Cassava, although being fourth largest source of calorie in the world, is severely affected economically in terms of yield loss due to various viral, bacterial and fungal infections. According to the FAOSTAT (2017) comparison data (Figure 1), there has been a significant reduction in yield of cassava in India from 38,581 kg/ha in 2012 to 22,323 kg/ha in 2016. Similarly, the appearance of two viral diseases, Cassava Mosaic Disease (CMD) and Cassava Brown Streak Disease (CBSD) seems to significantly constrain the productivity in East and Central African countries and is estimated to cause an annual loss of worth US\$1 billion

according to the reports by International Institute of Tropical Agriculture (IITA), 2014. The other pathogens causing infection in cassava includes bacterial species like *Xanthomonas axonopodis* pv. *manihotis*, which causes bacterial blight (Verdier *et al.*, 1998), *Erwinia carotovora* subsp. *carotovora* causing bacterial stem rot (Lozano *et al.*, 1978), fungal species like *Colletotrichum gleosporioides* f. sp. *manihotis* causing anthracnose (Fokunang *et al.*, 1997), *Fusarium oxysporum* causing Fusarium root rot (Kuldau *et al.*, 2000) etc. Oomycetes like *Pythium* spp. also causes infection in cassava in the form of Pythium root rot (Msikita *et al.*, 2005) and *Phytophthium* sp. is reported to cause storage root rot and foliage blight in cassava (Boari *et al.*, 2018).

Early identification plays a major role in managing the detection and spread of cassava diseases. Approaches for identification of diseases traditionally worked mainly with the support of agricultural extension organizations. But countries with low human infrastructure and logistical capacity seems to face limitations in applying this approach and their expensive scale-up also adds up to yet another disadvantage (Plucknett *et al.*, 1998). Hence it is very important for us to develop a better feasible technology that could possibly help in early disease detection. Fofana *et al.* (2004), showed that model plants like *Arabidopsis thaliana*, *Nicotiana benthamiana*, *N. tabacum*, *Lycopersicon esculentum* and others have well developed sequence databases and this could be used as a major source of information for functional genomic analysis in cassava. They constructed a gene silencing vector based on *African cassava mosaic virus* (ACMV) that carries a fragment from the *Nicotiana tabacum* sulfur gene (*su*). This was done to induce the silencing of the cassava orthologous gene that resulted in yellow-white spots, which is a characteristic of *su* expression inhibition. Modern technologies like transcriptome analysis of CBSD-resistant and susceptible cassava varieties infected with CBSV, based on RNAseq could identify genes involved in disease resistance (Maruthi *et al.*, 2014). Similarly, time series transcriptome analysis of cassava has been conducted in varieties challenged with *Ugandan cassava brown streak virus* (Amuge *et al.*, 2017).

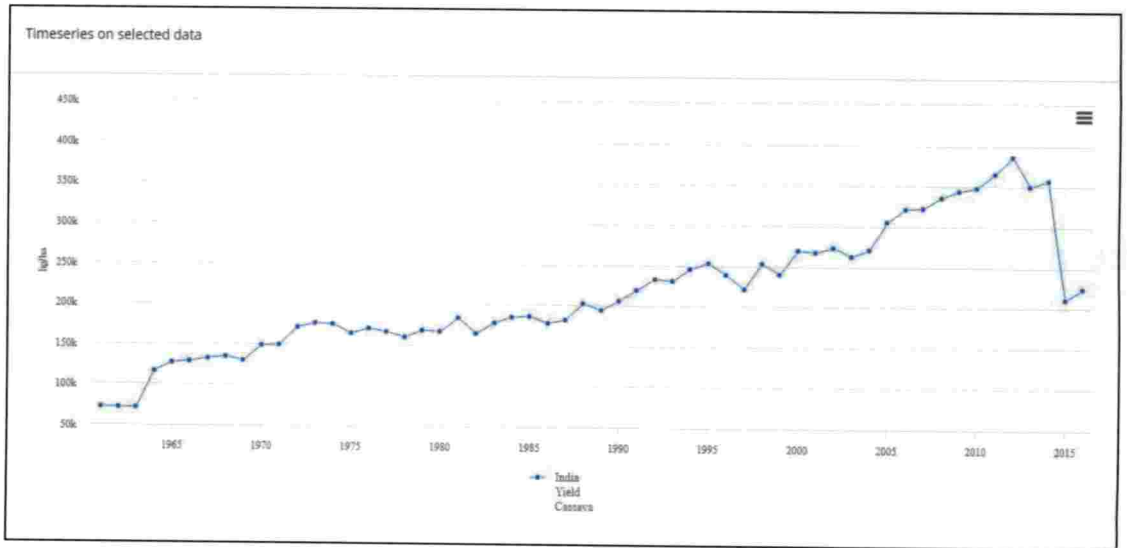


Figure 1. Comparison of Cassava Productivity in India (FAOSTAT, 2017)

A high dense genetic map of cassava was constructed by Soto *et al.* (2015), that contained 2,141 SNPs from which genes for 569 proteins related to immunity were localized. This was done based on the physical mapping data of the corresponding sequencing scaffolds. As a result of the *in-silico* screening for conserved domain, 1061 immune related protein coding genes were annotated.

This could provide data regarding the total number of genes that are currently annotated related to immunity in cassava as well as their organization and distribution in the cassava genome. Latest approaches for disease resistance and detection in cassava includes producing artificial microRNAs (amiRNAs) that could impart resistance to Cassava Brown Streak Disease (Wagaba *et al.*, 2016), deep learning approaches for image-based cassava disease detection (Ramcharan *et al.*, 2017), CRISPR-Cas9 in cassava to engineer resistance to African cassava mosaic virus (Mehta *et al.*, 2018) etc.

2.2 IMMUNITY IN CASSAVA

Despite of its defense mechanisms, the plants susceptibility to bacteria transmitted viral diseases (Boher and Verdier, 1994) and insect transmitted viral diseases (Hillocks and Jennings 2003; Patil and Fauquet, 2009) are very high. Cassava produces several compounds like cyanogenic glycosides, flavonoid glycosides, hydroxycoumarins etc which are seen to be involved in direct defences (Pinto-Zevallos *et al.*, 2016). Depending on the type of receptor involved in the plant-pathogen interaction, plant immunity is conferred in mainly two ways: first is by the microbe-associated molecular patterns (MAMPs) triggered immunity and second is by Effector triggered immunity imparted through R proteins or resistance protein. In MAMP or PAMP triggered immunity, Pattern Recognition Receptors (PRRs) recognize MAMPs and trigger immediate defence responses that lead to basal and nonhost resistance (Pérez-Quintero *et al.*, 2012). In plants, all known PRRs are plasma membrane resident proteins and hence they allow the perception of MAMPs at the cell surface. Leucine Rich Repeats (LRRs), LysM, kinases, WRKY, MAPK etc are examples for conserved domain associated with MAMP triggered immunity. Effector Triggered Immunity (ETI) acts by utilizing the host

cells evolutionarily conserved innate immune response, which can sense the pathogen through the activity of effectors produced by the pathogen and mount a robust immune response (Rajmohan *et al.*, 2014). The class of resistance protein includes TIR (Toll/interleukin-1 receptor), LRR, NB-ARC (Nucleotide-Binding domain shared by Apaf-1, R gene products, and CED-4) etc. In addition to these, Lectin Receptor Kinases and one of its major subtypes known as L-type Lectin Receptor Kinases are known to play major functions in plant development or abiotic stress tolerance (Wang *et al.*, 2017).

2.3 IMMUNITY RELATED GENES

2.3.1 ABC Transporters

ABC transporters serve many functions like contributing to plant growth, nutrition and development in plants, exhibiting proper response to abiotic stress, resistance against pathogens and also involved in plants interaction with its environment (Kang *et al.*, 2011).

They mediate the above ground and below ground secretion of anti-microbial secondary metabolites in plants, such as phenolics, cyanogenic glycosides, alkaloids, terpenoids and their derivatives and glucosinolates which forms an important first line of defence against pathogens that both host and non-host (Osborn *et al.*, 1996). They are characterized by two nucleotide-binding domains (NBD) and two transmembrane domains (TMDs) (Wilkins *et al.*, 2015).

2.3.2 Leucine Rich Repeats (LRRs)

The LRR structural motif consists of a conserved pattern of hydrophobic leucine residues. It acts as a platform for the mediation of several interactions between proteins, that are needed for exhibiting the dual role as sentry as well as activator of defence (Padmanabhan *et al.*, 2009). The LRR domain seems to have a slender, arc-shaped structure relative to globular proteins, with a high surface to volume ratio, that makes it suitable for involving in multiple interactions. LRR proteins serve to either act as resistance protein or proteins that are required for the functioning of required resistance proteins. It seems to provide resistance to a large

number of pathogens, including bacteria, viruses, nematodes, and fungi (Jones *et al.*, 1997). Serine/threonine kinases and polygalacturonase-inhibiting proteins were among the first LRR proteins to be described.

2.3.3 LysM

LysM is an ancient ubiquitous protein domain of about 40 amino acids in length and found in most living organisms except Archaea. Its structure contains 2 α -helices stacked onto one side of a double stranded antiparallel β -sheet (Bateman *et al.*, 2000). Evolution of this domain is closely related to an ancient sensor for N-acetylglucosamine (GlcNAc) (Buistet *et al.*, 2008). These glycans are believed to serve as immunogenic patterns activating LysM protein receptor mediated plant immunity and stopping microbial infection during early plant evolution (Zhang *et al.*, 2009). Further genetic studies also reveal the role of LysM-type receptor kinases for the establishment of legume-rhizobium symbiosis and also for plant mycorrhization.

2.3.4 NB-ARC

NB-ARC is a functional ATPase domain consisting of 3 subdomains: NB, ARC1 and ARC2. The nucleotide binding entity of NB-ARC regulates the activity of the R protein (van Ooijen *et al.*, 2008). The Nucleotide Binding subdomain is the catalytic core, the ARC1 subdomain acts as a scaffold for the intermolecular interaction with the Leucine Rich Repeats (LRR) and the ARS2b plays a major role in the regulating the transduction of pathogen perception into the activation of R-protein by the LRR (Rairdan and Moffett, 2006; Tameling *et al.*, 2006). NB-ARC domain adopts different conformations depending on the bound nucleotide (ATP or ADP) which will give a clear picture on the role that it plays as a molecular switch for the activation of R- protein (Tameling *et al.*, 2006).

2.3.5 NRAMP

Nramp stands for Natural Resistance- Associated Macrophage Protein gene seems to play an important role in modulating vertebrate natural resistance to intracellular pathogens (Jiang *et al.*, 2018). Host pathogen interface paves way for

several different responses and one among these is to withhold metals so that the growth of the microbes that invade them are retarded. This simple strategy can be effectively used for limiting infection by starving the invader of an essential element. Proteins of the Nramp domain are closely associated with conferring this “nutritional immunity” (Wessling *et al.*, 2015). Nramp domain proteins are generally categorized as members of a large family of divalent metal transporters that are evolutionarily conserved (Cellier *et al.*, 2012).

2.3.6 Protein Kinase

Protein Kinases become part of signal transduction pathways by modifying the protein function by phosphorylation. The two major properties that are essential for kinases to regulate multiple cellular responses include its high specificity for substrates and its sensitive means of regulation. PKs play a major role in MAMP triggered immunity with their representation as main PRRs in plasma membrane. They also have the ability to recognize cell wall signals and secondary danger-inducible plant peptides. These functions are particularly controlled by two types of RPKs, i.e. calcium-activated PKs and mitogen-activated PK (MAPK) cascades. Their signalling networks play a major role in controlling the activities and synthesis of enzymes, peptides, hormones, transcription factors (TF) and antimicrobial chemicals that contribute to resistance against bacteria, fungi and oomycetes (Tena *et al.*, 2011). Measurement of the phosphorylation of preferred substrate protein containing canonical serine-proline or threonine-proline phosphorylation sites can help in quantifying the activity of Kinases (Rodriguez *et al.*, 2010).

A characteristic ‘bean-like’ structure of the catalytic core of protein kinase domain is attributed by its small lobed N-terminal and large lobed C-terminal. This N-terminal and C-terminal are usually found as extensions from the catalytic core of protein kinase domain and often they fold back into the catalytic core leading to the formation of specific interactions.

ATP binding between the two almost will lead the adenine ring to lie deep in the cleft between the two lobes so that the γ -phosphate will be directed outwards (Biondi *et al.*, 2003). Almost one-third of the newly validated drugs that have emerged in the pharmaceutical industry is contributed by protein kinase targets, which contributes to its advancing applications in drug development.

2.3.7 TIR

The Toll/interleukin-1 receptor/resistance protein (TIR) domain refers to a protein-protein interaction domain, which consists of 125-200 residues in animals, plants and bacteria and absent in fungi, archaea and viruses (Jones *et al.*, 2017). TIR is an intracellular that is responsible for triggering immunity in plants when they perceive pathogen-associated molecular patterns (PAMPs) extracellularly. TIR domains are usually used by Plant Resistance proteins for pathogen detection and expression of genes involved in defense response inside the nucleus. They also seem to have a scaffold function in defense signalling. In plants, the TIR domains associated with intracellular immunity receptors are generally known as Nucleotide-binding oligomerization domain-Like Receptors (NLRs). General structure of plant TIR domain consists of a flavodoxin-like fold containing a central parallel β -sheet surrounded by an extended α D-helical region (Zhang *et al.*, 2017). The interaction of plant TIR domain with other TIR domain molecules acts as a key to the activation and hence the interface between these domains plays a key role in understanding its mechanism. (Thomas *et al.*, 2014).

2.3.8 WRKY

The WRKY transcription factor family is constituted by a large family of several plant transcription factors, acting both as repressors as well as activators. On the basis of the number of WRKY domain and certain zinc finger-like motif features, it is divided into three groups (Ishihama *et al.*, 2012). The members of the family play role in both repression and depression of important plant processes. DNA binding domain contributes to the most distinguishing feature of WRKY domain. It is called WRKY domain due to the presence of WRKY amino acid

sequence at the N-terminus. Several other amino acid sequences like WRRY, WSKY, WVKY, WKKY etc also seems to replace the WRKY proteins. The domain is about 60 residues in length containing a typical zinc-finger structure at the C-terminus (Eulgem *et al.*, 2000). The structure seen at the C-terminus is either C_{X4-5}C_{X22-23}HxH or C_{X7}C_{X23}HxC.

The typical structure of the WRKY domain consists of a four-stranded β -sheet forming a zinc-binding pocket with the zinc coordinating Cys/His residues. There is no existence of crystal structure in WRKY domain associated with its DNA-binding sites or for a full-length WRKY protein. The WRKY TFs are global regulators of host responses that help in regulates the defense gene expression at various levels, which seems to interact with key chromatin-remodelling factors, which together forms the WRKY network. MAP kinases in the nucleus, which are the key components of plant defense signalling are also associated with WRKY (Pandey *et al.*, 2009).

2.3.9 NBS-LRR

Plant proteins belonging to nucleotide-binding site (NBS) and leucine-rich repeats (LRRs) family, mostly encoded by R genes, are thought to be involved in pathogen detection. They recognize specialized pathogen effectors called avirulence (Avr) proteins, which provides virulence function in the absence of the cognate R gene (DeYoung and Innes *et al.*, 2006). They can be categorized into non-TIR and TIR classes based on the identity of the sequence preceding the NBS domain. The difference between TIR and non-TIR classes is contributed by the α -helical coiled-coil-like sequences at the amino terminal end of non-TIR and the amino terminal of TIR class seems to be homologous to the Toll and interleukin 1 receptors.

2.3.10 LECTIN

Lectins constitute an abundant multivalent group of proteins and/or glycoproteins, which are of non-immune origin that can reversibly bind to specific monosaccharides, oligosaccharides and glycoconjugates. Carbohydrate

Recognition Domain (CRD), which are the lectin binding sites on the carbohydrate, seems to be highly conserved in each type of lectin (Ni *et al.*, 1996). Lectins play a major role in endocytosis (Yi *et al.*, 2001) and intracellular transport of vector glycoprotein mechanisms (Yamamoto *et al.*, 2014), induction of apoptosis in tumoral cell (Kim *et al.*, 1993), blocking of HIV infection (Tanaka *et al.*, 2009), regulation of bacterial cell adhesion and migration (Tanne *et al.*, 2010) and control of protein levels in the blood. Lectins arrives as major contributors to the immune system by the recognition of carbohydrates that are found exclusively in pathogens, or that are inaccessible in host cells (Dias *et al.*, 2015).

2.4 GENE REGULATORY NETWORK (GRN)

A Gene-Regulatory Network (GRN) contributes to group of regulatory protein and their regulatory interactions that are involved in control and coordination of certain biological processes. The entire system consists of genes, cis-elements, and regulators. The regulation is mainly carried out by proteins, called transcription factors, and small molecules, like RNAs and metabolites. The level of gene expression during transcription is controlled by the interaction and binding of the regulators to cis-elements present on the cis-regions of the genes. The regulators mediate the aggregation of input signals which paves way for the specific gene expression signal. The gene network is constituted by the genes, regulators and the regulatory connections between them along with an interpretation scheme. GRN provides important information useful for drug design or medical-related fields hence the construction of GRN is a major focus in biological research. The generated network or module can serve as a working model for the formation of novel research hypotheses and assistance in experimental design.

2.5 METHODS FOR THE CONSTRUCTION OF GENE REGULATORY NETWORK

2.5.1 Probabilistic Boolean

They have been developed with an objective to study the logical interactions of genes without knowing specific details. Here the target gene is predicted by other

genes through a Boolean function. This method originally introduced by Kauffman (Shmulevich *et al.*, 2002) seems to be very useful to infer gene regulatory networks as it is able monitor the dynamic behaviour in complicated systems. This is achieved using large amounts of gene expression data. A stochastic extension of Boolean network, called Probabilistic Boolean Network (PBN), is made up of a family of networks corresponds to a contextual condition which is determined by variables outside the model. A structure-based method for fast simulation of PBNs was developed by Mizera *et al.* (2016), which initially performs a network reduction operation and then the nodes are divided into groups for parallel simulation (Mizera *et al.*, 2016). Probabilistic Boolean network (PBN) based on a network structure and desired steady-state properties have recently arisen to overcome the shortcomings of the earlier approach by using a matrix-based representation of PBN (Kobayashi and Hiraishi, 2017).

2.5.2 Dynamic Bayesian

Bayesian networks represent a general class of graphical models in which nodes are constituted by random variables and the lack of arcs amounts for conditional independence assumptions. The probabilistic nature of Bayesian network approaches paves way for its use in modeling genetic regulatory networks (Li *et al.*, 2007). Although Bayesian networks works well with static version of gene expression data, it faces a disadvantage in failing to capture temporal information and model cyclic networks. Zou *et al.*, 2018 presented an important approach for predicting the gene regulatory networks from time course expression data called DBN based approach. It seems to have several advantages like ability to model stochasticity, incorporation of prior knowledge and principled way of handling hidden variables and missing data. Here the number of potential regulators is limited to reduce search space.

As the quantities of biological data is limited, scientists are developing simulation approaches to improve DBN inference algorithms (Yu *et al.*, 2004). Even though likelihood maximization algorithms such as the Expectation-Maximization (EM) algorithm have been used to infer hidden parameters and deal with missing data, the effectiveness of current DBN methods is greatly reduced due

to low accuracy of prediction and excessive computational time.

2.5.3 Machine learning approach

This refers to a technique in the machines are programmed to learn patterns from data. Machine learning refers to a technique where the machines are programmed to learn patterns from data. The learning aims to develop a predictive model from a given dataset which is based on a set of mathematical rules and statistical assumptions. The above model could possibly predict any range of outputs like binary responses, categorical labels or continuous variables.

Machine learning methods are categorized into two, unsupervised learning and supervised learning (James *et al.*, 2013). The approaches are classified based on the availability of labels for the input data. Unsupervised methods are used if the labels on the input data are unknown. Here the learning takes place from the patterns in the features of the input data. The commonly used methods are principal components analysis (PCA) and hierarchical clustering. When the labels are available for the input data, supervised methods are applied. Here the labels are used for training the machine-learning model to recognize patterns that are able to predict the data labels.

Camacho *et al.* (2018) have discussed the opportunities and challenges faced at the intersection of machine learning and network biology. He also could possibly introduce a new approach for regulatory network construction based on deep learning. This seems to have drastic impact in disease biology, drug discovery, microbiome research, and synthetic biology. Ni *et al.* (2016) reports a machine learning approach to predict GRNs specific to developing *Arabidopsis thaliana* embryos. They developed the Beacon GRN inference tool which could predict GRNs occurring during seed development in *Arabidopsis* based on a support vector machine (SVM) model. Onik *et al.* (2018) have predicted a cancer-specific gene regulatory network using a simple and novel machine learning approach with linear regression and Pearson correlation coefficient.

2.5.4 Correlation based methods

Correlation based approach works based on the assumption that the interacting genes have correlated expression and methods like WGCNA (Weighted

Gene Correlation Network Analysis) which implements this methodology have proved to be consistently reliable and is widely adopted. Batushansky *et al.* (2016) introduces a series of methods for correlation-based network generation and analysis. This used freely available software that would allow the user to control each step of the network generation. It also provides an additional advantage in providing the flexibility in selection of correlation methods and thresholds.

2.5.5 Mutual Information based

Several information-theoretic approaches used in methods like as ARACNE (Margolin *et al.*, 2006b), CLR (Faith *et al.*, 2007) and minet (Meyer *et al.*, 2008) have found to be very successful in GRN construction. They compute the pairwise MIs between all possible pairs of genes, resulting in an MI matrix, which is then manipulated to identify the regulatory relationships (Altay and Emmert-Streib, 2010). Xing *et al.*, 2017 proposed a Candidate Auto Selection algorithm (CAS) based on mutual information. This algorithm automatically selects the neighbour candidates of each node before searching the best structure of GRN. It detects the breakpoint that can restrict the search space, that help in accelerating the learning process of Bayesian network.

2.6 TOOLS USED FOR GRN CONSTRUCTION

2.6.1.1 R packages

2.6.1.1 *parmigene* version 1.0.2

PARMIGENE, which stands for PARallel Mutual Information estimation for GENE NETWORK reconstruction, is used to infer large transcriptional networks using mutual information. It is an R package that implements a mutual information estimator based on k-nearest neighbour distances. This method is minimally biased compared to other methods and uses a parallel computing paradigm to reconstruct gene regulatory networks (Sales and Romuladi *et al.*, 2011). Parmigene seems to give more precise results compared to existing softwares with less computational costs.

The package along with reference manual is available at:

<https://cran.r-project.org/web/packages/parmigene/index.html>

2.6.1.2 wgcna version 1.63

The Weighted correlation network analysis (WGCNA) (Langfelder *et al.*, 2008), an R software package is a comprehensive collection of R functions that could perform various aspects of weighted correlation network analysis. This package includes functions for data simulation, network construction, gene selection, calculations of topological properties, module detection, visualization, and interfacing with external software. The package along with reference manual is available at:

<https://cran.r-project.org/web/packages/WGCNA/index.html>

2.6.1.3 GeneNet version 1.2.13

GeneNet is a package that analyses gene expression (time series) data with focus on the inference of gene networks with high accuracy (Opgen-Rhein *et al.*, 2006). It is computationally efficient and is appropriate for large scale data sets. The approach is based on dynamical correlation and covariance and it provides a similarity score for pairs of groups of randomly sampled curves.

The package along with reference manual is available at:

<https://cran.r-project.org/web/packages/GeneNet/index.html>

2.6.1.4 CoDiNA version 1.1

Co-expression Differential Network Analysis (CoDiNA) distinguishes between links that are common to all networks, links that are specific to only one of the compared networks, and links that are different in that their sign changes between networks. Basically, the package works based on a statistical framework that normalizes these different categories. The method identifies edges and links that are specific, differentiated or common to all networks, and it also includes an interactive tool for network visualization (Gysi *et al.*, 2018).

The package along with reference manual is available at:

<https://cran.r-project.org/web/packages/CoDiNA/index.html>

2.7 NETWORK VISUALIZATION TOOLS

2.7.1 Cytoscape

34

Cytoscape is a general-purpose open source software where biomolecular interaction networks with high-throughput expression data and other molecular states is integrated into a unified conceptual framework. Basic features such as network layout and mapping of the data attributes to visual display properties etc, could be handled by the Cytoscape core facility. The major functions include integration of the network with expression profiles, phenotypes, and other molecular states, linking the network to databases of functional annotations, and to layout and query the network (Shannon *et al.*, 2003). Cytoscape seems to be more powerful when used in conjunction with large databases of protein–protein interaction, protein–DNA interaction, and genetic interactions that are increasingly available for humans and model organisms.

2.7.2 VANTED

VANTED is a very important tool for modern biological research and plays a major role in analysing and interpreting biochemical data. Its functions include network loading and editing, importing any type of biochemical data (e. g. transcript, protein, metabolite) from different growth conditions and time-points, mapping of the data on the corresponding dynamic networks etc. A wide range of tasks including data visualization, network reconstruction, integration of various data types, network simulation, data exploration that serves to set the systems biology standards for visualization and data exchange (Rohn *et al.*, 2012). VANTED is a Java Web Start application that is platform-independent and available free of charge. It provides tremendous opportunities for visual exploration, statistical calculations (*t*-test, outlier identification, correlation analysis) (David *et al.*, 2014), data clustering with self-organizing maps, and much more. The various file formats supported by VANTED includes native formats like GML, GraphML, DAT (Kamp *et al.*, 2006), SBGN-ML (provided by the SBGN-ED add-on) and BioPAX. It also computes several topological properties like shortest paths between node pairs, network cycles and motifs.

2.7.3 Gephi

35

Gephi is an open source visualization module that uses a special 3D render engine to display large networks in real-time and to speed up the exploration. It serves to provide easy and broad access to network data allowing for spatializing, filtering, navigating, manipulating and clustering (Bastian *et al.*, 2009). It is exclusively used for the purpose of graph and network analysis. It is a network exploration and manipulation software that can import, visualize, spatialize, filter, manipulate and export all types of available networks. The technique involves the use of a computer graphic card, as video games do, and leaves the CPU free for other computing. As being built on a multi-task model, it makes use of its multi-core processors and hence deal with large networks (i.e. over 20,000 nodes). The extra features of Gephi includes that its node design can be personalized i.e. instead of a attaining a classical shape it can be a texture, a panel or a photo and also the graph window allows the highly configurable layout algorithms to be run in real-time.

2.7.4 BisoGenet

BisoGenet, a client server based multi-tier application is used for visualization and analysis of biomolecular relationships. It creates, visualizes and analyses biological networks and is designed according to a multi-tier architecture (Martin *et al.*, 2010). The system is constituted by three tiers: the data, the server and the client subsystems. It consists of an in-house database that stores genomics information, protein-DNA interactions, protein-protein interactions, gene ontology and metabolic pathways. It is a fast and user-friendly application which uses coding relations to distinguish between genes and their products. It creates, visualizes and analyses biological networks depending on the biological information provided by SysBiomics, an in-house database that integrated a wide range of omics information from multiple public data sources. It works as a Cytoscape plugin, that can be an easy interface for querying the server along with graph topology analysis and options for easy visualization and interpretation.

2.7.5 iDREM

The Dynamic Regulatory Events Miner (DREM) software is used to reconstruct dynamic regulatory network by integrating static protein-DNA interaction data with time series gene expression data. This would enable the user to interactively visualize the resulting model (Ding *et al.*, 2018). iDREM implements its regulatory model prediction part in Java and the interactive visualization part is implemented in Javascript with D3.js and Google charts. The users will need to only run the java program `idrem.jar` to get all results including the interactive visualization.

2.8 NETWORK VALIDATION STRATEGIES

2.8.1 Network Comparison Test (NCT)

Network Comparison Test (NCT) uses permutation tests for the comparison of network structures from two independent cross-sectional data sets (Borkulo *et al.*, 2017). The method is currently implemented for handling networks derived from continuous and binary data. The empirical dataset used is selected based on invariance in three parameters- Network structure, edge (connection) strength and global strength.

The R package for Network Comparison Test (Burkulo *et al.*, 2018) is available at: <https://cran.r-project.org/web/packages/NetworkComparisonTest/index.html>.

2.8.2 Module Validation Approaches

2.8.2.1 Topology based Approaches (TBA)

In this method, several topological features such as connectivity (Dong *et al.*, 2007), modularity (Newman *et al.*, 2006), clustering coefficient, degree, density (Georgii *et al.*, 2009), edge betweenness etc are focussed and the presence of modular structure for the identified modules is determined. The validity of a module is determined by a single or composite topological index. Uniform distribution of data will increase the value of entropy hence, a good quality module is expected to have a low entropy (Rau *et al.*, 2013).

2.8.2.2 *Statistics- Based Approaches (SBA)*

37

The module's stability, phenotypic correlation or significance of consistency is assessed in this approach. A binary or mixed integer linear programming models can be used for the validation of causal or dependent relations between network modules and biological phenotypes for module biomarker identification (Shi *et al*, 2010).

MATERIALS AND METHODS

3. MATERIALS AND METHODS

The study entitled “Comparative evaluation of tools for Gene Regulatory Network prediction and network reconstruction using genomic data” was carried out at the Section of Extension and Social Sciences, ICAR-Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram during 2017-2018. In this chapter, details regarding experimental materials and methodology used in the study are elaborated.

3.1 CONSTRUCTION OF GENOMIC DATASETS

3.1.1 Collection of Cassava genome resources

The Cassava genome resource used for the study was obtained from Phytozome, the Plant Comparative Genomics portal of the Department of Energy's Joint Genome Institute. In the latest release v12.1.6, Phytozome hosts 93 assembled and annotated genomes, from 82 Viridiplantae species of which more than half of the genomes have been sequenced, assembled and/or annotated with JGI Plant Science program resources. The selected sequence data were generated from a partially inbred (third generation self, or S3, of MCOL1505) line called AM560-2 which was generated at CIAT (International Center for Tropical Agriculture) in Cali, Colombia. The whole genome assembly (approximately 582.25 Mb arranged on 18 chromosomes plus 2,001 scaffolds) and whole genome annotation (33,033 genes) of AM560-2 genotype of *Manihot esculenta* v6.1 (Cassava) were downloaded from Phytozome v12.1. (<https://phytozome.jgi.doe.gov/pz/portal.html>) (Bredeson *et al.*, 2016).

3.1.2 Identification of immunity related genes

Genes coding for canonical immune protein domains which includes WRKY, TIR, LRR, Kinase, NBS, LysM, Lectin, NB-ARC etc. that are related to MAMP triggered and ETI triggered immunity in cassava were downloaded from Pfam 30.0 (<https://pfam.xfam.org/>) (Finn *et al.*, 2016). These domains were searched in proteomes of cassava (*Mesculenta* 305 v6.1 protein.fa.gz) using HMMER v3.1b2 (Finn *et al.*, 2015) by generating Hidden Markov Model (HMM) corresponding to

the several selected Pfam families. The entire workflow for the reconstruction of regulatory network using genomic data is depicted in Figure 4.

3.1.3 Protein domain search and analysis- HMMER suite version v3.1b2

HMMER mainly serves the purpose of searching sequence databases for homologs of protein or DNA sequences, and to make sequence alignments. It is a free, commonly used software package that carries out Bio sequence analysis by generating profile Hidden Markov Models. Several programs included in the HMMER v3.1b2 suite provides the core functionality for protein domain analysis and annotation pipelines.

Steps for installation:

- Download hmmer-3.1b2.tar.gz from <http://hmmer.org/>
- Unpack it

```
> wget ftp://selab.janelia.org/pub/software/hmmer3/3.1b2/hmmer-3.1b2.tar.gz
```

```
> tar xf hmmer-3.1b2.tar.gz
```

```
> cd hmmer-3.1b2
```

```
> ./configure
```

```
> make install
```

3.1.3.1 *hmmbuild*- Building profile HMM

Synopsis: `hmmbuild [options]<hmmfile_out><msafile>`

<hmmfile_out> represents output file name and <msafile> represents multiple sequence alignment file. A multiple sequence alignment file (*msa file*) generated using *Clustal omega* (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) online with “.clustal” extension, is used to create a hmm file.

3.1.3.2 *hmmsearch*- Searches profile HMM against protein database

Synopsis: `hmmsearch [options] <hmmfile><seqdb>`

<seqdb> refers to the sequence database. All sequence hits with an E value $< 1 \times 10^{-20}$ were selected to acquire a high-quality cassava protein set conferring immunity using the raw profile HMM

3.1.3.3 *hmmalign*- Align sequences to profile HMM

Synopsis: *hmmalign* [options]<hmmfile><seqfile>

hmmalign performs a multiple sequence alignment by aligning all the sequences individually to the profile HMM to obtain output in Stockholm format. It creates a high confidence immunity related gene set in cassava which can be used to build cassava immunity specific hmm (using *hmmbuild*). This is further applied into whole genome annotation to improve the sensitivity of the method by pre-checking the location of the sequence in cassava genome. This new cassava-specific HMM was used and proteins with a reporting threshold (E-value) less than 0.01 was selected and used for further analysis. Several other parameters like gap penalties, filter etc. were also considered.

3.1.4 Filtering genes for high competence cassava specificity

3.1.4.1 *Manual curation*

Each set of proteins were separately checked for homology with other plant R genes and genes unrelated to immunity and less than 250 amino acids were filtered. A plant R gene database was constructed with already identified immunity related genes and other plant R genes and BLASTP was performed in the ubuntu terminal against the duplication removed R gene database. The sequence IDs were sorted and the output was retrieved in FASTA format.

3.1.4.2 *PRGDB 3.0*

Online BLASTP was performed against the PR-proteins specific to *Manihot esculenta* stored in Pathogen Resistance Genes Database (PRGdb 3.0) (Osuna-Cruz *et al.*, 2017). Both the results were compared and analysed for obtaining high confidence immunity related protein dataset in cassava.

3.1.5 Mapping and functional annotation

Blast2GO 5 is a comprehensive bioinformatics software for the functional annotation and analysis of genome scale sequence datasets (Conesa and Gotz *et al.*, 2008). It is meant to provide a user-friendly interface for Gene Ontology annotation.

Blast2GO 5 uses CloudBlast, which is a high-performance, cost-optimized and secure method for massive sequence alignment tasks. It allows you to execute standard NCBI Blast+ searches directly from Blast2GO PRO in a dedicated computing cloud. Gene Ontology Mapping performed retrieves the GO terms associated to the Hits obtained by the BLAST search. Gene Ontology Annotation selects the GO terms from the GO pool obtained by the Mapping step and assigns them to query sequences.

InterPro annotations in Blast2GO 5, functions to retrieve domain/motif information in a sequence wise manner and also transfers the corresponding GO terms to the sequences and merges it with already existing GO terms.

3.1.6 Detection of molecular function and pathway mapping

The identified genes were then further annotated in InterPro, Gene Ontology (GO) (Figure 10), EC (Enzyme Code), and Kyoto Encyclopedia of Genes and Genomes (KEGG) database with Blast2GO 5.

Blast2GO 5 provides EC annotation in which the sequences with GO annotations will eventually show EC numbers so that the GO annotation accuracy can be made extensive to Enzyme annotations. An extensive search of all KEGG maps containing the EC numbers of the selected sequences will be done, which would allow the display of enzymatic functions (Figure 11) in the context of the metabolic pathways in which the particular protein participates.

3.2 PREDICTION AND ANALYSIS OF GRN USING DIFFERENT COMPUTATIONAL METHODS

3.2.1 Different computational approaches for GRN construction

3.2.1.1 Weighted Gene Co-expression Network Analysis (WGCNA)

Weighted Gene Co-expression Network Analysis (WGCNA) describes the gene pattern correlations across microarray samples. It is used to find clusters (modules) of genes that are highly correlated, to summarize such clusters using the module eigengene or an intramodular hub gene, to relate modules to one another and to external sample traits (using eigengene network methodology), and to calculate module membership measures. Correlation networks helps for network-based gene screening methods that helps us to identify candidate biomarkers or therapeutic targets.

The construction of correlation networks is on the basis of correlations between quantitative measurements that can be described by an $n \times m$ matrix, $X = [x_{ij}]$ where, the row indices represents network nodes ($i=1, \dots, n$) and the column indices ($l=1, \dots, m$) represents sample measurements:

$$X = [x_{ij}] = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

The i -th row x_i is referred to as the i -th node profile across m sample measurements. A quantitative measure, referred to as sample trait, T which is a vector with m components that correspond to the columns of the data matrix X is used to define a node significance measure.

$$T:GS_i = |\text{cor}(x_i, T)|$$

A p-value based node significance measure can be defined by a correlation test p-value or a regression-based p-value. It is used for assessing the statistical significance between x_i and the sample trait T .

$$GS_i = -\log P_i$$

The correlation network methodology uses network language to describe the pairwise relationships (correlations) between the rows of X .

3.2.1.1.1 *Tool used for WGCNA*

The WGCNA R software package (Langfelder *et al.*, 2008) was used. The package along with its source code and additional material are freely available at: <https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/>

The various steps involved in WGCNA using R software is as follows:

- i Cleaning and input of data
- ii Construction of network and detection of module
- iii Relating modules and identification of important genes
- iv Network interface analysis with other data such as functional annotation and gene ontology
- v Network visualization using WGCNA functions
- vi Export of networks to external software

3.2.1.2 *Bayesian K2 Algorithm*

Bayesian K2 presents a Bayesian algorithm for the construction of a probabilistic network from a database (Cooper and Herskovits *et al.*, 1992). The advantage of using Bayesian network is that it is capable of dealing with the noise in experimental measurements and it has the ability to handle the missing data and incomplete knowledge about the biological system. The algorithm begins by assuming that a node has no parents, and then starts adding parents to that node, in which the addition better improves the probability of the resulting network. It reduces the computational complexity by the requirement of a prior ordering of nodes as an input, from which the network structure will be constructed. Algorithm K2 uses greedy search as the search strategy. The ordering of nodes in the algorithm is very important because this technique only considers the nodes that have already been filled with parents as the parent nodes. Therefore, the first node to be considered

will always be empty of parents and the second one will only be able to have the first node as the parent node. Log metrics such as MDL, AIC and Entropy need to be minimized instead of maximized. In order to do that, when the algorithm is applied, the sign of these metrics has to be changed without changing the algorithm.

3.2.1.2.1 Tool used for Bayesian K2

Cytoscape Network Inference Toolbox (Cyni), a Cytoscape application that functions in version 3.0 or newer versions and made available in the Cy3 App Store is used (<http://proteomics.fr/Sysbio/CyniProject>).

3.2.1.3 *Mutual Information based*

Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) is a novel algorithm based on information-theory which mainly serves the purpose of reverse engineering the transcriptional networks from microarray data (Margolin *et al.*, 2006). Here the candidate interactions are identified by estimating pairwise gene expression profile mutual information. Mutual information measures the degree of statistical dependency between two variables.

For a pair of random variables, x and y , Mutual information could be defined as,

$$I(x, y) = S(x) + S(y) - S(x, y)$$

where $S(t)$ represents the entropy of an arbitrary variable t .

For network construction, the mutual information for each pair of elements is computed and the two nodes representing the two biological elements would get connected in the network if their mutual information value is above a certain particular threshold. If the value lies below the threshold, it will remain unconnected. In this algorithm, discrete data is used to calculate the mutual information, hence, all the continuous data will have to be discretized before using.

3.2.1.3.1 Tool used for Mutual Information based

Cytoscape Network Inference Toolbox (Cyni), a Cytoscape application that functions in version 3.0 or newer versions and made available in the Cy3 App Store

is used (<http://proteomics.fr/Sysbio/CyniProject>). To calculate mutual information, the entropy used is joint entropy and the log base is 2.

3.2.1.4 Basic Correlation Algorithm

This algorithm is mainly used to explain the observed correlations between biological elements such as genes by the presence of other biological elements. Here, the networks are inferred by computation of similarity measures for each pair of elements. If the similarity value is above a certain threshold, the two biological elements represented by two nodes get connected in the network and if the value is below threshold, it remains unconnected.

3.2.1.4.1 Tool used for Basic correlation algorithm

Cytoscape Network Inference Toolbox (Cyni), a Cytoscape application that functions in version 3.0 or newer versions and made available in the Cy3 App Store is used (<http://proteomics.fr/Sysbio/CyniProject>). The list of correlation metrics that are made available in Cyni so far includes:

- i **Pearson Correlation:** Here, the value is obtained by division of the covariance of the two variables by the product of their standard deviations.
- ii **Spearman's rank Correlation:** This is a non-parametric measure which represents the correlation between the two rows of data. In this metric, the current value row values are replaced by their ranks and then the liner correlation coefficient is applied to this data.
- iii **Kendall's Tau Correlation:** This is also a non-parametric measure. Here, instead of using the numerical difference of ranks, only the relative ordering of the ranks is used.

3.2.2 Generation of simulated dataset

SynTReN is used for the generation of synthetic gene expression data for the design and analysis of structure learning algorithm. The benchmark data set for which the underlying network is known, used for the validation of the constructed networks is generated using SynTReN (Van den Bulcke *et al.*, 2006).

3.2.3 Comparison of different methods

47

Accuracy studies of a particular regulatory network construction strategy addresses how well they determine the particular interaction. Sensitivity, Specificity, Predictive values and likelihood ratios (LRs) are all different ways of expression of the performance of a particular method. Accuracy studies are conducted by comparing the network generated by each method with a reference standard consisting of actual interactions. A Table (Table 1) is created with the index results on one side of the table and those of the reference standard on the other side.

Table 1 Parameters for determining interaction statistics

TEST	Present	Absent	TOTAL
Positive	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)	TP+FP
Negative	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)	FN+TN
TOTAL	TP+FN	FP+TN	

The following values are calculated for measuring the performance:

- Sensitivity refers to the probability that the presence of the interaction will give a positive test result (%).

$$\text{SENSITIVITY} = \text{TP} / (\text{TP} + \text{FN})$$

- Specificity refers to the probability that the absence of an interaction will give a negative test result (%).

$$\text{SPECIFICITY} = \text{TN} / (\text{FP} + \text{TN})$$

- Positive Likelihood Ratio (PLR) is the ratio between the probability of obtaining a positive test result in the presence of the interaction and the probability of obtaining positive test result in the absence of the interaction

$$\text{PLR} = \text{True Positive Rate} / \text{False Positive Rate} = \text{Sensitivity} / (1 - \text{Specificity})$$

- Negative Likelihood Ratio (NLR) is the ratio between the probability of obtaining a negative test result in the presence of the interaction and the probability of obtaining a negative test result in the absence of the interaction

$$\text{NLR} = \frac{\text{False Negative Rate}}{\text{True Negative Rate}} = \frac{1 - \text{Sensitivity}}{\text{Specificity}}$$
- Positive Predictive Value (PPV) is the probability of getting a positive test result in the presence of the interaction (%)

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
- Negative Predictive Value (NPV) is the probability of getting a negative test result in the absence of the interaction (%)

$$\text{NPV} = \frac{\text{FN}}{\text{FN} + \text{TN}}$$

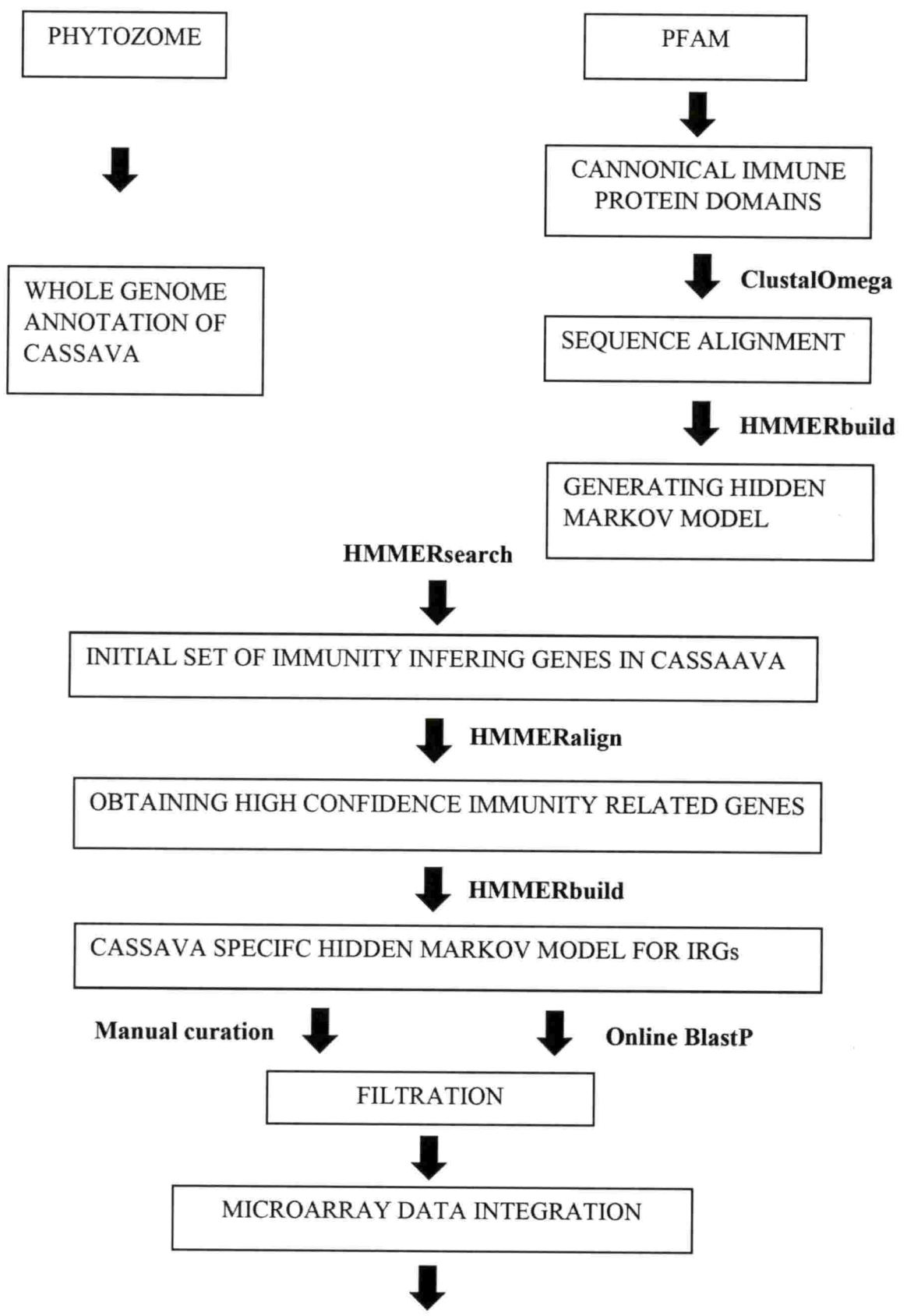
3.3 PROTEIN-PROTEIN INTERACTION NETWORK CONSTRUCTION

The STRING v10.5 (<https://string-db.org/cgi/input.pl>) was used to predict the protein-protein association data of the identified genes. It collects and reassesses available experimental data on protein-protein interactions available in the database, and imports the known pathways and protein complexes from other curated databases. Statistical analysis results are also by default obtained in STRING v10.5 (Szklarczyk *et al.*, 2017). The enrichment tests done for a variety of classification systems (Gene Ontology, KEGG, Pfam and InterPro) seems to provide functional characterization of the set of protein. They also seem to employ Fisher's exact test followed by a correction for multiple testing.

3.4 DATA INTEGRATION AND VISUALIZATION

3.4.1 Microarray data integration

From a Cassava cDNA microarray dataset constructed based on a large cassava EST database to study the interaction incompatibility between cassava and *Xanthomonas axonopodis* pv. *manihotis* (*Xam*) strain CIO151 by Lopez *et al.* (2005), a total of 199 genes were found to be differentially expressed (126 up-regulated and 73 down-regulated). The GenBank accession IDs of all the upregulated and down regulated genes were collected and a database was constructed. BLASTX was done using the identified immunity related genes in cassava as query against the constructed nucleotide database.



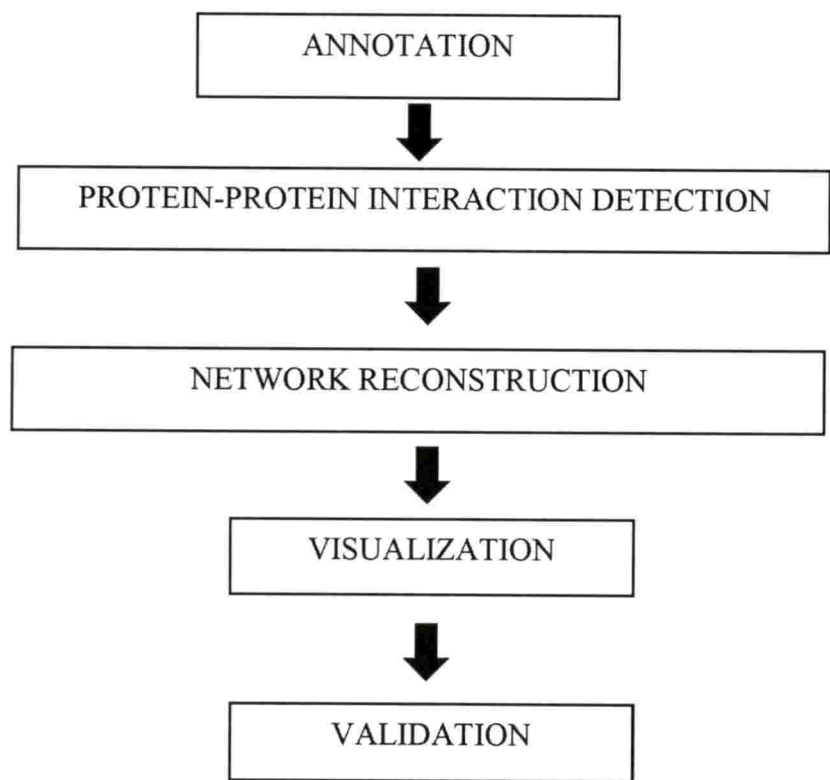


Figure 4. Workflow for the construction of regulatory network of immunity related genes in cassava

3.4.2 Network Visualization

Cytoscape 3.6.1 (<http://www.cytoscape.org/>) helps to visualize the molecular interaction networks and biological pathways, and enables the integration of these networks with gene expression profiles, annotation and other state data (Shannon *et al.*, 2003). It also paves way for changing the visual styles upon the application of algorithm for clustering, enrichment analysis, network layout and network analysis.

The interaction data is imported into Cytoscape in various formats like CSV (comma-separated values), TSV (tab-separated values) in Excel, along with network-specific formats such as SIF (simple interaction file), OpenBEL (Open Biological Expression Language) XGMML (Extensible Graph Markup and Modeling Language), BioPAX (Biological Pathway Exchange), GML (Graph Modelling Language), PSI MI (Proteomics Standards Initiative–Molecular Interaction format), and SBML (Systems Biology Markup Language). Cytoscape also provides extended service through its various plugin softwares available at Cytoscape App store. Various Cytoscape plugins include Metscape plugin which are used to generate metabolic networks based on information in the Kyoto Encyclopedia for Genes and Genomes (KEGG), BioCycPlugin, which provides access to the BioCyc metabolic network database (<http://biocyc.org/>), and ReConn, which provides access to Reactome (<http://reactome.org/>). Cytoscape integration with STRING v10.5 has paved way for explicit opportunities in analysis and visualization of large scale networks (Szklarczyk *et al.*, 2017).

RESULTS

4. RESULT

The results of the study “Comparative evaluation of tools for gene regulatory network prediction and network reconstruction using genomic data” carried out at the Section of Extension and Social Sciences, ICAR- Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram during 2017-2018 are presented in this chapter.

5.1 CONSTRUCTION OF GENOMIC DATASETS

The genomic dataset used for the study was generated by searching the canonical immune protein domains obtained from Pfam30.0 (<https://pfam.xfam.org/>) (Figure 3) in the whole genome sequence of AM560-2 genotype of *Manihot esculenta* v6.1 (Cassava), derived from Phytozomev12.1 (<https://phytozome.jgi.doe.gov/pz/portal.html>) (Figure 2) by using Hidden Markov Models (HMM). The Phytozome data was constituted by a whole genome assembly of cassava made up of 582.25 Mb sequence arranged on 18 chromosomes plus 2001 scaffolds. It consisted a total of 8348 alternatively spliced transcripts with 35.9% GC content. The cassava proteome sequence was taken from the Whole genome annotation of cassava composed of 33,033 total loci containing protein coding transcripts.

The main protein domains involved in conferring Microbe-Associated Molecular Patterns (MAMPs) triggered immunity and Effector Triggered Immunity (ETI) in plants were identified from literature. These particular domains and other selected domains related to immunity in plants were selectively downloaded from Pfam. The selected protein domains were those of ABC Transporters (PF00005), Lectin-c (PF00059), LRR1 (PF00560), LRR3 (PF07725), LRR9 (PF14580), LRR4 (PF12799), LRR5 (PF13306), LRR6 (PF13516), LRR8 (PF13855), LysM (PF01476), NB-ARC (PF00931), Nrap (PF01566), PKinase (PF00069), TIR (PF01582), and WRKY (PF03106). Table 2. gives details regarding the total number of genes present in each domain and the identified number of genes that specifically confers resistance in cassava.



Figure 2. Phytozome v12.1 database (<https://phytozome.jgi.doe.gov/pz/portal.html>)



Figure 3. Pfam database (<https://pfam.xfam.org/>)

Table 2. Screening of Immunity related genes from Immune protein domains

Immune protein domains	Pfam ID	Total No. of genes	No. of genes identified	No. of predicted interactions
ABC Trans	PF00005	36973	104	57
Lectin-C	PF00059	17879	Nil	0
LRR1	PF00560	16878	Nil	0
LRR3	PF07725	1274	Nil	0
LRR9	PF14580	5476	2	0
LRR4	PF12799	6525	221	89
LRR5	PF13306	17281	Nil	0
LRR6	PF13516	56000	16	10
LRR8	PF13855	133230	248	97
LysM	PF01476	29030	1	0
NB-ARC	PF00931	24904	207	28
Nramp	PF01566	5393	13	5
Pkinase	PF00069	236455	1320	517
TIR	PF01582	6950	40	15
WRKY	PF03106	6320	109	43

```

# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b2 (February 2015); http://hmmerr.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
#
# query HMM file:                /home/bioinfolab-5/RESHMA /data/ABC Trans/file.hmm
# target sequence database:      /home/bioinfolab-5/RESHMA /data/ABC Trans/annotation.fa
# sequence reporting threshold:  E-value <= 1e-20
#
Query:      ABC [M=147]
Scores for complete sequences (score includes all domains):
--- full sequence ---   --- best 1 domain ---   -#dom-
E-value  score  bias  E-value  score  bias  exp  N  Sequence              Description
-----
 2e-74  249.9  0.4  7.1e-37  128.1  0.1  2.5  2  cassava4.1_024510m    pacid=17991711 transcript=cassava4.1_0245
7.9e-74  248.0  0.1  1.1e-35  124.3  0.0  2.4  2  cassava4.1_000398m    pacid=17960517 transcript=cassava4.1_0003
 1e-73  247.6  0.3  9.1e-37  127.8  0.1  2.5  2  cassava4.1_000369m    pacid=17964690 transcript=cassava4.1_0003
1.6e-73  247.1  0.0  7.2e-36  124.9  0.0  2.6  2  cassava4.1_000306m    pacid=17968622 transcript=cassava4.1_0003
2.1e-73  246.7  0.5  3.1e-35  122.8  0.1  2.4  2  cassava4.1_033109m    pacid=17964654 transcript=cassava4.1_0331
2.8e-73  246.2  0.2  6.3e-37  128.3  0.0  2.6  2  cassava4.1_021429m    pacid=17993610 transcript=cassava4.1_0214
2.9e-73  246.2  0.4  1.1e-35  124.3  0.1  2.4  2  cassava4.1_029060m    pacid=17991709 transcript=cassava4.1_0290
2.9e-73  246.2  0.6  1.7e-35  123.6  0.1  2.6  2  cassava4.1_000359m    pacid=17963755 transcript=cassava4.1_0003
6.5e-73  245.0  3.4  3.3e-36  126.0  0.0  3.0  2  cassava4.1_030988m    pacid=17964040 transcript=cassava4.1_0309
8.3e-73  244.7  0.4  3.6e-36  125.8  0.1  2.5  2  cassava4.1_000345m    pacid=17963748 transcript=cassava4.1_0003
1.1e-72  244.3  0.4  9.3e-36  124.5  0.0  2.5  2  cassava4.1_001064m    pacid=17962284 transcript=cassava4.1_0010
5.6e-72  242.0  0.2  6.9e-36  124.9  0.1  2.5  2  cassava4.1_000399m    pacid=17985450 transcript=cassava4.1_0003
2.1e-71  240.1  0.1  4.3e-34  119.1  0.0  2.6  2  cassava4.1_025247m    pacid=17966639 transcript=cassava4.1_0252
2.8e-70  236.5  0.0  1e-33  117.9  0.0  2.6  2  cassava4.1_000384m    pacid=17984309 transcript=cassava4.1_0003
1.3e-69  234.4  0.0  3.4e-33  116.2  0.0  2.6  2  cassava4.1_033075m    pacid=17959721 transcript=cassava4.1_0330
3.1e-69  233.1  0.0  5.3e-33  115.6  0.0  2.7  2  cassava4.1_000386m    pacid=17977806 transcript=cassava4.1_0003
3.2e-69  233.1  0.1  4e-34  119.2  0.0  2.6  2  cassava4.1_026648m    pacid=17992500 transcript=cassava4.1_0266
 5e-69  232.4  0.0  6.7e-33  115.2  0.0  2.7  2  cassava4.1_000385m    pacid=17975374 transcript=cassava4.1_0003
6.1e-69  232.2  0.0  1.8e-33  117.1  0.0  2.6  2  cassava4.1_000410m    pacid=17964245 transcript=cassava4.1_0004
5.5e-66  222.6  0.1  6.7e-32  112.0  0.0  2.5  2  cassava4.1_000409m    pacid=17985941 transcript=cassava4.1_0004
2.7e-61  207.3  0.6  2.4e-36  126.4  0.2  2.5  2  cassava4.1_034121m    pacid=17964955 transcript=cassava4.1_0341
2.7e-59  200.8  1.3  9.6e-31  108.2  0.0  3.3  2  cassava4.1_002469m    pacid=17959752 transcript=cassava4.1_0024
 4e-59  200.3  1.4  2.1e-30  107.1  0.0  2.9  2  cassava4.1_002838m    pacid=17983025 transcript=cassava4.1_0028
9.9e-58  195.8  1.7  9.5e-28  98.5  0.0  3.2  2  cassava4.1_002648m    pacid=17989132 transcript=cassava4.1_0026
1.1e-56  192.4  0.0  7.8e-31  108.5  0.0  2.4  2  cassava4.1_000427m    pacid=17989065 transcript=cassava4.1_0004
2.1e-56  191.5  0.1  4.2e-28  99.7  0.0  2.7  2  cassava4.1_000219m    pacid=17960822 transcript=cassava4.1_0002
7.4e-56  189.7  0.0  1.8e-29  104.1  0.0  2.4  2  cassava4.1_000424m    pacid=17970289 transcript=cassava4.1_0004
4.4e-54  183.9  0.3  6.3e-29  102.3  0.0  3.0  2  cassava4.1_004081m    pacid=17968002 transcript=cassava4.1_0040
3.4e-53  181.6  0.1  7.0e-37  95.5  0.1  2.4  2  cassava4.1_000315m    pacid=17975111 transcript=cassava4.1_0003

```

Figure 6. A model of the **hmmsearch** result

HMMER v3.1b2 that makes Hidden Markov Model (HMM) corresponding to the selected Pfam families were used to search the occurrence of selected immune domains in the cassava proteome. A Hidden Markov model (Figure 8) of the pfam domain data was created separately for each domains and screening was carried out using **hmmsearch** (Figure 6), **hmmalign** and **hmmbuild**. The domain annotation for each sequence as well as the alignments for each domain (Figure 5) was acquired. The characteristic parameters of the Hidden Markov Model initially created is given in Table 3.

Domain annotation for each sequence (and alignments):

```
>> cassava4.1_006823m pacid=17963166 transcript=cassava4.1_006823m
locus=cassava4.1_006823m.g ID=cassava4.1_006823m.v4.
# score bias c-Evaluei-Evaluehmmifrom hmm to alifromali to envfrom env to acc
-----
1! 95.2 4.2 1.3e-29 4e-27 1 58 [] 191 247.. 191 247.. 0.97
2! 94.2 4.7 2.6e-29 8.2e-27 1 58 [] 363 420.. 363 420.. 0.98
```

Alignments for each domain:

== domain 1 score: 95.2 bits; conditional E-value: 1.3e-29

```
wrky 1 dDgyqWrKYGqkkikgskfprsYYrCthqgCpakKqVqrsdedpsvlevtYegeHthp 58
dDgy+WrKYGqk++kgs+fprsYY+Cth++Cp+kK+V+rs d++v+e++Y+g+H+h+
cassava4.1_006823m 191 DDGYNWRKYGQKQVKGSEFPRSYYKCTHPSPVKKKVERSL-
DGQVTEIYYKGQHNHQ 247
7*****9.667*****5 PP
```

== domain 2 score: 94.2 bits; conditional E-value: 2.6e-29

```
wrky 1 dDgyqWrKYGqkkikgskfprsYYrCthqgCpakKqVqrsdedpsvlevtYegeHthp 58
dDgy+WrKYGqk +kg+++prsYY+Ct+++gC+++K+V+r ++dp+++++tYeg+H+h+
cassava4.1_006823m 363
DDGYRWRKYGQKVVKGNPYPRSYYKCTTSGCTVRKHVERAATDPRAVITTYEGKHND
420
7*****5 PP
```

Figure 5. Domain annotation and Alignments in HMMER v3.1b2

Table 3. Characteristic parameters of Hidden Markov Model generated

DOMAIN NAME	MODEL LENGTH	NSEQ	EFFN	CKSUM
ABC-Tran	147	55	4.273376	2707433464
Lectin-C	111	57	4.783630	8927388
LRR1	16	2407	2407.000000	2159447960
LRR3	20	56	56.000000	1928672352
LRR4	45	276	276.000000	4106556335
LRR9	175	8	0.714844	3904491559
LRR5	129	165	5.294724	1114312103
LRR6	24	80	80.000000	2192239453
LRR8	61	63	15.615417	1728073999
LysM	46	155	27.931976	2347244879
NB-ARC	289	9	1.937988	3208044688
Nramp	355	92	4.152466	3994122721
PKinase	257	38	2.593018	808671746
TIR	172	24	2.944336	490769738
WRKY	58	305	3.290329	2758810124

```

HMMER3/f[3.1b2 | February 2015]
NAME LRR3
LENG 20
ALPH amino
RF no
MM no
CONS yes
CS no
MAP yes
DATE Wed Mar 21 09:33:02 2018
NSEQ 56
EFFN 56.000000
CKSUM 1928672352
STATS LOCAL MSV -6.1661 0.72777
STATS LOCAL VITERBI -6.5033 0.72777
STATS LOCAL FORWARD -4.1027 0.72777
HMM A C D E F G H I K L M N P
Q R S T V W Y
m->m m->i m->d i->m i->i d->m d->d
COMPO 4.07240 4.45323 3.39042 2.35533 4.02443 2.97498 3.59148 3.36997 2.31458
1.66830 3.43509 2.90708 3.98677 3.09032 3.35652 2.71783 3.36669 2.72852 3.08032 3.61275
2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690
2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
0.04003 7.62288 3.25058 0.61958 0.77255 0.00000 *
1 4.68164 3.12360 2.60234 3.49294 2.08448 5.45485 2.18923 6.02609 2.25450 5.49930
6.23430 1.27135 5.84839 3.08808 3.20842 4.64358 3.59918 5.58491 7.62581 2.55583 1 n
---
2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690
2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
0.00076 7.58360 8.30595 0.61958 0.77255 0.46255 0.99337
2 8.59541 9.24832 9.01107 9.02068 7.92651 8.11683 9.46891 7.42265 9.15923 0.00370
7.78773 9.37551 8.65187 9.22207 8.89497 9.27028 8.87411 7.83868 9.53767 9.07040 2 L
---
2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690
2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
0.00074 7.60889 8.33123 0.61958 0.77255 0.28758 1.38660
3 2.26483 6.87469 7.65744 7.10661 3.39210 6.97062 7.44308 2.56632 3.38302 3.77901
6.05469 7.14002 7.28120 7.13054 7.00821 6.32591 3.48573 0.37763 7.88804 6.68966 3 V
---
2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690
2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
0.00073 7.62288 8.34522 0.61958 0.77255 0.48576 0.95510
4 4.71900 7.22807 1.46011 1.06201 3.29817 5.49218 3.65905 2.63927 2.22181 5.53297
6.26898 3.05042 5.88598 4.77628 3.05483 4.68148 3.60428 4.17788 7.66096
//

```

Figure 8. A Model of the generated HMM file

Further filtration of the genes for cassava specificity was done by creating a R gene database manually by incorporating already identified immunity related genes and R genes from other plants. Blast p was performed against this resistance database and the sequences with lower E-value as selected and proteins made up of less than 250 amino acid sequences were omitted. This result as compared with the results obtained through online Blastp in PRGDB to make a final list of immunity related genes in cassava. Further the identified genes were again Blasted, Mapped and Annotated in Blast2go to determine the process or pathway in which the identified genes are involved.

A total of 1919 immunity related genes were identified (Figure 7) in cassava out of which 22 of them seems to specifically offer virus resistance in cassava (Table 4). The identified genes were predicted to be involved in various biological processes (Figure 9) like intracellular signal transduction, oxidation-reduction process, transmembrane transport, regulation of catalytic activity etc. Most of the genes functions to be involved in protein phosphorylation and very few in response to stress. The sequence distribution data predicted the localisation of the greatest number of genes in the membrane and a very few in the cytoplasm (Figure 13). The major metabolic pathways (Figure 12) in which the predicted genes are involved are purine metabolism, thiamine metabolism, aminobenzoate degradation, Th1 and Th2 cell differentiation, T cell receptor signalling pathway, biosynthesis of antibiotics, pentose phosphate pathway and Glutathione metabolism.

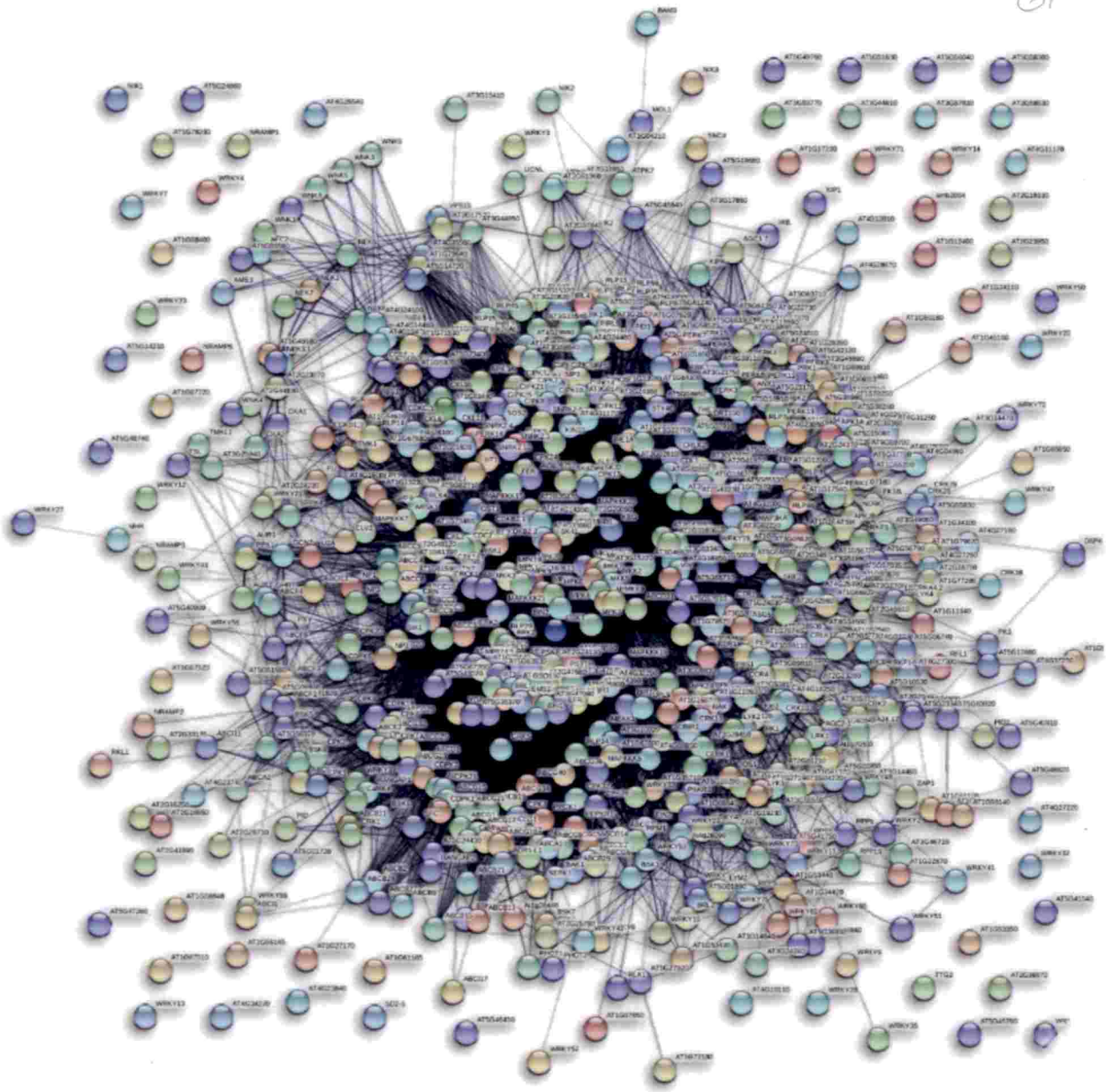


Figure 7. Interactions of the predicted genes obtained from STRING v10.5

Sequence Distribution [Biological Process]

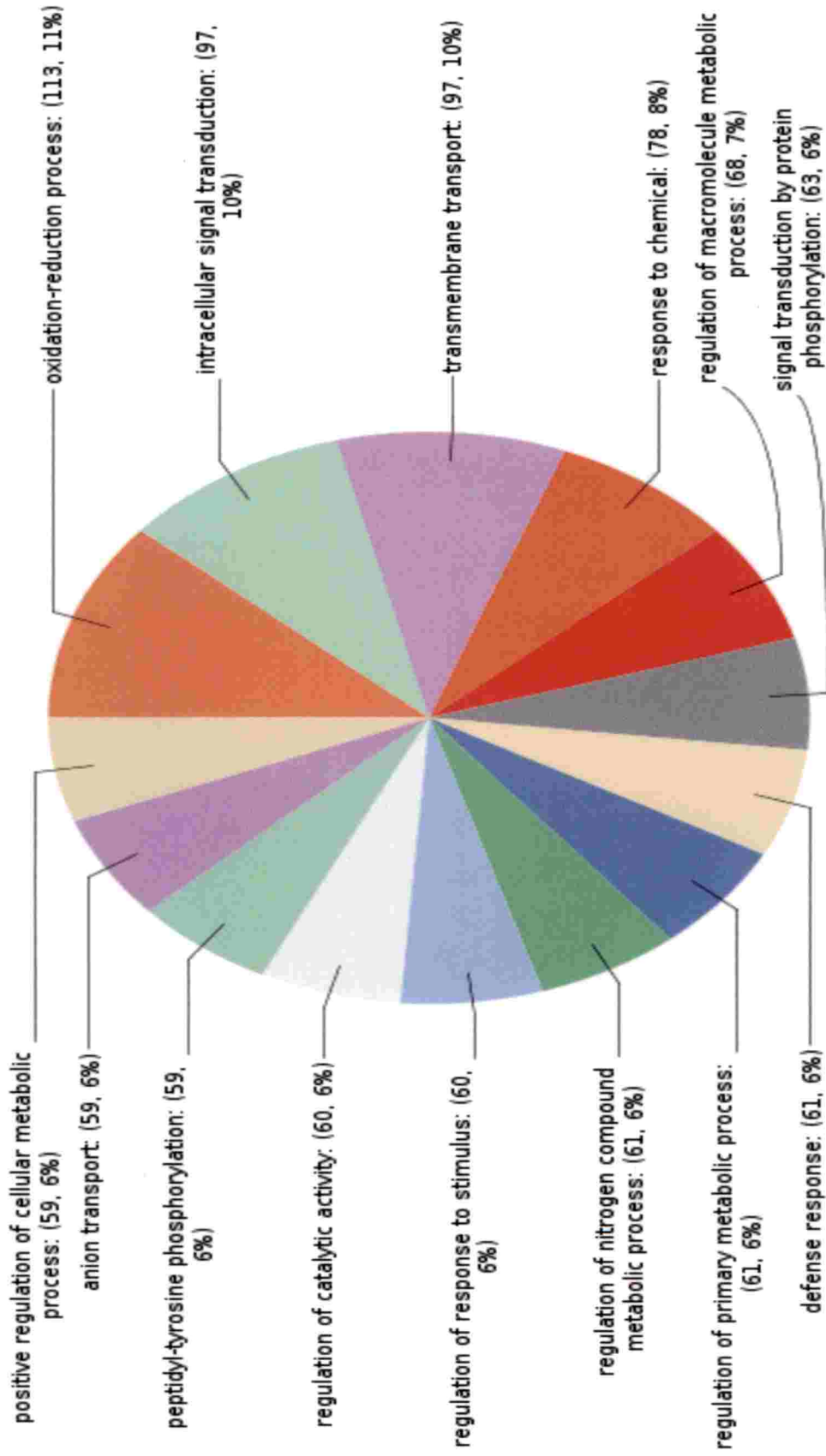


Figure 9. Distribution of sequences corresponding to Biological Process

Direct GO Count (BP) [reshma]

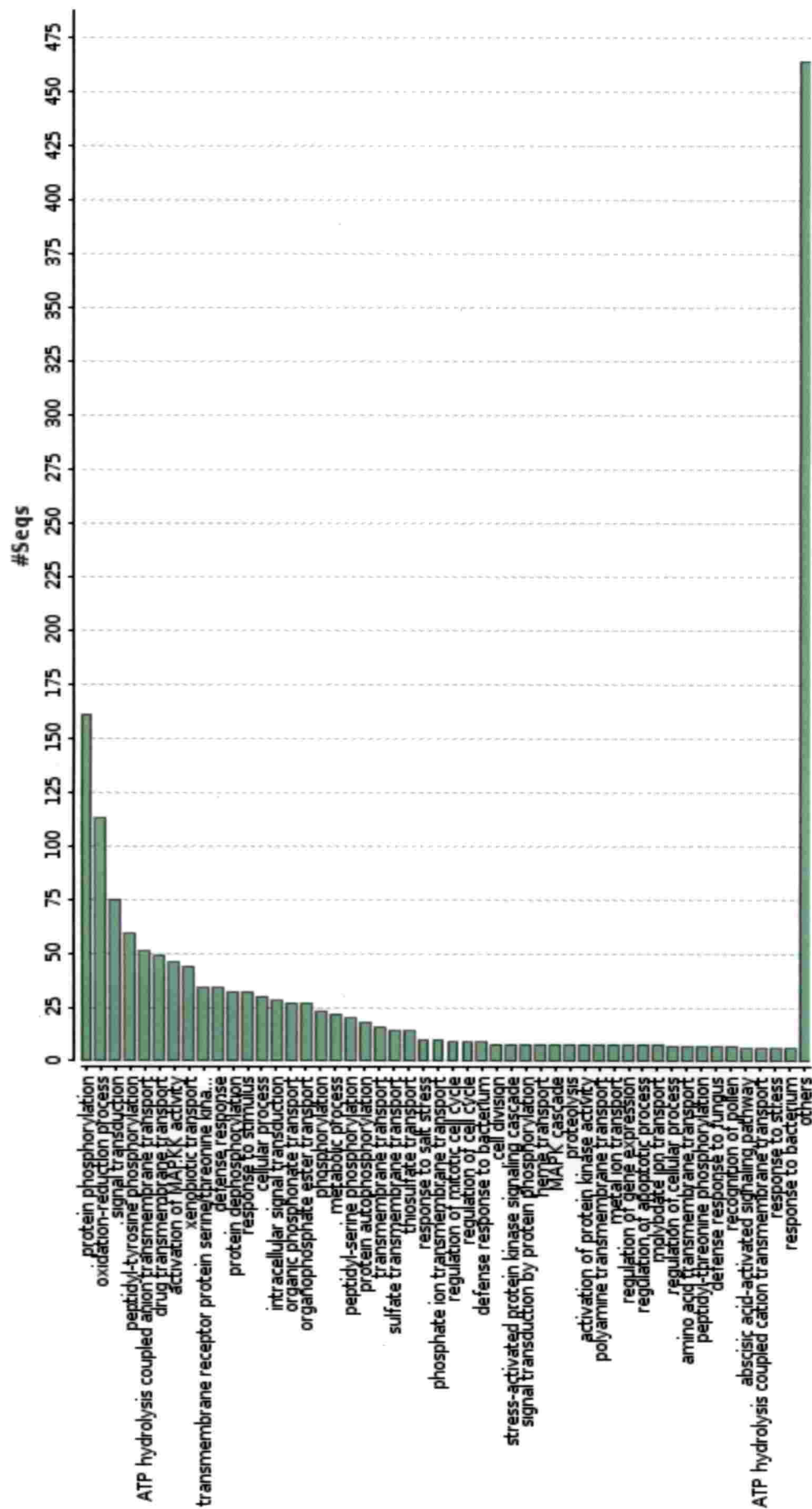


Figure 10. Gene Ontology Mapping result

Graph Level 5 Pie Chart of #Seqs [Molecular Function]

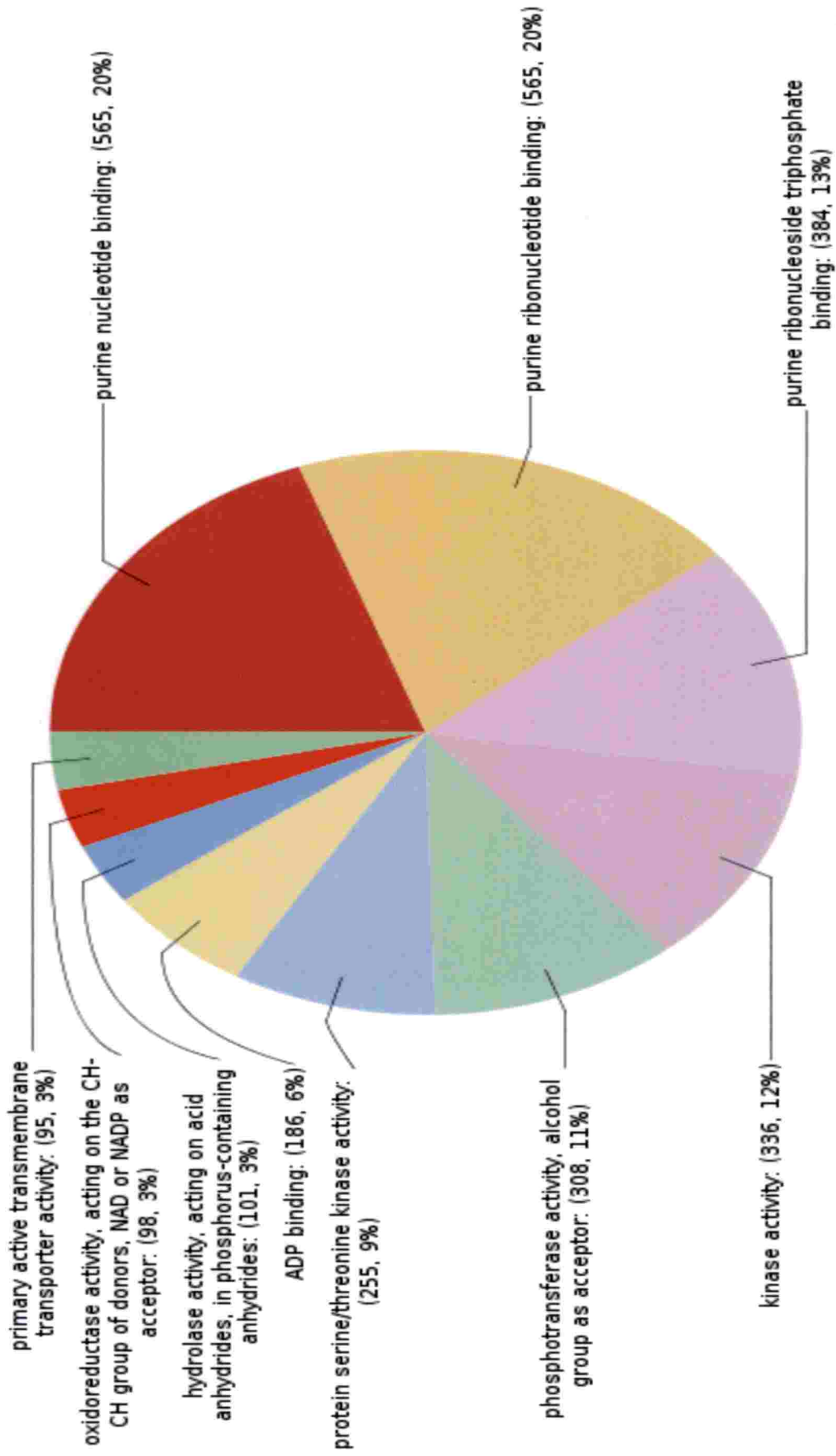


Figure 11. Distribution of sequences according to Molecular Function

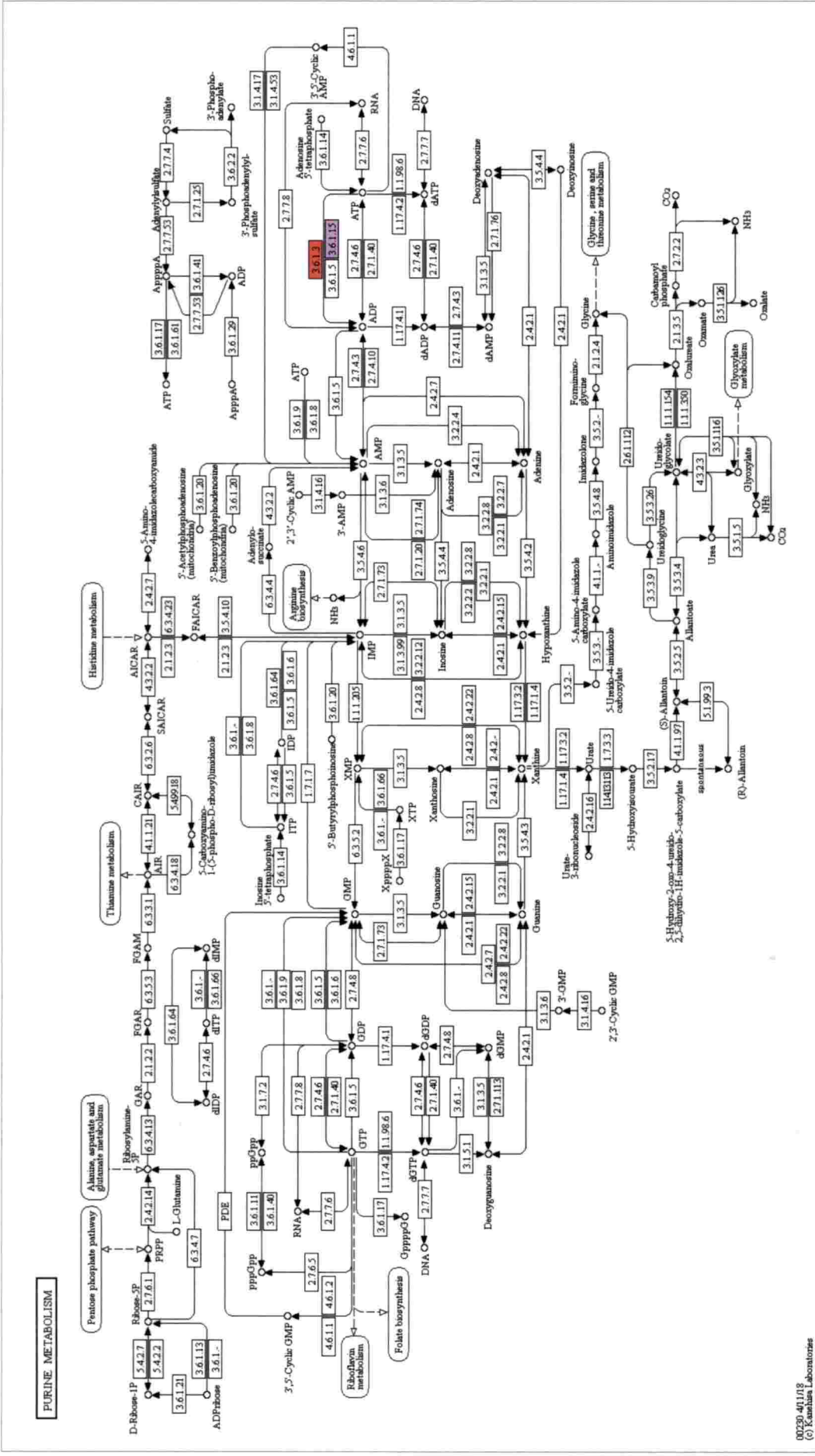


Figure 12. KEGG Pathway Mapping of identified sample identified gene

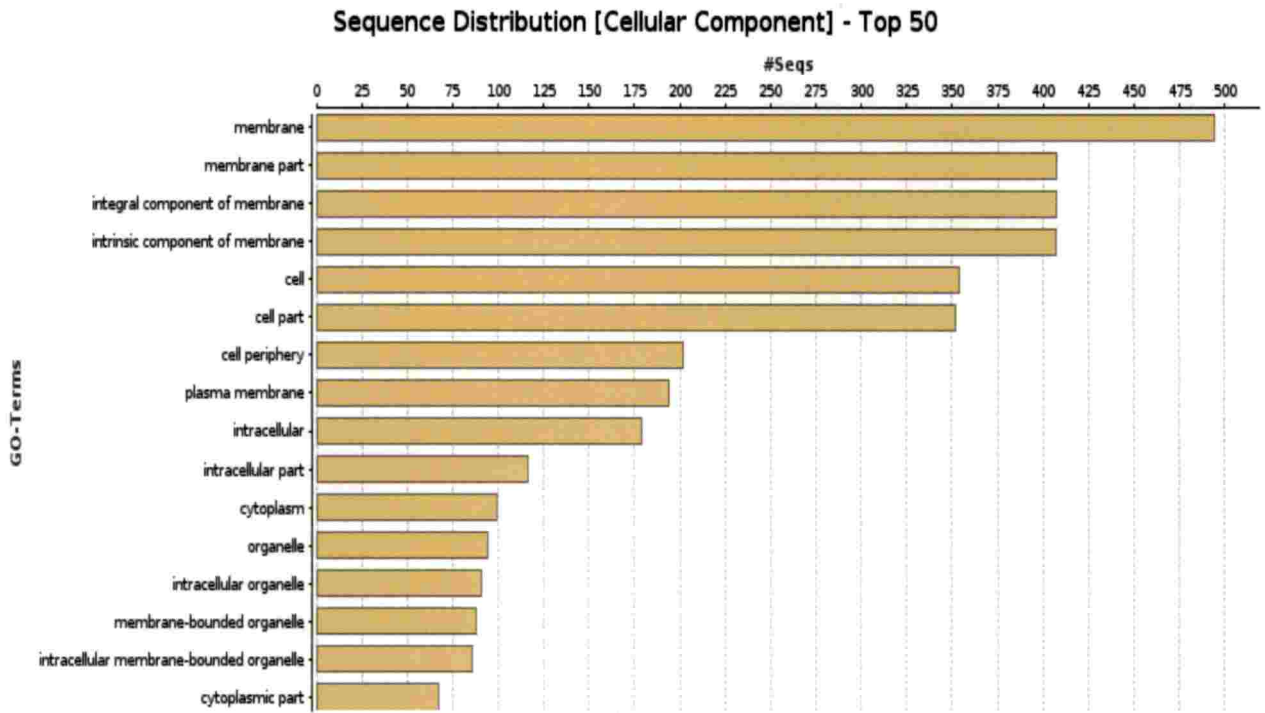


Figure 11. Distribution of sequences in the cellular component

Table 4. Genes predicted to confer virus resistance in Cassava

SEQUENCE ID	LENGTH	SIM MEAN (%)
Cassava4.1_000748m	1049	62.49
Cassava4.1_031760m	861	68.21
Cassava4.1_000507m	1168	68.96
Cassava4.1_031334m	528	69.20
Cassava4.1_034433m	1036	66.84
Cassava4.1_022519m	824	70.17
Cassava4.1_032178m	1160	66.48
Cassava4.1_033689m	1050	60.24
Cassava4.1_000627m	1101	65.36
Cassava4.1_020952m	992	58.96
Cassava4.1_022344m	443	76.42
Cassava4.1_025806m	972	63.46
Cassava4.1_031978m	941	62.46
Cassava4.1_000798m	1029	66.13
Cassava4.1_032695m	837	70.68
Cassava4.1_001696m	834	62.25
Cassava4.1_028330m	1064	67.65
Cassava4.1_000585m	1119	68.12
Cassava4.1_000944m	991	62.91
Cassava4.1_028973m	1580	82.75
Cassava4.1_022814m	2050	81.62
Cassava4.1_033473m	1008	67.08

5.2 PREDICTION AND ANALYSIS OF GRN USING DIFFERENT COMPUTATIONAL METHODS

5.2.1 WGNCA

Weighted Gene Expression Correlation Network (WGCNA), a network construction approach based on correlation among genes, was carried out for a sample microarray dataset. The WGCNA package in R version 3.5.1 was used to carry out the method. The sample microarray dataset used was obtained from Ghazalpour *et al.* (2006) (Integrating Genetics and Network Analysis to Characterize Genes Related to Mouse). The steps that were used to carry out WGCNA in R is given below in Figure 14. Separate R code is used for each step and the program is run in R studio version 1.0.143, which provides an open source programming environment for R. The final result file obtained is visualized in cytoscape to view the interaction network. A network consisting of 409 nodes with a characteristic path length of 1.550 was created with a network density of 0.45. A network of radius 2 was created with a clustering coefficient of 0.777. The time taken for analysis was about 0.64 sec.

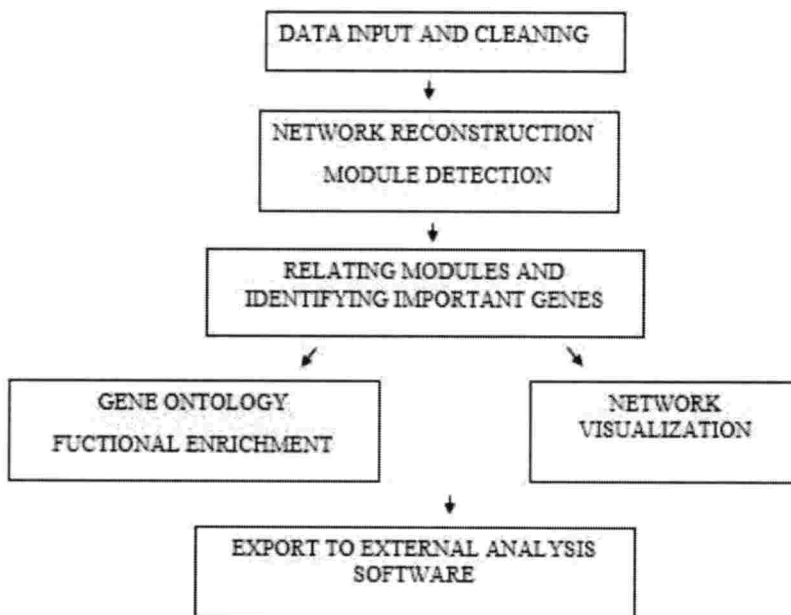


Figure 14. Flowchart for WGCNA

5.2.2 ARACNE

Algorithm for the Reconstruction of Accurate Cellular Networks works on the basis of a mutual information-based approach and the tool used here is ARACNE Java reference implementation in Cytoscape. Regulatory network was created using a sample microarray dataset and the obtained genomic dataset. The ARACNE tool available in Cytoscape 3.4.2 was used for analysing both the datasets. The sample microarray dataset was used to construct a network made up of 405 nodes having a network density of 0.121. The network diameter seems to be 2 and the clustering coefficient was observed to be 0.252. the time taken for analysis was about 0.094 sec.

5.2.3 Bayesian K2

Bayesian K2 is a probabilistic method-based network construction approach. The Bayesian K2 Algorithm is carried out using Cynitools, which is a Cytoscape App available at <http://proteomics.fr/Sysbio/CyniProject>. The number of parents is set as default and the analysis is carried out by selecting all data attributes as sources. The row order options are set as the default Cytoscape order. The result obtained is shown with different network sizes like 50,100 and 150.

5.2.4 Mutual Information based

Mutual Information method is carried out by Cyni algorithm-based Inference tool. The threshold to add new edge is set to be 0.6. If the network associated to table data is used, it is possible to set parameters like selection of nodes so that the data attributes of the selected nodes are considered as the source for network inference.

5.2.5 Basic Correlation based

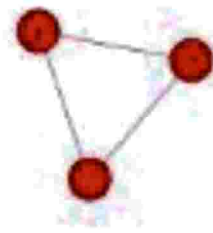
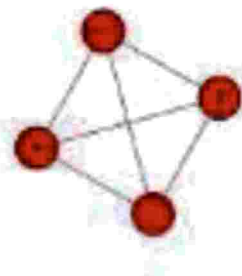
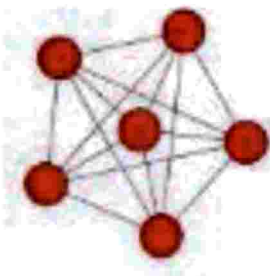
The Basic Correlation Inference Algorithm enables to carry out network inference based on different types of Correlations like positive, negative and absolute value. Different metric settings can be also given which works based on Pearson's Correlation, Kendall Tau Correlation or Spearman Rank Correlation. Cynitool, which is a Cytoscape v3.3 Plugin was used to carry out the method.

5.2.6 Generation of simulated dataset

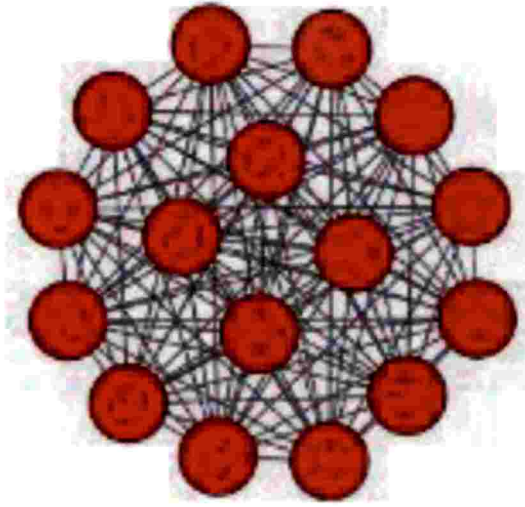
SynTRen, implemented in Java 5 is used to generate a synthetic Transcriptional Regulatory Networks (TRN) and the corresponding microarray data sets. In the network obtained, the nodes corresponds to the genes and the edges is used to represent the regulatory interactions at transcriptional level between the genes. The conditions set for generating datasets include Burnin period of 2000, number of experiments 10, number of nodes 100, experimental noise 0.1, random seed 1300 etc. The simulated dataset generated (Figure 19) is considered as the true interaction and is used for comparing the networks generated by different approaches. The synthetic Transcriptional Regulatory Network created is shown in Figure.

5.2.7 Comparison of different methods using network parameters

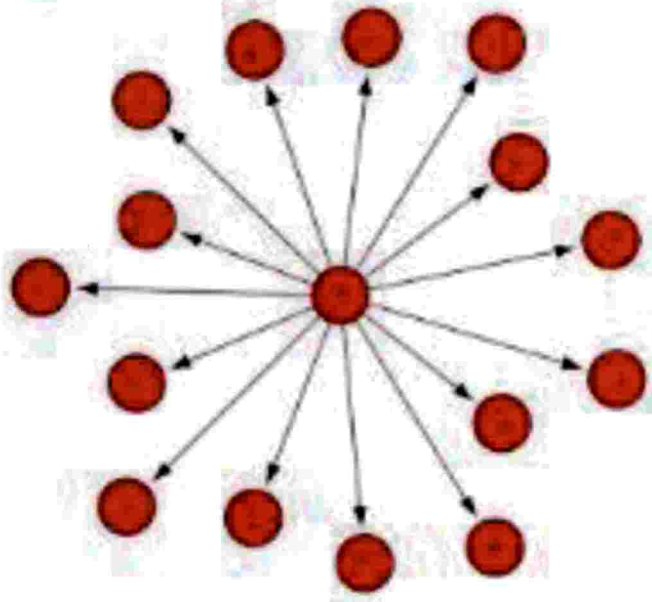
Different network parameters like Clustering coefficient, network diameter, network heterogeneity, isolated nodes, network density etc. is compared for networks of different sizes like 50,100 and 150 (Table 5,6 and 7). The clustering coefficient seems to be higher for mutual information and basic correlation approach. Even when the number of genes increases, the clustering coefficient remains the same. The obtained networks are depicted in Figure 15, 16 and 17 corresponding to Network size, $N=50,100$ and 150. And networks were also generated for $N=200,500,1000$ and 1500 as shown in Figure 18.



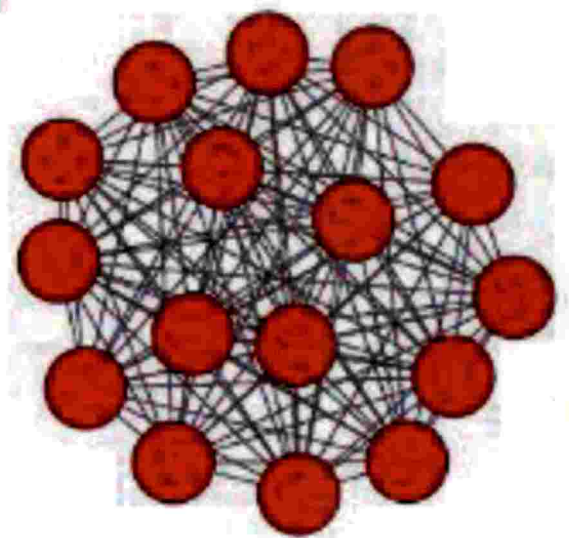
15a. ARACNE



15b. Mutual Information



15c. Bayesian K2



15d. Correlation

Figure 15. Networks with a size of N=50 constructed using different algorithms

Table 5. Network Statistics for N=50

72

NETWORK STATISTICS (N=50)				
PARAMETERS	CORRELATION	BAYESIAN K2	MUTUAL INFORMATION	ARACNE
Clustering Coefficient	1	0	1	0.867
Connetcted components	1	1	1	4
Network diameter	1	2	1	1
Network radius	1	1	1	1
Network centralization	0	1	0	0.137
Shortest paths	240 (100%)	240 (100%)	240 (100%)	23 (50%)
Characteristic path length	1	1.875	1	1
Avg. number of neighbours	15	1.875	15	0.333
Number of nodes	16	16	16	15
Network density	1	0.125	1	0.238
Network heterogeneity	0	1.807	0	0.447
Isolated nodes	0	0	0	0
Number of self-loops	0	0	0	0
Multi-edge node pairs	0	0	0	0
Analysis time (sec)	0.012	0.01	0.011	0.01

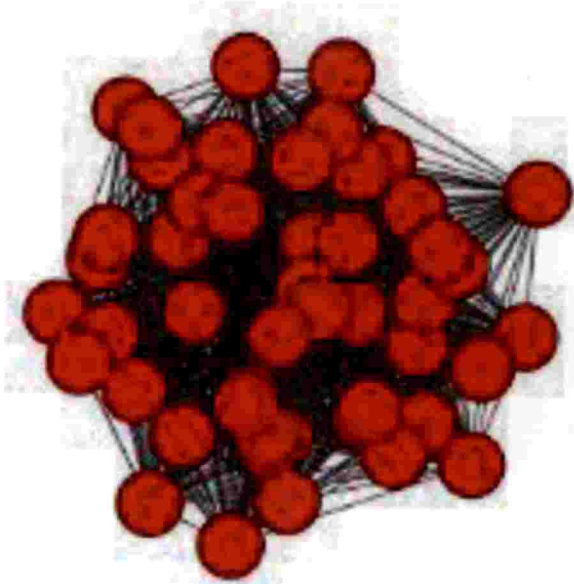
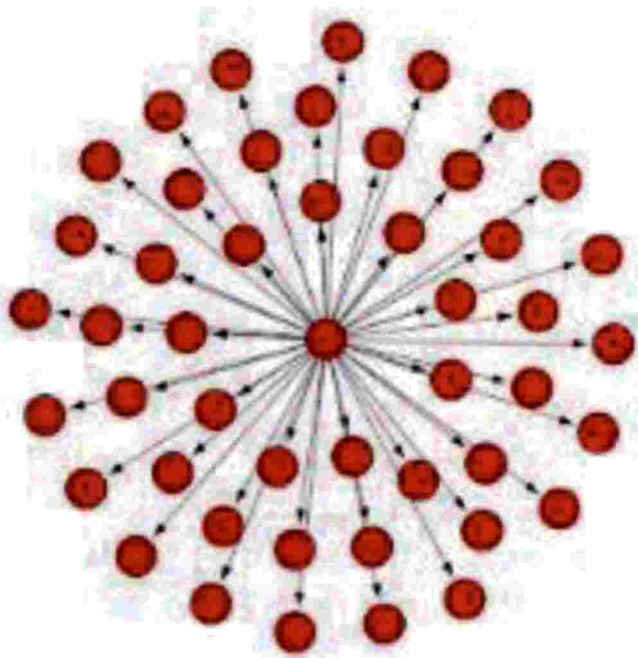
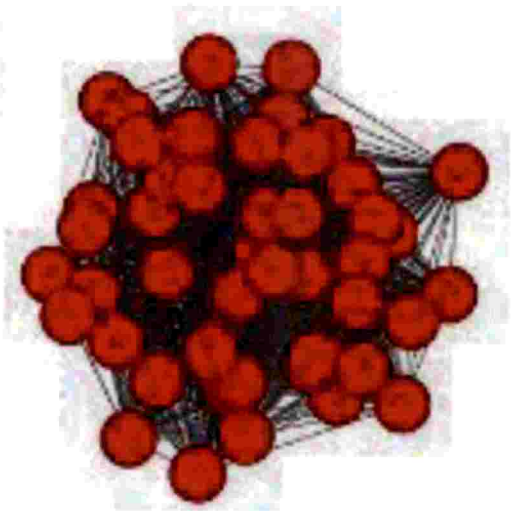
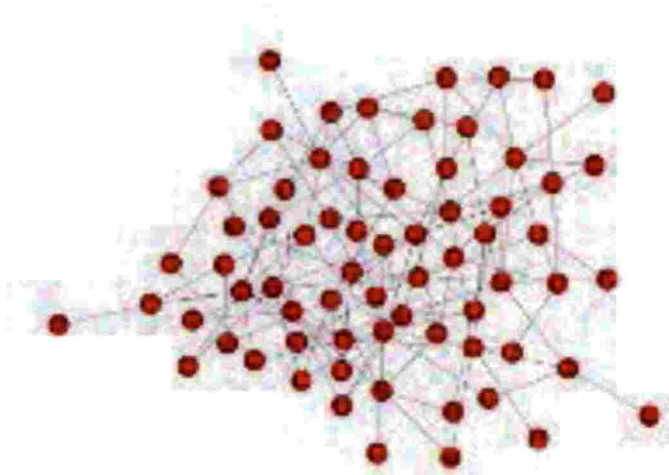


Figure 16. Networks with a size of $N=100$ constructed using different algorithms

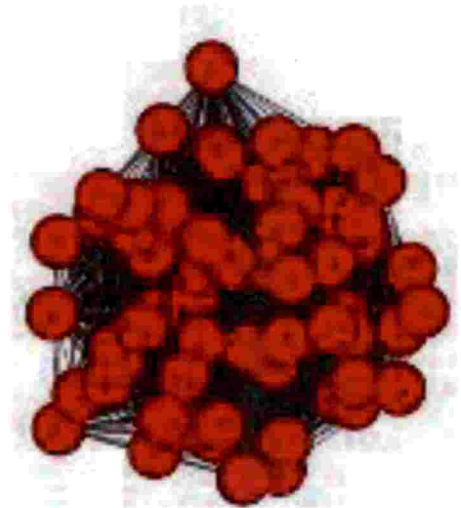
Table 6. Network Statistics for N=100

74

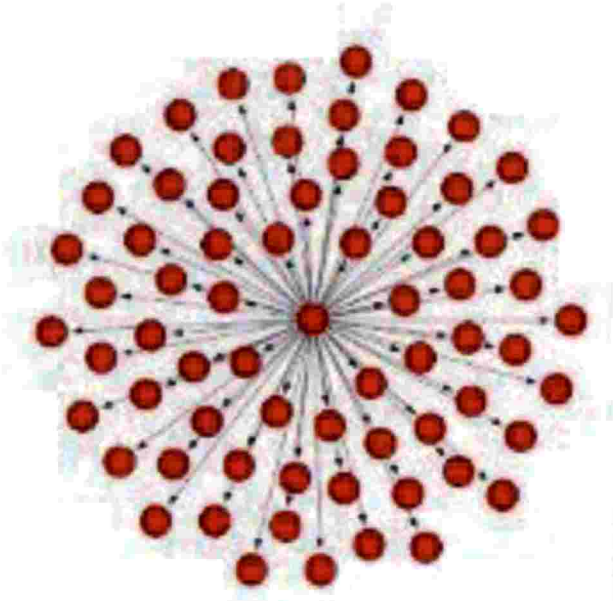
NETWORK STATISTICS (N=100)				
PARAMETERS	CORRELATION	BAYESIAN K2	MUTUAL INFORMATION	ARACNE
Clustering Coefficient	1	0	1	0
Connected components	1	1	1	2
Network diameter	1	2	1	8
Network radius	1	1	1	1
Network centralization	0	1	0	0.162
Shortest paths	2256 (100%)	2256 (100%)	2256 (100%)	1642 (90%)
Characteristic path length	1	1.958	1	3.495
Avg. number of neighbours	47	1.958	47	2.512
Number of nodes	48	48	48	43
Network density	1	0.042	1	0.06
Network heterogeneity	0	3.355	0	0.77
Isolated nodes	0	0	0	0
Number of self-loops	0	0	0	0
Multi-edge node pairs	0	0	0	0
Analysis time (sec)	0.049	0.013	0.032	0.019



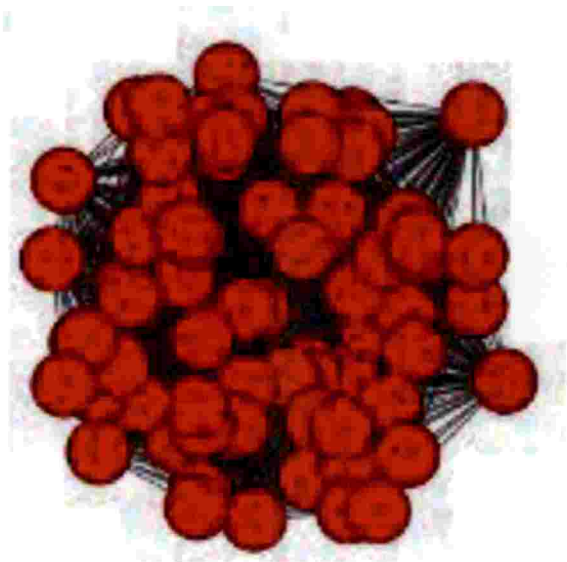
17a. ARACNE



17b. Mutual Information



17c. Bayesian K2



17d. Correlation

Figure 17. Networks with a size of $N=150$ constructed using different algorithms

Table 7. Network Statistics for N=150

76

NETWORK STATISTICS (N=150)				
PARAMETERS	CORRELATION	BAYESIAN K2	MUTUAL INFORMATION	ARACNE
Clustering Coefficient	1	0	1	0
Connected components	1	1	1	1
Network diameter	1	2	1	6
Network radius	1	1	1	4
Network centralization	0	1	0	0.126
Shortest paths	4692 (100%)	4692 (100%)	4422 (100%)	4556 (100%)
Characteristic path length	1	1.971	1	3.177
Avg. number of neighbours	68	1.971	66	3.824
Number of nodes	69	69	67	68
Network density	1	0.029	1	0.057
Network heterogeneity	0	4.062	0	0.562
Isolated nodes	0	0	0	0
Number of self-loops	0	0	0	0
Multi-edge node pairs	0	0	0	0
Analysis time (sec)	0.033	0.016	0.031	0.01

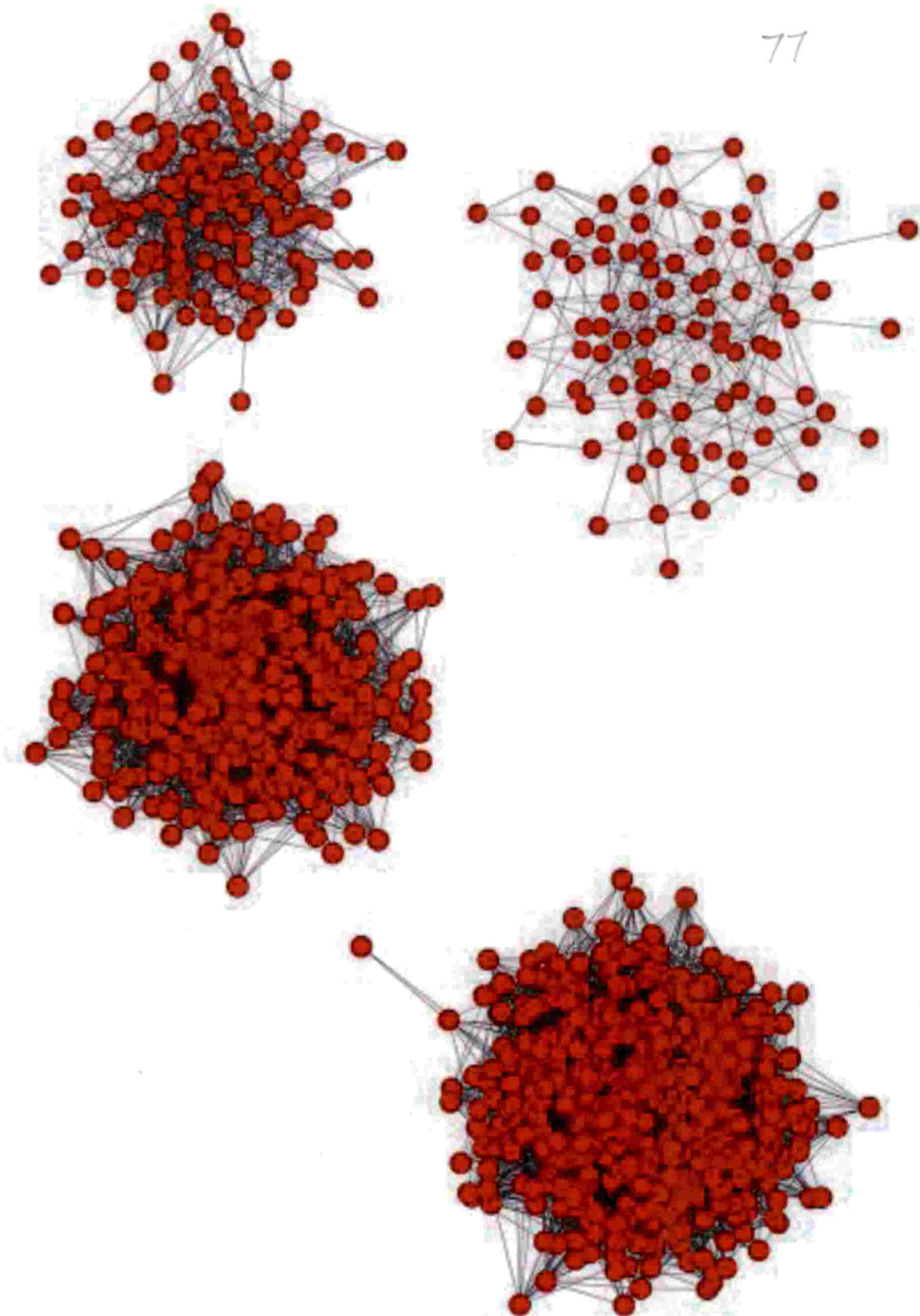


Figure 18. Networks constructed with ARACNE having different network sizes
N=500, 1000, 1500

5.2.8 Network Analysis

The analysis of the constructed network was carried out by calculating the parameters like sensitivity, specificity, PLR (Positive Likelihood Ratio), NLR (Negative Likelihood Ratio), PPV (Positive Predicted Value) and NPV (Negative Predicted Value) as shown in Table 9 using certain other parameters used for interaction studies (Table 8). Specificity and Sensitivity seems to be highest for Mutual information-based method and lowest for Bayesian K2, which indicates that Mutual Information based method is of more efficiency. ROC curve was plotted and AUC determined to correctly evaluate the efficiency of the selected methods. The accuracy measurements are calculated (Figure 20) and plotted as shown in Figure 21.

Table 8. Parameters for Interaction Statistics

	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)	TRUE NEGATIVE (TN)	FALSE NEGATIVE (FN)
ARACNE	22	6	20	9
Mutual Information	25	7	27	8
Correlation	16	6	17	7
Bayesian K2	20	12	22	9



Figure 19. Synthetic Transcriptional Regulatory Network Constructed using SynTren

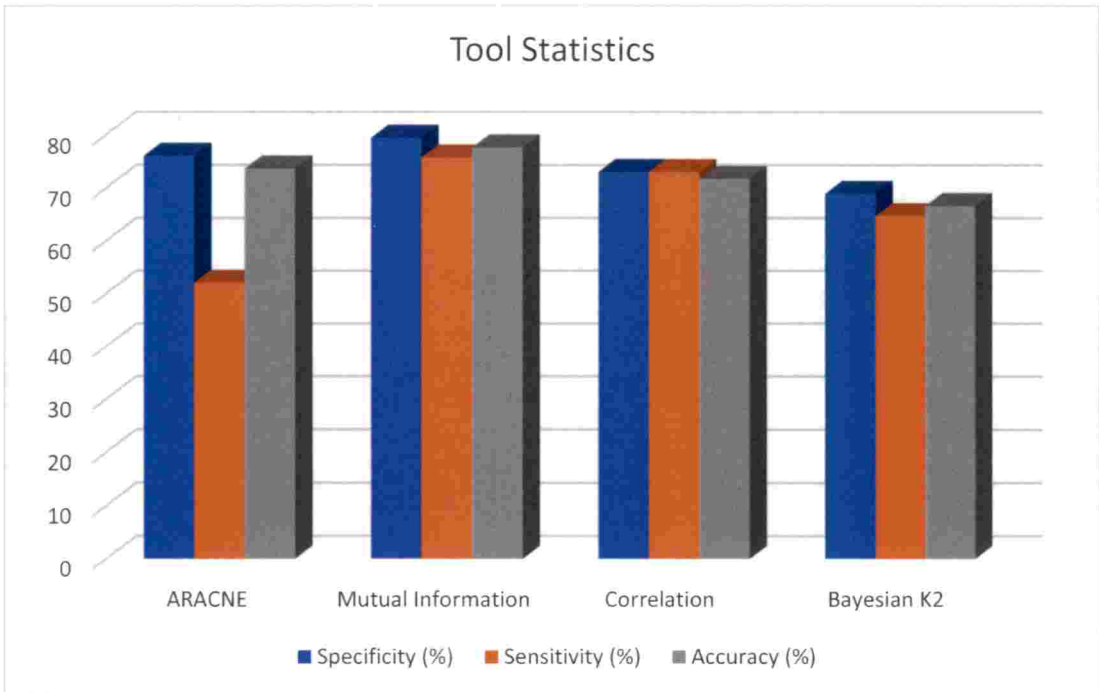
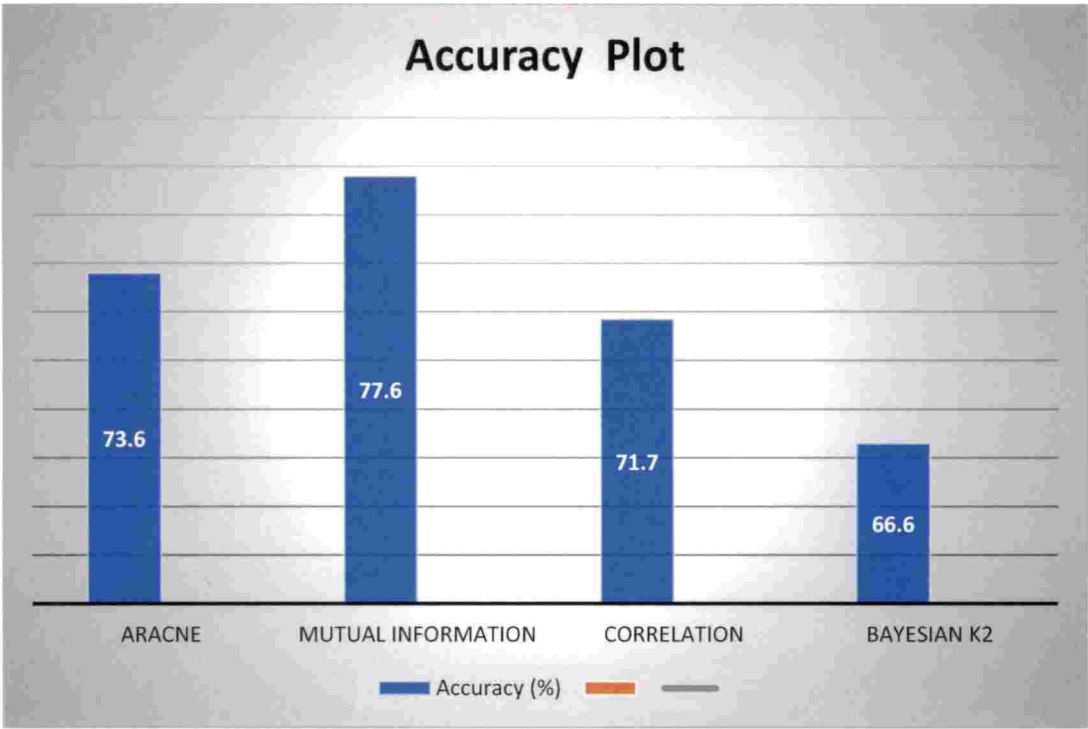


Table 9. Parameters for comparing the performance of the methods

	Specificity (%)	Sensitivity (%)	PLR	NLR	PPV (%)	NPV (%)
ARACNE	76	52	2.16	0.63	78.5	31
Mutual Information	79.4	75.7	3.60	0.30	78	23
Correlation	73	73	2.70	0.36	76	20
Bayesian K2	68.9	64.7	1.96	0.48	62.5	26

5.3 MICROARRAY DATA INTEGRATION

Cassava cDNA microarray dataset of genes related to *Xanthomonas axonopodis* pv. *manihotis* (*Xam*) infection in Cassava causing CBB (Cassava Bacterial Blight), obtained from ArrayExpress (E-GEOD-29379) was used to screen the genes for resistance to CBB. The corresponding sequences of dataset consisting of 199 genes obtained from GEO (GSE 29379) were used to search for homology with sequences of the identifies genes which gave a resulting 727 genes which are proved to have specific resistance to Bacterial Blight in Cassava.

5.4 PROTEIN-PROTEIN INTERACTION NETWORK CONSTRUCTION

String v10.5 was used to determine the protein protein interaction of the obtained genes (Figure 22). Out of the total 727 genes identified to be specific to Bacterial blight resistance in cassava, 324 of them were identified to have predicted interactions. The obtained result predicts the network statistics to be constituted by 324 nodes, 3140 edges, average node degree 19.4, PPI enrichment value $< 1.0e-16$ and average local clustering coefficient is estimated to be 0.495. Gene Neighbourhood data was also plotted in STRING as shown in Figure 23.

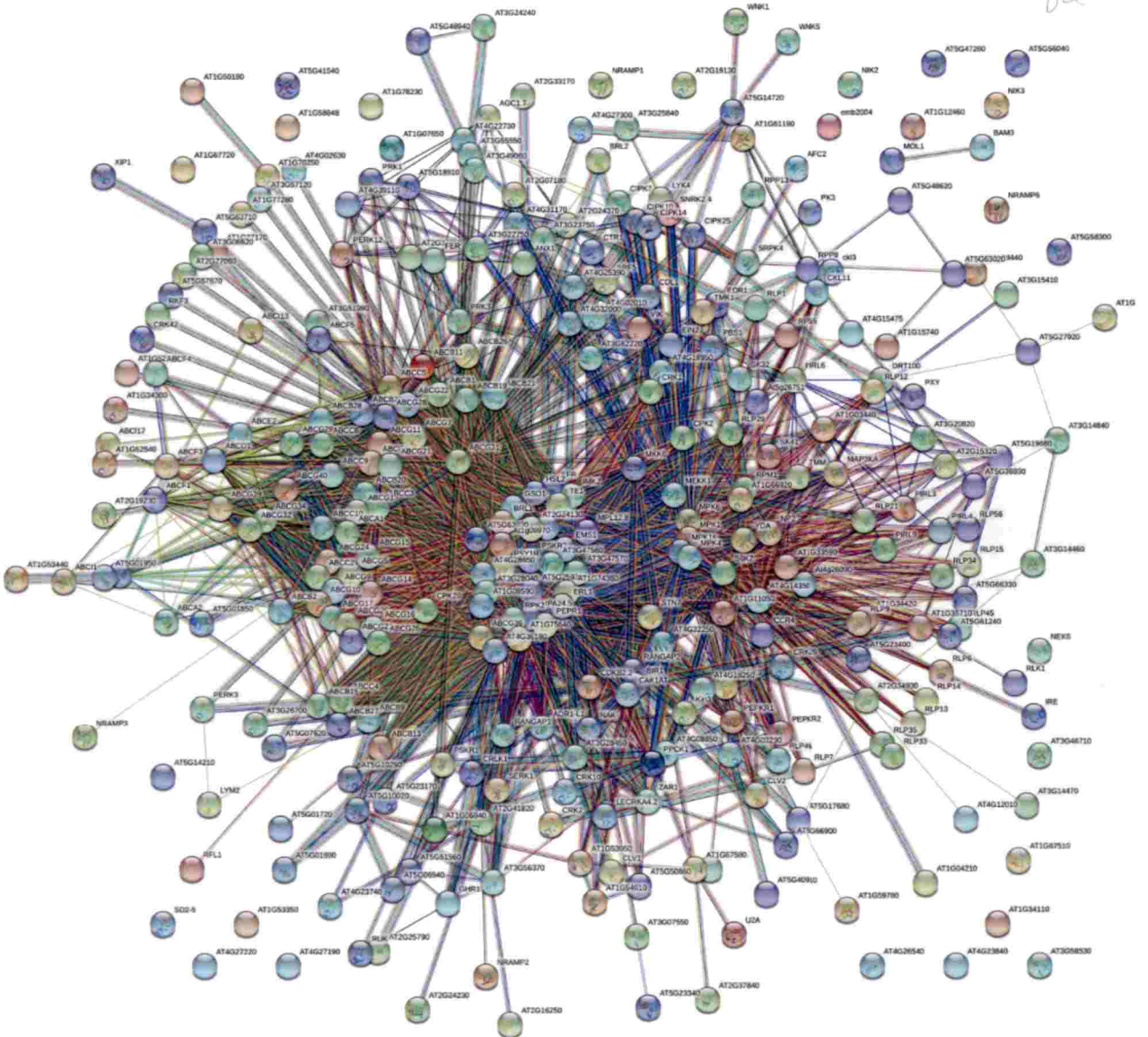


Figure 12. Protein- Protein interactions predicted from STRING v10.5

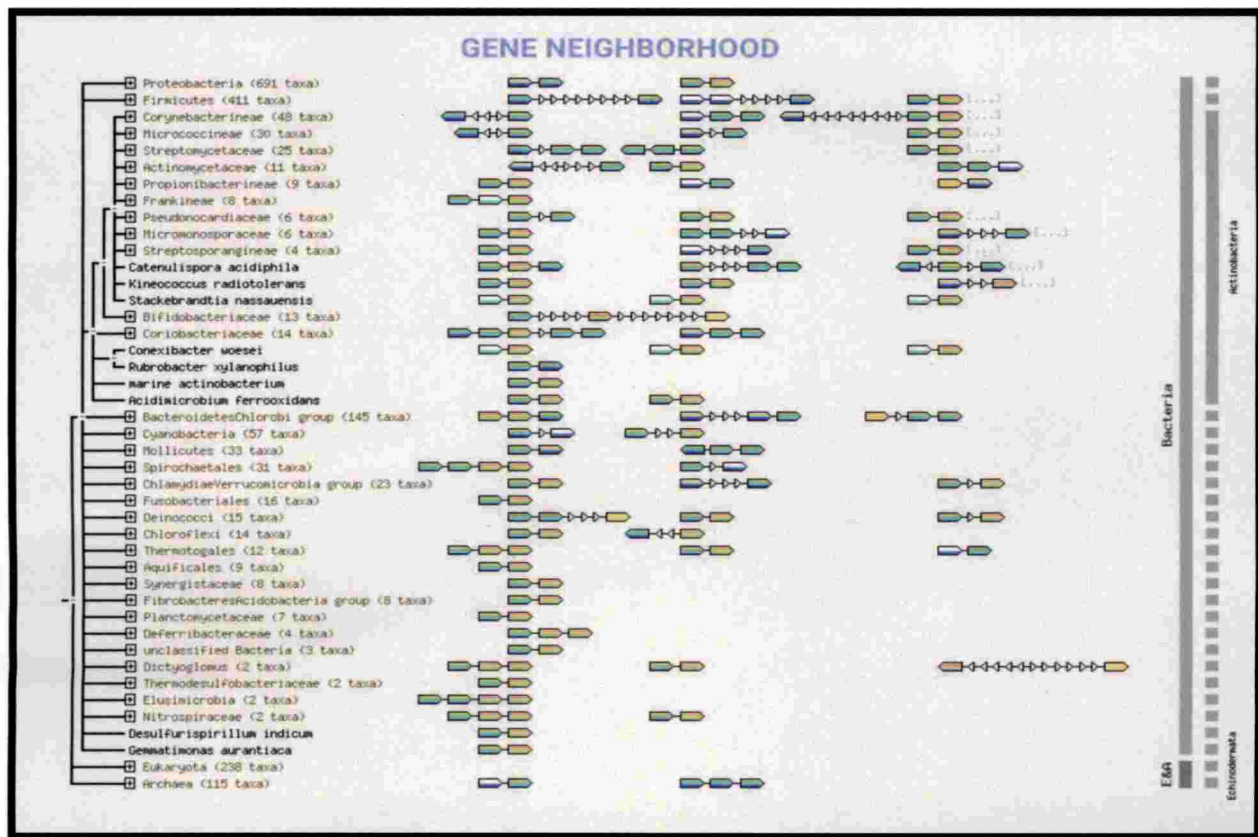


Figure 13. Neighbourhood pattern of the identified genes obtained from STRING database

5.5 VISUALIZATION AND VALIDATION

Visualization tool used was Cytoscape v3.6.1. The reconstructed regulatory network was validated by comparing with the Synthetic Transcriptional Regulatory Network created by SynTRen (Appendix II) and several network parameters were compared to evaluate the accuracy of the generated network depicted in Figure 24.

5.6 DEVELOPMENT OF ONLINE VISUALIZATION TOOL

We have tried to develop an interactive web app directly from R using Shiny in RStudio. A comparative evaluation of different tools as shown in Table 10, could help in adapting the better feasible strategy for GRN prediction. Shiny is an open source R package which provides an elegant frame work for web applications. Shiny applications seem to have two components, a user interface and a server function that creates a Shiny app object from server pair. The application is made only with a few lines of code, without the requirement of JavaScript. Other packages used along with Shiny were reshape, reldist etc. Reactive Programming is an important feature of Shiny. It has a reactive programming library that can be used to structure the application logic. googleViz is incorporated for visualization. The complete structure of the tool is not yet elucidated and the full R code is yet to be constructed.

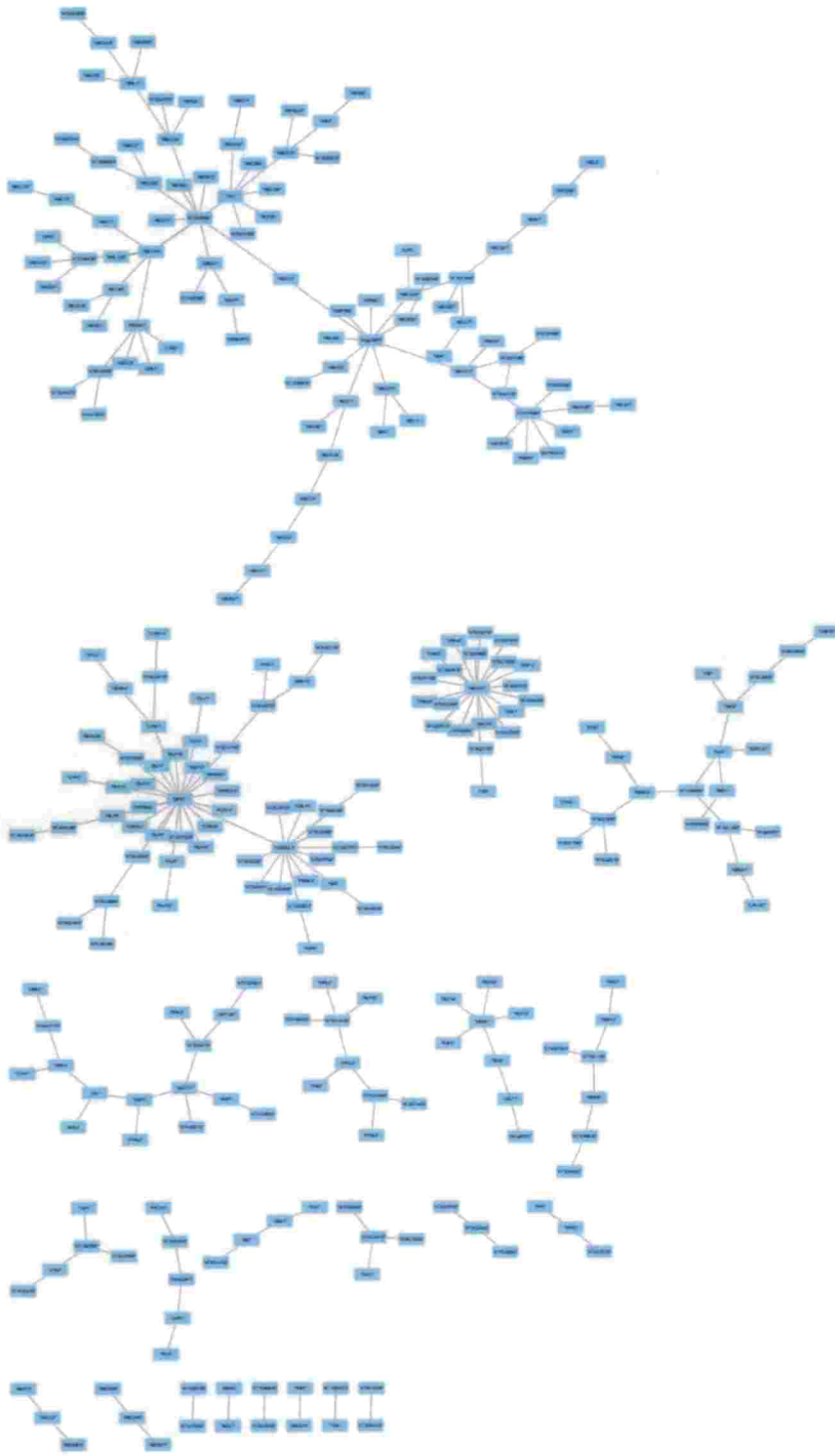


Figure 24. The reconstructed pathway as visualized in Cytoscape v3.6.1.

Table 10. Tools used for the Prediction of Gene Regulatory Networks

TOOL	METHOD	OS	URL	CITATION
BeaconGRN inference tool	Machine learning	Unix/Linux	https://omictools.com/beacon-grn-inference-tool	Aghamirzaie <i>et al.</i> , 2016
GeNESIS	Machine learning (GA)	Unix-like, Windows, macOS	http://genomics.iab.keio.ac.jp/genesis.html	Kratz <i>et al.</i> , 2008
SIRENE	Machine learning (SVM)	Windows	http://cbio.ensmp.fr/sirene	Mordelet <i>et al.</i> , 2008
ARACNE	Mutual Information	Unix, Linux Windows, macOS	https://omictools.com/aracne-tool	Margolin <i>et al.</i> , 2006
GeRNet	BiHEA and GRNCOP2	Unix/Linux	https://omictools.com/gernet-tool	Dussaut <i>et al.</i> , 2017
SEBINI	Inference and discretization algorithms	Unix/Linux, Windows	https://omictools.com/sebini-tool	Taylor <i>et al.</i> , 2006
BANJO	Bayesian Inference	Unix/Linux, Mac OS, Windows	https://omictools.com/banjo-tool	Smith <i>et al.</i> , 2006
WGCNA in R	Correlation based	Unix/Linux, Windows	https://cran.r-project.org/web/packages/WGCNA/index.html	Langfelder <i>et al.</i> , 2008
GEPASI	Pearson Partial Correlation	Windows	https://omictools.com/gepasi-tool	Mendes <i>et al.</i> , 1993
GRRANN	Artificial Neural Network	Unix/Linux, Mac OS, Windows	https://omictools.com/grrann-tool	Kang <i>et al.</i> , 2017
GENIE3	Random Forest method	Unix/Linux, Mac OS, Windows	https://omictools.com/genie3-tool	Irrthum <i>et al.</i> , 2010
SCENIC in R	Clustering based	Unix/Linux, Mac OS, Windows	https://omictools.com/scenic-tool	Aibar <i>et al.</i> , 2017
FANOVA	Gaussian Process Based	Unix/Linux	https://omictools.com/fanova-tool	Darnell <i>et al.</i> , 2017
sgnesR	Stochastic Simulation Algorithm	Unix/Linux, Mac OS, Windows	https://omictools.com/sgnesr-tool	Tripathi <i>et al.</i> , 2017
ENNET	Regression based method	Unix/Linux	https://omictools.com/enet-tool	Slawek <i>et al.</i> , 2013
MICMIC	Correlation based	Unix/Linux	https://omictools.com/micmic-tool	Tong <i>et al.</i> , 2018
HOCCLUS	Biclustering based	Unix/Linux	https://omictools.com/hocclus-tool	Pio <i>et al.</i> , 2014
SINCERITIES	Linear Regression based	Mac OS, Windows	https://omictools.com/sincerities-tool	Papili Gao <i>et al.</i> , 2017
GNW	Community profiling based	Unix/Linux	https://omictools.com/gnw-tool	Schaffter <i>et al.</i> , 2011
ARIMAI-VBEM	Bayesian Inference	Unix/Linux	https://omictools.com/arimai-vbem-tool	Sanchez-Castillo <i>et al.</i> , 2017

DISCUSSION

5. DISCUSSION

The study entitled “Comparative evaluation of tools for Gene Regulatory Network prediction and Network reconstruction using genomic data” was conducted to develop a better feasible methodology for the construction of gene regulatory network using genomic data and hence identify the immunity related genes and their interactions in cassava for better understanding of the defence mechanism of the crop. The study also includes a comparison of different network construction approaches and the evaluation of their performance which could specifically screen the best available method. The results of this study presented in chapter 4 are discussed here.

Cassava production although being economically feasible and industrially profitable is severely affected by several infectious pathogens. Depending on the locality, the infection and the infectious agent will vary. When Cassava Mosaic Geminiviruses (CMGs) and Cassava Brown Streak Virus (CBSV) infects sub-Saharan Africa (Legg *et al.*, 2003), phytoplasmal disease called Cassava witches broom (CWB) is reported to infect 64% of study plots in South-east Asian countries (Graziosi *et al.*, 2016). In such a situation where the infection and the infectious agent becomes less predictable by a single defence strategy, it becomes important to develop more feasible and less time-consuming strategies for the identifying the defence mechanisms associated with disease infection.

Developing regulatory network of genes controlling traits which are of importance economically, commercially and academically are gaining much importance in present times. GRN's provide an insight into the transcriptional mechanisms that regulate the robust and stochastic gene expression and their relationship with the phenotypic variability that can be utilized for better crop improvement strategies. For example, when we take the case of Maize, in 2012, Dong *et al* developed a gene regulatory model for the floral transition of the shoot apex in maize which proposed the genetic control of flowering time in maize that could facilitate maize breeding and transgenic product development. Similarly,

Wils *et al.*, 2017 developed gene regulatory network controlling inflorescence and flower development in *Arabidopsis thaliana*. Jiang *et al.*, 2018 analyses Gene Regulatory Network of Maize in response to nitrogen Artificial Neural Network Analysis.

Here, we have tried to develop the gene regulatory network related to *Xanthomonas axonopodis* pv. *manihotis* (*Xam*) infection causing Bacterial Blight in cassava. The mechanism of CBB resistance is analysed and checked for homology with the total defence mechanism in Cassava. From the total 1919 immunity related genes identified, 727 of them were selectively predicted to infer CBB resistance, out of which 324 of them had predicted interactions in STRING v10.5. Earlier methods for screening cassava genotypes for bacterial blight disease were based on approaches like stem inoculation methods (Banito *et al.*, 2010). Bioinformatics approaches for gene identification could improve the efficiency and economic feasibility and hence is better adopted. Bioinformatic identification of miRNAs differentially expressed in cassava in response to infection by *Xanthomonas axonopodis* pv. *manihotis* could prove the critical role played by miRNAs in defense against *Xam*. (Perez-Quintero *et al.*, 2012).

The genomic dataset used for the study consists of 1919 immunity related genes in cassava, identified by a simple and effective strategy developed by a combination of two methods. Leal LG *et al.*, 2013 used a holistic approach to combine their own microarray and RNA-seq data with public genome data from *Arabidopsis* and cassava in order to acquire biological knowledge about the proteins encoded by immunity related genes and other genes. They constructed a network of immunity related genes in *Arabidopsis* using a kernel-based correlation approach. Lozano *et al.*, 2015 identified NBS-LRR type genes in cassava by searching for Pfam domains in the cassava genome and manual curation of the cassava gene annotations. They identified 228 NBS-LRR type genes and 99 partial NBS genes. A combination of both these approaches was tried to develop the immunity related genes network in cassava by converting the protein-protein interaction parameters into network parameters. This approach could develop a less



time consuming and feasible approach for gene regulatory network construction using genomic data.

The genes identified separately in different canonical immune protein domains were combined in later stages for network reconstruction. Effective annotation obtained from Blast2GO could provide several valuable data regarding the identified genes. Among the total genes identified, 22 of them were specifically found to confer virus resistance. Most of the genes identified were found to be involved in imparting immunity by altering the metabolic pathway of purines. Upon mapping with the KEGG pathways, it was observed that certain genes that plays role in providing immunity in cassava are part of the sulfur metabolic pathway. Sulfur metabolism in plants offers several ways to combat fungal attack. Haneklaus *et al.* (2007) presents a model that reflects the synthesis of sulfur metabolites and related biochemical pathways which are putatively triggered by Sulfur Induced Resistance (SIR) in oilseed rape in a chronological manner. Sulfur Induced Resistance (SIR) is the mechanism of stimulation of metabolic processes involving sulfur by targeted sulfate-based and soil-applied fertilizer strategies which could reinforce the natural resistance of plants against fungal pathogens. The identified genes could be probably part of a Sulfur Induced Resistance Mechanism (SIR) in cassava. Unravelling the mechanism causing SIR in plants is significant for sustainable agricultural production as the input of fungicides can be minimized by crop specific Sulfur fertilization and a higher resistance due to Sulfur will not be rapidly broken by new pathotypes (Bloem *et al.*, 2015). Rausch *et al.*, 2005 also discusses the crucial role of Sulfur-containing Defence Compounds (SDCs) for the survival of plants under biotic and abiotic stress.

The various approaches used for gene regulatory network construction was mutual information based, correlation based, probabilistic method, coexpression based methods using various tools like ARACNE, cynitools, WGCNA package in R etc. The network developed by these methods were evaluated by visualizing in Cytoscape. Although numerous methods have been developed for inferring gene regulatory networks from expression data. Both their absolute and comparative

performance remains poorly understood. Current inference methods seem to be affected to various degrees by different types of systematic prediction errors. Hence the performance of community-wide challenge within the context of DREAM (Dialogue on Reverse Engineering Assessment and Methods) project has been proved more reliable than individual inference methods (Marbach *et al.*, 2010). Here, they systematically form communities composed of the top two methods, the top three methods, the top four methods, etc., until the last community, which contains all applied methods of a particular subchallenge.

The present research focusses on exploring the relatively unexplored fourth dimension of gene regulatory networks i.e. time. Varala *et al.*, 2018 applied a time-based machine learning method to learn the temporal transcriptional logic underlying dynamic nitrogen (N) signalling in plants. A dynamic regulatory network of N-responsive genes was constructed to identify 155 candidate TFs that could improve nitrogen use efficiency with potential agricultural applications. One of the recent approaches in investigating biological networks is based on the network modules like communities, clusters, and subnetworks etc. (Wang *et al.*, 2010). Several algorithms like network clustering (Ihmels *et al.*, 2005), heuristic search (Dittrich *et al.*, 2008), seed extension (Ulitsky *et al.*, 2009), topology network (Chin *et al.*, 2010), and matrix decomposition (Li *et al.*, 2006) have been proposed to for the identification of modules. In contradiction to the large number of methods for module detection, only few methods have been developed for module evaluation and validation. Topology based approach for module validation has been incorporated in the study which includes detection of modularity, connectivit, density, clustering coefficient, degree, and edge betweenness.

SUMMARY

6. SUMMARY

93

The study entitled “Comparative evaluation of tools for Gene Regulatory Network prediction and Network reconstruction using genomic data” was carried out at the Section of Extension and Social Sciences, ICAR-Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram during 2017-2018. The objectives of the study were to reconstruct the regulatory network of immunity related genes in cassava using genomic data, to compare different computational methods for Gene Regulatory Network prediction and analysis and to develop an online visualization tool using the appropriate method.

The study had mainly two objectives, GRN prediction tool evaluation and network reconstruction using genomic data. The initial genomic dataset used for the study was derived by homology search using HMMER. The immune related protein domains were searched in whole genome resource of cassava available publicly at Phytozome. Hidden Markov Models corresponding to particular immune domain was generated and compared. Protein domain search and analysis carried out using HMMER suite version v3.1b2 resulted in identification of an initial set of immunity related genes in cassava. These genes were further filtered for high competence cassava specificity in two ways. Initially by constructing a plant R gene database with already identified immunity related genes and other plant R genes, BLASTP was performed and those with E-value less than 0.01, unrelated to immunity and less than 250 amino acids were filtered. Online BLASTP in PRGDB 3.0 was also performed and both results were compared to get a dataset consisting of immune specific genes in cassava. A set of 1919 immunity related genes were identified in cassava out of which 22 of them were found to confer virus resistance in specific. Gene Ontology Mapping and Annotation of the identified genes were carried out using Blast2GO. InterPro annotations in Blast2GO 5 was also done to retrieve domain/ motif information in a sequence wise manner. Further annotation with Gene Ontology (GO), Enzyme Code (EC) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database eventually allowed the display of

enzymatic functions and the metabolic pathways related to the identified genes. Most of the genes identified were found to be involved in purine metabolism and sulfur metabolism which can pave way for a possible Sulfur-Induced Resistance mechanism in Cassava.

The predicted gene sequences were screened for bacterial blight resistance by incorporating the sequences of genes identified from a microarray experiment that listed a set of genes that were upregulated and down regulated in cassava upon *Xanthomonas axonopodis* pv. *manihotis* (*Xam*) infection (ArrayExpress: E-GEOD-29379; GEO: GSE 29379). Our identified gene sequences were searched for homology with these sequences to screen out 727 genes related to bacterial blight resistance in cassava. The protein-protein interactions of these genes were predicted using STRING v10.5 which collects and reassesses available experimental data on protein-protein interactions, and imports known pathways and complexes from curated databases. 324 predicted interactions were identified and a network was constructed using the protein- protein interaction parameters as input source. The network was visualized in Cytoscape v3.6.1 and the network parameters were determined. Validation was done by Topology based method of module detection by comparing certain parameters like clustering coefficient, network density etc. A synthetic Transcriptional Regulatory Network was constructed using SynTRen implemented in Java 5. The simulated expression dataset was generated and compared with the reconstructed network.

The latter part of the work consists of comparison of different approaches for Gene Regulatory Network construction and analysis of the reconstructed networks using simulated dataset. Different tools for GRN prediction like WGCNA, ARACNE etc were valued and the network was compared through statistical analysis by plotting ROC and determining AUC. The different approaches used were mutual information based, correlation based, ARACNE and Bayesian K2. Network parameters for each method were determined for different network sizes like N, number of genes=50, 100 and 150. The corresponding data was compared and inferred. For a particular method, the clustering coefficient,

which is measure of accuracy doesn't seem to alter with increase in network size. As expected the network heterogeneity, number of nodes and network density increases with increase in network size. The analysis of the constructed networks was carried out by calculating the parameters like sensitivity, specificity, PLR (Positive Likelihood Ratio), NLR (Negative Likelihood Ratio), PPV (Positive Predicted Value) and NPV (Negative Predicted Value). The results of the calculations show that mutual information-based approaches perform better than all other methods with a value for specificity as 79.4% and sensitivity as 75.7%. The plotted ROC and determined AUC also supports the obtained results and hence it can be concluded that among the several approaches developed for GRN construction, the tools focusing on mutual information-based method seems to work with more accuracy for predicting network interactions.

Gene Regulatory Networks provide insight into the transcriptional mechanisms of genes controlling traits which are of importance economically and commercially. Hence understanding these mechanisms in plants play major role in crop improvement and management. GRNs have also proved to be successful in detecting the pre-diseased state of a disease, that can help in adapting better prevention strategies. Construction of networks of immunity related genes in cassava can contribute to better understanding of defence mechanism in cassava and evaluation of the different tools for GRN construction will help in adapting a better method with more accuracy.

174552



REFERENCES

7. REFERENCE

99

- Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J. C., Geurts, P., Aerts, J., van den Oord, J., and Atak, Z. K. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat. methods* 14(11): 1083-1086.
- Allen, J. D., Xie, Y., Chen, M., Girard, L. and Xiao, G. 2012. Comparing statistical methods for constructing large scale gene networks. *PloS one* 7(1): 236-248.
- Altay, G. and Emmert-Streib, F. 2010. Inferring the conservative causal core of gene regulatory networks. *BMC Syst. Biol.* 4(1): 132p.
- Altenbach, D. and Robatzek, S. 2007. Pattern Recognition Receptors: From the cell surface to Intracellular Dynamics. *MPMI.* 20(9): 1031-1039.
- Amuge, T., Berger, D. K., Katari, M. S., Myburg, A. A., Goldman, S. L. and Ferguson, M. E. 2017. A time series transcriptome analysis of cassava (*Manihot esculenta* Crantz) varieties challenged with Ugandan cassava brown streak virus. *Sci. Rep.* 7(1): 9747p.
- Banito, A., Kpémoua, K. E. and Wydra, K. 2010. Screening of cassava genotypes for resistance to bacterial blight using strain \times genotype interactions. *J. of Plant Pathol.* pp.181-186.
- Bastian, M. and Heymann, S. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. In: Jacomy, M. (ed.), *Proc. of the Third Int. ICWSM Conference*, Gephi, WebAtlas, Paris, France, pp.361-362.
- Batushansky, A., Toubiana, D. and Fait, A. 2016. Correlation-based network generation, visualization, and analysis as a powerful tool in biological

- studies: a case study in cancer cell metabolism. *BioMed Res. Int.* 2016, 36p.
- Biondi, R. M. and Nebreda, A. R. 2003. Signalling specificity of Ser/Thr protein kinases through docking site mediated interactions. *Biochem J.* 372: 1-13.
- Bloem, E., Haneklaus, S., and Schnug, E. 2015. Milestones in plant sulfur research on sulfur-induced-resistance (SIR) in Europe. *Frontiers in Plant Sci.* 5(17): 345-356.
- Boari, A. J., Cunha, E. M., Quadros, A. F. F., Barreto, R. W., and Fernandes, A. F. 2018. First report of *Phytophthora* sp. causing storage root rot and foliage blight of cassava in Brazil. *Plant Dis.* 102(5): 1042-1042.
- Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., Grimwood, J., Schmutz, J., Rabbi, I. Y., Egesi, C. and Nauluvula, P. 2016. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* 34(5): 562p.
- Buist, G., Steen, A., Kok, J. and Kuipers, O. P. 2008. LysM, a widely distributed protein motif for binding to (peptido)glycans. *Mol. Microbiol.* 68(4): 838-847.
- Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. and Collins, J. J. 2018. Next-Generation Machine Learning for Biological Networks. *Cell* 14(8):1045-1058.
- Chandrasekara, A. and Josheph Kumar, T. 2016. Roots and tuber crops as functional foods: a review on phytochemical constituents and their potential health benefits. *Int. J. of Food Sci.* 16(8): 56-59.

- Chin, C. H., Chen, S. H., Ho, C. W., Ko, M. T., and Lin, C. Y. 2010. A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles. *BMC Bioinforma.* 11(1): S25p.
- Cock, J. H. 1982. Cassava: a basic energy source in the tropics. *Sci.* 218(4574): 755-762.
- Conesa., A. and Gotz., S. 2008. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *Int. J. of Plant Genomics* 8(7): pp.1-13.
- Darnell, C. L., Tonner, P. D., Gulli, J. G., Schmidler, S. C., and Schmid, A. K. 2017. Systematic Discovery of Archaeal Transcription Factor Functions in Regulatory Networks through Quantitative Phenotyping Analysis. *MSyst.* 2(5): pp.e00032-17.
- DeYoung, B. J. and Innes, R. W. 2006. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat Immunol.* 7(12): 1243-1249.
- Dias, R. O., Machado, L. S., Migliolo, L., and Franco, O. L. 2015. Insights into Animal and Plant Lectins with Antimicrobial Activities. *Molecules* 20: 519-541.
- Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. 2008. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinforma.* 24(13): pp.i223-i231.
- Dong, J. and Horvath, S. 2007. Understanding network concepts in modules. *BMC Syst. Biol.* 1(1): 24p.

- Dong, Z., Danilevskaya, O., Abadie, T., Messina, C., Coles, N., and Cooper, M. 2012. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. *PLoS One*: 7(8): e43450p.
- Dussaut, J. S., Gallo, C. A., Cravero, F., Martínez, M. J., Carballido, J. A., and Ponzoni, I. 2017. GeRNet: a gene regulatory network tool. *Biosystems* 162: pp.1-11.
- FAOSTAT (2016) Food and agriculture organizations statistics database. FAO, Rome. <http://www.fao.org/faostat/en.html>. [09 June 2017].
- FAOSTAT (2017) Food and agriculture organizations statistics database. FAO, Rome. <http://www.fao.org/faostat/en.html>. [07 April 2017].
- Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. 2007. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5(1): e8p.
- Finn, R. D., Coggil, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J, Mitchell, A. L., Potter, S. C., Punta, M., Quereshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44: D279-D285.
- Finn, R. D., Clements, J., and Eddy, S. R. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39(2): pp.W29-W37.
- Fofana, I. B., Sangaré, A., Collier, R., Taylor, C., and Fauquet, C. M. 2004. A geminivirus-induced gene silencing system for gene function validation in cassava. *Plant Mol. Biol.* 56(4): pp.613-624.

- Fokunang, C. N., Ikotun, T., Dixon, A. G. O., and Akem, C. N. 1997. First report of *Colletotrichum gloeosporioides* f. sp. *manihotis*, cause of cassava anthracnose disease, being seed-borne and seed-transmitted in cassava. *Plant Dis.* 81(6): pp.695-695.
- Georgii, E., Dietmann, S., Uno, T., Pagel, P., and Tsuda, K. 2009. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinforma.* 25(7): pp.933-940.
- Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E. E., Drake, T. A., Lusic, A. J., and Horvath, S. 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* 2(8): e130p.
- Graziosi, I., Minato, N., Alvarez, E., Ngo, D. T., Hoat, T. X., Aye, T. M., Pardo, J. M., Wongtiem, P., and Wyckhuys, K. A. 2016. Emerging pests and diseases of South-east Asian cassava: a comprehensive evaluation of geographic priorities, management options and research needs. *Pest Manag. Sci.* 72(6): pp.1071-1089.
- Gysi, D. M., Fragoso, T. M., Almaas, E., and Nowick, K. 2018. CoDiNA: an R Package for Co-expression Differential Network Analysis in n Dimensions. *arXiv preprint arXiv:1802.00828*.
- Haneklaus, S., Bloem, E., De Kok, L. J., Yang, Z., Wang, S., and Schnug, E. 2007. The potential of sulfur induced resistance against plant diseases of oilseed rape. In: *THE 12th INTERNATIONAL RAPESEED CONGRESS*; p.43.
- Ihmels, J., Bergmann, S., Berman, J., and Barkai, N. 2005. Comparative gene expression analysis by a differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genet.* 1(3): e39p.

- Irrthum, A., Wehenkel, L., and Geurts, P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PloS one*. 5(9): e12776p.
- Ishihama, N. and Yoshioka, H. 2012. Post-translational regulation of WRKY transcription factors in plant immunity. *Current Opinion in Plant Biol.* 15(4): 431-437.
- Jansson, C., Westerbergh, A., Zhang, J., Hu, X., and Sun, C. 2009. Cassava, a potential biofuel crop in (the) People's Republic of China. *Appl. Energy* 86: pp.S95-S99.
- Jiang, L., Ball, G., Hodgman, C., Coules, A., Zhao, H., and Lu, C. 2018. Analysis of Gene Regulatory Networks of Maize in Response to Nitrogen. *Genes* 9(3): 151p.
- Jiang, Y., Mao, Y., Lv, Y., Tang, L., Zhou, Y., Zhong, H., Xiao, J., and Yan, J. 2018. Natural Resistance Associated Macrophage Protein is involved in Immune response of Blunt Snout Bream *Megalobrama amblycephala* 7:27.
- Jones, J. D. and Banfield, M. J. 2017. Two-faced TIRs trip the immune switch. *PNAS*. 114(10):2445-2446.
- Kang, J., Park, J., Choi, H., Burla, B., Kretschmar, T., Lee, Y., and Martinoia, E. 2011. Plant ABC Transporters e0153. doi: 10.1199/tab.0153.
- Kang, T., Ding, W., Zhang, L., Ziemek, D., and Zarringhalam, K. 2017. A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. *BMC Bioinforma.* 18(1): 565p.

- Karlebach, G. and Shamir, R. 2008. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* 9(10): 770p.
- Kobayashi, K. and Hiraishi, K. 2017. Design of probabilistic Boolean networks based on network structure and steady-state probabilities. *IEEE Trans. on neural networks and learning Syst.* 28(8): pp.1966-1971.
- Kratz, A., Tomita, M., and Krishnan, A. 2008. GeNESiS: gene network evolution simulation software. *BMC Bioinforma.* 9(1): 541p.
- Krouk, G., Lingeman, J., Colon, A. M., Coruzzi, G., and Shasha, D. 2013. Gene regulatory networks in plants: learning causality from time and perturbation. *Genome Biol.* 14: 123.
- Kuldau, G. A. and Yates, I. E. 2000. Evidence for Fusarium Endophytes. *Microbial endophytes* p.85.
- Langfelder, P. and Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9(1): 559p.
- Leal, L. G., Perez, A., Quintero, A., Bayona, A., Ortiz, J. F., Gangadharan, A., Mackey, D., Lopez, C., and Lopez-Kleine, L. 2013. Identification of immunity-related genes in Arabidopsis and Cassava using genomic data. *Genomics, proteomics & Bioinforma.* 11(6): pp.345-353.
- Legg, J. P. and Thresh, J. M. 2003. Cassava virus diseases in Africa. In: *Proceedings of the First International Conference on Plant Virology in Sub-Saharan Africa*, 4–8 June 2001, Ibadan, Nigeria. IITA, Ibadan, Nigeria, pp. 517-522.
- Li, H., Sun, Y., and Zhan, M. 2006. The discovery of transcriptional modules by a two-stage matrix decomposition approach. *Bioinforma.* 23(4): pp.473-479.

- Lozano, J. C. and Bellotti, A. 1978. *Erwinia carotovora* var. *carotovora*, causal agent of bacterial stem rot of cassava: etiology, epidemiology and control. *PANS*. 24(4): pp.467-479.
- Lozano, R., Hamblin, M. T., Prochnik, S., and Jannink, J. L. 2015. Identification and distribution of the NBS-LRR gene family in the Cassava genome. *BMC genomics*. 16(1): 360p.
- Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. 2010. Revealing strengths and weaknesses of methods for gene network inference. *Proc. of the Natl. Acad. of Sci.* 23-26 January 2014, Nigeria, pp.2145-2149.
- Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In: *BMC Bioinforma.* BioMed Cent. 7(1): S7
- Maruthi, M. N., Bouvaine, S., Tufan, H. A., Mohammed, I. U., and Hillocks, R. J. 2014. Transcriptional response of virus-infected cassava and identification of putative sources of resistance for cassava brown streak disease. *PloS one*, 9(5): e96642.
- Mehta, D., Stürchler, A., Hirsch-Hoffmann, M., Grussem, W., and Vanderschuren, H., 2018. CRISPR-Cas9 interference in cassava linked to the evolution of editing-resistant geminiviruses. *bioRxiv*. 314542p.
- Mendes, P. 1993. GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Bioinforma.* 9(5): pp.563-571.

- Meyer, P. E., Lafitte, F., and Bontempi, G. 2008. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinforma.* 9(1):461p.
- Mordelet, F. and Vert, J. P. 2008. SIRENE: supervised inference of regulatory networks. *Bioinforma.* 24(16): pp.i76-i82.
- Msikita, W., Bissang, B., James, B. D., Baimey, H., Wilkinson, H. T., Ahounou, M., and Fagbemissi, R. 2005. Prevalence and severity of *Nattrassia mangiferae* root and stem rot pathogen of cassava in Benin. *Plant Dis.* 89(1): pp.12-16.
- Nassar, N. M. 2002. Cassava, *Manihot esculenta* Crantz, genetic resources: origin of the crop, its evolution and relationships with wild relatives. *Genet. Mol. Res.* 1(4): pp.298-305.
- Nassar, N. M., Hashimoto, D. Y. C., and Fernandes, S. D. C., 2008. Wild *Manihot* species: botanical aspects, geographic distribution and economic value. *Genet. Mol. Res.* 7(1): pp.16-28.
- Newman, M. E., 2006. Modularity and community structure in networks. *Proc. of the Natl. Acad. of Sci.* 103(23): pp.8577-8582.
- Nhassico, D., Muquingue, H., Cliff, J., Cumbana, A., and Bradbury, J. H. 2008. Rising African cassava production, diseases due to high cyanide intake and control measures. *J. of the Sci. of Food and Agric.* 88(12): pp.2043-2049.
- Ni, Y., Aghamirzaie, D., Elmarakeby, H., Collakova, E., Li, S., Grene, R., and Heath, L. S. 2016. A machine learning approach to predict gene regulatory networks in seed development in *Arabidopsis*. *Frontiers in Plant Sci.* 7:1936p.

- Olsen, K. M. and Schaal, B. A. 1999. Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. *Proc. of the Natl. Acad. of Sci.* 96(10): pp.5586-5591.
- Onik, M. M. H., Nobin, S. A., Ashrafi, A.F., and Chowdhury, T. M. 2018. Prediction of a gene regulatory network from gene expression Profiles with Linear Regression and Pearson Correlation Coefficient. *arXiv preprint arXiv:1805.01506*.
- Osbourn, A. E. 1996. Preformed Antimicrobial Compounds and Plant Defense against Fungal attack. *The Plant Cell* 8: 1821-1831.
- Osuna-Cruz, C. M., Paytuyi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Aiese Cigliano, R., Sanseverino, W., and Ercolano, M. R. 2017. PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.* (D1): pp. D1197-D1201.
- Padmanabhan, M., Cournoyer, P., and Dinesh-Kumar, S. P. 2009. The leucine-rich repeat domain in plant innate immunity: a wealth of possibilities. *Cell Microbiol.* 11(2): 191-198.
- Pandey, S. P. and Somissich, I. E. 2009. The Role of WRKY Transcription Factors in Plant Immunity *Cell* 150: 1648-1655.
- Papili Gao, N., Ud-Dean, S. M., Gandrillon, O. and Gunawan, R. 2017. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinforma.* 34(2): pp.258-266.
- Pérez-Quintero, Á.L., Quintero, A., Urrego, O., Vanegas, P. and López, C., 2012. Bioinformatic identification of cassava miRNAs differentially

expressed in response to infection by *Xanthomonas axonopodis* pv. *manihotis*. *BMC plant biology*, 12(1), p.29.

Pio, G., Malerba, D., D'Elia, D., and Ceci, M. 2014. Integrating microRNA target predictions for the discovery of gene regulatory networks: a semi-supervised ensemble learning approach. *BMC Bioinforma.* 15(1): S4p.

Plucknett, D. L., Phillips, T. P., and Kagbo, R. B., 1998. A global development strategy for cassava: Transforming a traditional tropical root crop. *Draft report prepared for the Int. Fund for Agric. Development* 99p. (work in progress).

Prochnik, S., Marri, P. R., Desany, B., Rabinowicz, P. D., Kodira, C., Mohiuddin, M., Rodriguez, F., Fauquet, C., Tohme, J., Harkins, T., and Rokhsar, D.S., 2012. The cassava genome: current progress, future directions. *Trop. Plant Biol.* 5(1): pp.88-94.

Rairdan, G. J. and Moffett, P. 2006. Distinct Domains in the ARC Region of the Potato Resistance Protein Rx Mediate LRR Binding and Inhibition of Activation. *The Plant Cell* 18: 2082-2093.

Rajamuthiah, R. and Eleftherios M. 2014. Effector Triggered Immunity: Activation of Innate Immunity in Metazoans by Bacterial Effectors. *Virulence* 5.7 (2014): 697–702. *PMC*.

Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., and Hughes, D.P. 2017. Deep learning for image-based cassava disease detection. *Frontiers in Plant Sci.* 8: 1852p.

Rau, C. D., Wisniewski, N., Orozco, L. D., Bennett, B., Weiss, J. N. and Lusi, A. J. 2013. Maximal information component analysis: a novel non-linear network analysis method. *Frontiers in Genet.* 4: 28p.

- Rausch, T. and Wachter, A. 2005. Sulfur metabolism: a versatile platform for launching defence operations. *Trends in Plant Sci.* 10(10): pp.503-509.
- Rodriguez, M. C. S., Peteresen, M and Mundy, J. 2010. Mitogen-Activated Protein Kinase Signaling in Plants. *Annu. Rev. Plant Biol.* 61: 621-649.
- Sales, G. and Romualdi, C. 2011. parmigene—a parallel R package for mutual information estimation and gene network reconstruction. *Bioinforma.* 27(13): pp.1876-1877.
- Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M. C., and Huang, Y. 2017. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinforma.* 34(6): pp.964-970.
- Schaffter, T., Marbach, D., and Floreano, D. 2011. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinforma.* 27(16): pp.2263-2270.
- Shi, Z., Derow, C. K., and Zhang, B. 2010. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Syst. Biol.* 4(1): 74p.
- Shmulevich, I., Dougherty, E. R., and Zhang, W. 2002. From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proc. of the IEEE.* 90(11): pp.1778-1792.
- Simoës, R. M. and Emmert-Streib, F. 2012. Bagging Statistical Network Inference from Large-Scale Gene Expression Data. *Bionforma.* 7(3): e33624p.
- Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. 2005. ROCr: visualizing classifier performance in R. *Bioinforma.* 21(20): pp.3940-3941.

- Sławek, J. and Arodź, T. 2013. ENNET: inferring large gene regulatory networks from expression data using gradient boosting. *BMC Syst. Biol.* 7(1):106p.
- Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J., and Jarvis, E.D. 2006. Computational inference of neural information flow networks. *PLoS Computational Biol.* 2(11): e161p.
- Sonego, P., Kocsor, A., and Pongor, S. 2008. ROC analysis: applications to the classification of biological sequences and 3D structures. *Briefings in Bioinforma.* 9(3): pp.198-209.
- Soto, J. C., Ortiz, J. F., Perlaza-Jiménez, L., Vásquez, A. X., Lopez-Lavalle, L. A. B., Mathew, B., León, J., Bernal, A. J., Ballvora, A., and López, C. E. 2015. A genetic map of cassava (*Manihot esculenta* Crantz) with integrated physical mapping of immunity-related genes. *BMC Genomics* 16(1):190p.
- Tameling, W. L., Vossen, J. H., Albrecht, M., Lengauer, T., Berden, J. A., and Haring, M. A., Cornelissen, B. C., and Takken, F. L. W. 2006. Mutations in the NB-ARC Domain of I-2 That Impair ATP Hydrolysis Cause Autoactivation. *Plant Physiol.* 140: 1233-1245.
- Tanaka, H., Chiba, H., Inokoshi, J., Kuno, A., Sugai, T., Takahashi, A., Ito, Y., Tsunoda, M., Suzuki, K., Takenaka, A., Sekiguchi, T., Umeyama, H., Hirabayashi, J., and Mura, S. O. 2009. Mechanism by which the lectin actinohivin blocks HIV infection of target cells. *PNAS.* 106(37): 15633-15638.
- Tanne, A. and Neyrolles, O. C-type lectins in immune defense against pathogens: the murine DC-SIGN homologue SIGNR3 confers early protection against *Mycobacterium tuberculosis* infection. *PNAS.* 1(4): 285-290.

- Taylor, R. C., Shah, A., Treatman, C., and Blevins, M. 2006. SEBINI: software environment for biological network inference. *Bioinforma.* 22(21): pp.2706-2708.
- Tena, G., Boudsocq, M., and Sheen, J. 2011. Protein Kinase signalling networks in plant innate immunity. *Curr. Opinion Plant Biol.* 14(5): 519-529.
- Tong, Y., Sun, J., Wong, C. F., Kang, Q., Ru, B., Wong, C. N., Chan, A. S., Leung, S. Y., and Zhang, J. 2018. MICMIC: identification of DNA methylation of distal regulatory regions with causal effects on tumorigenesis. *Genome Biol.* 19(1):73p.
- Tripathi, S., Lloyd-Price, J., Ribeiro, A., Yli-Harja, O., Dehmer, M., and Emmert-Streib, F. 2017. sgenesR: An R package for simulating gene expression data from an underlying real gene network structure considering delay parameters. *BMC Bioinforma.* 18(1):325p.
- Unteawati, B. and Fatih, C. 2018, March. Consumer's market analysis of products based on cassava. In *IOP Conf. Series: Earth and Environ. Sci.* 141(1):012033p. IOP Publishing.
- Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., De Moor, B., and Marchal, K. 2006. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinforma.* 7(1):43p.
- Varala, K., Marshall-Colón, A., Cirrone, J., Brooks, M. D., Pasquino, A. V., Léran, S., Mittal, S., Rock, T. M., Edwards, M. B., Kim, G. J., and Ruffel, S. 2018. Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants. *Proc. of the Natl. Acad. of Sci.* 115(25): pp.6494-6499.

- Verdier, V., Mosquera, G., and Assigbétsé, K. 1998. Detection of the cassava bacterial blight pathogen, *Xanthomonas axonopodis* pv. *manihotis*, by polymerase chain reaction. *Plant Dis.* 82(1): pp.79-83.
- Wagaba, H., Patil, B. L., Mukasa, S., Alicai, T., Fauquet, C. M., and Taylor, N. J. 2016. Artificial microRNA-derived resistance to Cassava brown streak disease. *J. of Virol. Methods* 231:pp.38-43.
- Wang, J., Li, M., Deng, Y., and Pan, Y. 2010. Recent advances in clustering methods for protein interaction networks. *BMC Genomics* 11(3): S10p.
- Wang, Y. and Bouwmeester. 2018. K. L-type lectin receptor kinases: New forces in plant immunity. *PLoS Pathol.* 13(8): e1006433.
- Wessling-Resnick, M. 2015. Nramp1 and other transporters involved in metal withholding during infection. Published, JBC Papers in Press, June 8, 2015, DOI 10.1074/jbc.R115.643973.
- Wilkins, S. 2015. Structure and mechanism of ABC transporters. *Prime Rep.* 7:14 (doi:10.12703/P7-14)
- Wils, C.R. and Kaufmann, K. 2017. Gene-regulatory networks controlling inflorescence and flower development in *Arabidopsis thaliana*. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1860(1): pp.95-105.
- Xing, L., Guo, M., Liu, X., Wang, C., Wang, L. and Zhang, Y. 2017. An improved Bayesian network method for reconstructing gene regulatory network based on candidate auto selection. *BMC Genomics* 18(9): 844p.
- Yamamoto, K. 2014. Intracellular lectins are involved in quality control of glycoproteins. *Proc. Jpn. Acad., Ser. B* 90.

- Yi, S. M. P., Harson, R. E., Zabner, J., and Welsh, M. J. 2001. Lectin binding and endocytosis at the apical surface of human airway epithelia. *Gene Therapy*. 8: 1826-1832.
- Zhang, S., Chen, X., Lu, C., Ye, J., Zou, M., Lu, K., Feng, S., Pei, J., Liu, C., Zhou, X., Ma, P., Li, Z., Liu, C., Liao, Q., Xia, Z., and Wang, W. 2018. Genome-Wide Association Studies of 11 Agronomic Traits in Cassava (*Manihot esculenta* Crantz). *Original Res.* doi: 10.3389/fpls.2018.00503.
- Zhang, X., Bernoux, M., Bentham, A. R., Newman, T. E., Ve, T., Casey, L. W., Raaymakers, T. M., Hu, J., Croll, T. I., Schreiber, K. J., Staskawicz, B. J., Anderson, P. A., Sohn, K. H., Williams, S. J., Dodd, P. N., and Kobe, B. 2014. Multiple functional self-association interfaces in plant TIR domains. www.pnas.org/cgi/doi/10.1073/pnas.1621248114.
- Zhang, X., Cannon, S. B., and Stacey, G. 2009. Evolutionary genomics of LysM genes in land plants. *BMC Evolutionary Biol.* 9: 183p.

APPENDICES

APPENDIX I LIST OF PREDICTED IMMUNITY RELATED GENES			STRING PROTEIN
GENE IDENTIFIED	FUNCTION ANNOTATED		
cassava4.1_002469m	ABC transporter F family member 4-like [Manihot esculenta]		ABCF4
cassava4.1_000762m	pleiotropic drug resistance protein 1-like [Manihot esculenta]		ABCG40
cassava4.1_033810m	pleiotropic drug resistance protein 1-like [Manihot esculenta]		ABCG40
cassava4.1_000398m	ABC transporter B family member 13-like [Manihot esculenta]		ABCB13
cassava4.1_000219m	ABC transporter C family member 10 [Manihot esculenta]		ABCC10
cassava4.1_003332m	ABC transporter B family member 29, chloroplastic [Manihot esculenta]		ABCB29
cassava4.1_001064m	ABC transporter B family member 15-like [Manihot esculenta]		ABCB15
cassava4.1_022144m	ABC transporter G family member 17-like [Manihot esculenta]		ABCG17
cassava4.1_032403m	hypothetical protein MANES_12G015700 [Manihot esculenta]		ABCC3
cassava4.1_000345m	ABC transporter B family member 11-like [Manihot esculenta]		ABCB11
cassava4.1_031477m	ABC transporter G family member 28-like isoform X1 [Manihot esculenta]		ABCG28
cassava4.1_030988m	ABC transporter B family member 9 [Manihot esculenta]		ABCB9
cassava4.1_000410m	putative ABC transporter B family member 8 [Manihot esculenta]		ABCB15
cassava4.1_033109m	ABC transporter B family member 11-like [Manihot esculenta]		ABCB11
cassava4.1_000369m	ABC transporter B family member 11-like [Manihot esculenta]		ABCB11
cassava4.1_002950m	ABC transporter G family member 11-like, partial [Manihot esculenta]		ABCG11
cassava4.1_003511m	ABC transporter G family member 11-like [Manihot esculenta]		ABCG11
cassava4.1_002690m	ABC transporter G family member 11-like [Manihot esculenta]		ABCG11
cassava4.1_023286m	ABC transporter G family member 31 [Manihot esculenta]		ABCG31
cassava4.1_025247m	ABC transporter B family member 15-like [Manihot esculenta]		ABCB15
cassava4.1_002808m	ABC transporter G family member 15-like isoform X1 [Manihot esculenta]		ABCG15
cassava4.1_004081m	ABC transporter F family member 1 [Manihot esculenta]		ABCF1
cassava4.1_025467m	ABC transporter G family member 17-like [Manihot esculenta]		ABCG16
cassava4.1_000182m	ABC transporter C family member 5 isoform X2 [Manihot esculenta]		ABCC5
cassava4.1_013263m	ABC transporter I family member 11, chloroplastic [Manihot esculenta]		ABCI11
cassava4.1_000306m	ABC transporter B family member 1 [Manihot esculenta]		ABCB1
cassava4.1_033024m	hypothetical protein MANES_11G148900, partial [Manihot esculenta]		ABCG28

cassava4.1_002954m	ABC transporter G family member 21 [Manihot esculenta]	ABCG21
cassava4.1_002589m	ABC transporter F family member 3 isoform X1 [Manihot esculenta]	ABCF3
cassava4.1_003397m	ABC transporter F family member 3 isoform X2 [Manihot esculenta]	ABCF3
cassava4.1_023079m	hypothetical protein MANES_02G013700 [Manihot esculenta]	ABCC10
cassava4.1_002812m	ABC transporter G family member 15-like isoform X2 [Manihot esculenta]	ABCG15
cassava4.1_004508m	ABC transporter G family member 25-like [Manihot esculenta]	ABCG25
cassava4.1_025786m	ABC transporter C family member 5-like [Manihot esculenta]	ABCC5
cassava4.1_015426m	ABC transporter I family member 1 [Manihot esculenta]	ABCI1
cassava4.1_000238m	pleiotropic drug resistance protein 2-like isoform X1 [Manihot esculenta]	ABCG39
cassava4.1_025593m	ABC transporter G family member 1-like [Manihot esculenta]	ABCG1
cassava4.1_025724m	hypothetical protein MANES_04G133800 [Manihot esculenta]	ABCG36
cassava4.1_000220m	ABC transporter C family member 8 [Manihot esculenta]	ABCC8
cassava4.1_024602m	pleiotropic drug resistance protein 2-like [Manihot esculenta]	ABCG34
cassava4.1_003442m	ABC transporter G family member 5-like isoform X1 [Manihot esculenta]	ABCG5
cassava4.1_002248m	ABC transporter G family member 6-like [Manihot esculenta]	ABCG16
cassava4.1_001113m	ABC transporter A family member 2-like [Manihot esculenta]	ABCA2
cassava4.1_000213m	ABC transporter C family member 3-like [Hevea brasiliensis]	ABCC3
cassava4.1_000215m	ABC transporter C family member 3-like [Manihot esculenta]	ABCC3
cassava4.1_005504m	ABC transporter B family member 11-like [Manihot esculenta]	ABCB11
cassava4.1_025686m	hypothetical protein MANES_08G061500 [Manihot esculenta]	ABCB21
cassava4.1_000260m	ABC transporter G family member 32 [Manihot esculenta]	ABCG32
cassava4.1_000385m	ABC transporter B family member 19-like [Manihot esculenta]	ABCB19
cassava4.1_001419m	ABC transporter C family member 2-like [Manihot esculenta]	ABCC2
cassava4.1_021533m	ABC transporter C family member 12-like [Manihot esculenta]	ABCC1
cassava4.1_002526m	ABC transporter G family member 7 isoform X1 [Manihot esculenta]	ABCG7
cassava4.1_000311m	ABC transporter D family member 1-like [Manihot esculenta]	ABCD1
cassava4.1_013337m	ABC transporter I family member 17-like isoform X1 [Manihot esculenta]	ABCI17
cassava4.1_013402m	ABC transporter I family member 17-like isoform X2 [Manihot esculenta]	ABCI17
cassava4.1_000386m	ABC transporter B family member 19 [Manihot esculenta]	ABCB19
cassava4.1_032720m	putative ABC transporter C family member 15 [Manihot esculenta]	ABCC3
cassava4.1_002804m	ABC transporter G family member 11 [Manihot esculenta]	ABCG11

cassava4.1_032702m	ABC transporter G family member 9 [Manihot esculenta]	ABCG9
cassava4.1_004013m	ABC transporter B family member 1 [Manihot esculenta]	ABCB1
cassava4.1_003953m	ABC transporter E family member 2 [Manihot esculenta]	ABCE2
cassava4.1_003954m	ABC transporter E family member 2 [Manihot esculenta]	ABCE2
cassava4.1_023380m	ABC transporter E family member 2-like isoform X2 [Manihot esculenta]	ABCE2
cassava4.1_032500m	ABC transporter G family member 10-like [Manihot esculenta]	ABCG10
cassava4.1_000583m	ABC transporter G family member 24-like [Manihot esculenta]	ABCG24
cassava4.1_021886m	ABC transporter C family member 4-like [Manihot esculenta]	ABCC4
cassava4.1_003297m	ABC transporter G family member 23 [Manihot esculenta]	ABCG23
cassava4.1_029148m	ABC transporter A family member 1 isoform X1 [Manihot esculenta]	ABCA1
cassava4.1_026769m	hypothetical protein MANES_03G143200 [Manihot esculenta]	ABCG26
cassava4.1_013814m	ABC transporter I family member 17-like [Manihot esculenta]	ABCI17
cassava4.1_002838m	ABC transporter F family member 4-like [Manihot esculenta]	ABCF4
cassava4.1_010982m	protein TRIGALACTOSYLDIACYLGLYCEROL 3, chloroplastic-like [Manihot esculenta]	ABCI13
cassava4.1_032931m	pleiotropic drug resistance protein 1-like isoform X2 [Hevea brasiliensis]	ABCG40
cassava4.1_000241m	pleiotropic drug resistance protein 1-like isoform X1 [Manihot esculenta]	ABCG40
cassava4.1_031766m	ABC transporter G family member 29-like [Manihot esculenta]	ABCG29
cassava4.1_000384m	ABC transporter B family member 15-like [Manihot esculenta]	ABCB15
cassava4.1_000399m	ABC transporter B family member 2-like [Manihot esculenta]	ABCB28
cassava4.1_002547m	ABC transporter B family member 28 [Manihot esculenta]	ABCB28
cassava4.1_002998m	ABC transporter B family member 26, chloroplastic isoform X1 [Manihot esculenta]	ABCB26
cassava4.1_023047m	ABC transporter G family member 10 [Manihot esculenta]	ABCG10
cassava4.1_027677m	pleiotropic drug resistance protein 1-like isoform X2 [Manihot esculenta]	ABCG40
cassava4.1_000229m	hypothetical protein MANES_11G063400 [Manihot esculenta]	ABCG40
cassava4.1_028674m	pleiotropic drug resistance protein 1-like [Manihot esculenta]	ABCG40
cassava4.1_012274m	hypothetical protein MANES_02G056700 [Manihot esculenta]	ABCB27
cassava4.1_003515m	ABC transporter G family member 14-like [Manihot esculenta]	ABCG14
cassava4.1_020897m	ABC transporter G family member 25 [Manihot esculenta]	ABCG25
cassava4.1_000209m	putative ABC transporter C family member 15 isoform X1 [Manihot esculenta]	ABCC9

cassava4.1_000205m	ABC transporter C family member 4-like [Manihot esculenta]	ABCC4
cassava4.1_000427m	ABC transporter B family member 20-like [Manihot esculenta]	ABCB20
cassava4.1_002648m	ABC transporter F family member 5 [Manihot esculenta]	ABCF5
cassava4.1_032786m	ABC transporter G family member 11-like [Manihot esculenta]	ABCG11
cassava4.1_000251m	pleiotropic drug resistance protein 1-like [Manihot esculenta]	ABCG40
cassava4.1_024510m	ABC transporter B family member 21-like [Manihot esculenta]	ABCB21
cassava4.1_003496m	ABC transporter G family member 14-like [Manihot esculenta]	ABCG14
cassava4.1_002331m	ABC transporter G family member 22-like [Manihot esculenta]	ABCG22
cassava4.1_026648m	ABC transporter B family member 19-like [Manihot esculenta]	ABCB19
cassava4.1_003665m	ABC transporter G family member 5-like [Manihot esculenta]	ABCG5
cassava4.1_010779m	protein TRIGALACTOSYLDIACYLGLYCEROL 3, chloroplastic-like [Manihot esculenta]	ABCI13
cassava4.1_021429m	hypothetical protein MANES_05G196700 [Manihot esculenta]	ABCB1
cassava4.1_001856m	ABC transporter G family member 22-like isoform X2 [Manihot esculenta]	ABCG22
cassava4.1_027732m	probable LRR receptor-like serine/threonine-protein kinase At4g36180 [Manihot esculenta]	RLP9
cassava4.1_000670m	probable LRR receptor-like serine/threonine-protein kinase At1g34110 [Manihot esculenta]	AT1G34110
cassava4.1_001097m	leucine-rich repeat receptor-like tyrosine-protein kinase PXC3 [Manihot esculenta]	AT1G34420
cassava4.1_000658m	LRR receptor-like serine/threonine-protein kinase [Manihot esculenta]	AT5G56040
cassava4.1_011715m	probable leucine-rich repeat receptor-like protein kinase At1g35710 [Manihot esculenta]	AT5G56040
cassava4.1_000983m	receptor protein kinase CLAVATA1 [Manihot esculenta]	CLV1
cassava4.1_029345m	hypothetical protein MANES_17G067000 [Manihot esculenta]	MPL12.8
cassava4.1_024812m	receptor like protein 30-like [Manihot esculenta]	RLP12
cassava4.1_022591m	receptor-like protein 12 [Manihot esculenta]	EFR
cassava4.1_032707m	receptor-like protein 12 [Manihot esculenta]	RLP6
cassava4.1_033177m	LRR receptor-like serine/threonine-protein kinase GSO1 [Manihot esculenta]	GSO1
cassava4.1_002234m	PREDICTED: probable LRR receptor-like serine/threonine-protein kinase IRK [Nicotiana tabacum]	AT3G56370

cassava4.1_026412m	LRR receptor-like serine/threonine-protein kinase GSO2 [Manihot esculenta]	AT2G34930
cassava4.1_000978m	receptor-like protein kinase HAIKU2 [Manihot esculenta]	At1g09970
cassava4.1_0011107m	leucine-rich repeat receptor-like protein kinase PXC2 [Manihot esculenta]	AT5G01890
cassava4.1_034159m	serine/threonine-protein kinase BRI1-like 2 [Manihot esculenta]	BRL2
cassava4.1_031896m	probable LRR receptor-like serine/threonine-protein kinase At4g36180, partial [Manihot esculenta]	RLP14
cassava4.1_030257m	receptor-like protein kinase 2 [Manihot esculenta]	AT3G24240
cassava4.1_000689m	probable leucine-rich repeat receptor-like protein kinase At2g33170 [Manihot esculenta]	AT2G34930
cassava4.1_020964m	probable LRR receptor-like serine/threonine-protein kinase At4g36180, partial [Manihot esculenta]	RLP13
cassava4.1_033985m	hypothetical protein MANES_14G136900 [Manihot esculenta]	RLP45
cassava4.1_000967m	leucine-rich repeat receptor-like serine/threonine-protein kinase BAM3 isoform X1 [Manihot esculenta]	MPA24.5
cassava4.1_030027m	hypothetical protein MANES_S047800 [Manihot esculenta]	RLP15
cassava4.1_031576m	probable LRR receptor-like serine/threonine-protein kinase At1g34110 [Manihot esculenta]	AT2G34930
cassava4.1_001023m	receptor protein-tyrosine kinase CEPR2-like [Manihot esculenta]	AT1G72180
cassava4.1_021898m	probable inactive leucine-rich repeat receptor kinase XIAO [Manihot esculenta]	AT2G34930
cassava4.1_030238m	LRR receptor-like serine/threonine-protein kinase FLS2 [Hevea brasiliensis]	MPL12.8
cassava4.1_032461m	probably inactive leucine-rich repeat receptor-like protein kinase IMK2 [Manihot esculenta]	IMK2
cassava4.1_025624m	receptor-like protein 12 [Manihot esculenta]	RLP33
cassava4.1_000873m	leucine-rich repeat receptor-like protein kinase TDR [Manihot esculenta]	PXY
cassava4.1_001117m	hypothetical protein MANES_06G138900 [Manihot esculenta]	ERL1
cassava4.1_034192m	phytosulfokine receptor 1-like [Manihot esculenta]	PSKR1
cassava4.1_029173m	hypothetical protein MANES_07G034400 [Manihot esculenta]	AT2G34930
cassava4.1_001861m	probably inactive leucine-rich repeat receptor-like protein kinase IMK2 [Manihot esculenta]	IMK2
cassava4.1_032357m	receptor-like protein kinase 5 [Manihot esculenta]	AT5G25930

cassava4.1_001164m	receptor protein kinase TMK1-like [Manihot esculenta]	TMK1
cassava4.1_025332m	receptor-like protein 12 [Manihot esculenta]	RLP6
cassava4.1_028371m	receptor-like protein 12 [Manihot esculenta]	RLP7
cassava4.1_025772m	hypothetical protein MANES_06G040100 [Manihot esculenta]	RLP45
cassava4.1_001257m	receptor protein-tyrosine kinase CEPR2-like [Manihot esculenta]	AT1G72180
cassava4.1_000567m	receptor-like protein 12 [Manihot esculenta]	RLP7
cassava4.1_002496m	leucine-rich repeat receptor-like protein kinase PXC2 [Manihot esculenta]	AT5G01890
cassava4.1_004649m	plant intracellular Ras-group-related LRR protein 4-like [Manihot esculenta]	PIRL4
cassava4.1_008123m	LRR receptor-like serine/threonine-protein kinase FLS2 [Manihot esculenta]	AT5G66330
cassava4.1_021460m	probable LRR receptor-like serine/threonine-protein kinase At4g36180 isoform X1 [Manihot esculenta]	AT4G36180
cassava4.1_000430m	receptor-like protein kinase BRI1-like 3 [Manihot esculenta]	BRL1
cassava4.1_026118m	probable LRR receptor-like serine/threonine-protein kinase At4g36180 [Manihot esculenta]	RLP56
cassava4.1_000643m	LRR receptor-like serine/threonine-protein kinase [Manihot esculenta]	AT5G56040
cassava4.1_028544m	hypothetical protein MANES_11G060100 [Manihot esculenta]	RLP13
cassava4.1_021773m	LRR receptor-like serine/threonine-protein kinase RPK2 [Manihot esculenta]	RPK2
cassava4.1_033957m	MDIS1-interacting receptor like kinase 1-like [Manihot esculenta]	AT4G28650
cassava4.1_001201m	receptor protein-tyrosine kinase CEPR1-like [Manihot esculenta]	XIP1
cassava4.1_001017m	receptor-like protein kinase HAIKU2 [Manihot esculenta]	At1g09970
cassava4.1_026823m	putative receptor-like protein kinase At3g47110 [Manihot esculenta]	AT3G47570
cassava4.1_000663m	receptor-like protein kinase 2 [Manihot esculenta]	AT5G48940
cassava4.1_023549m	receptor like protein 30-like [Manihot esculenta]	GSO1
cassava4.1_003608m	LRR receptor-like serine/threonine-protein kinase GSO1 [Manihot esculenta]	RLP56
cassava4.1_033390m		RLP1
cassava4.1_000624m	probable LRR receptor-like serine/threonine-protein kinase At1g74360 [Manihot esculenta]	AT1G74360
cassava4.1_022906m	LRR receptor-like serine/threonine-protein kinase GSO1 [Manihot esculenta]	AT2G34930
cassava4.1_025890m	probable LRR receptor-like serine/threonine-protein kinase At4g36180, partial [Manihot esculenta]	ABCF3

cassava4.1_029680m	plant intracellular Ras-group-related LRR protein 9-like [Manihot esculenta]	PIRL9
cassava4.1_033286m	hypothetical protein MANES_06G042100 [Manihot esculenta]	RLP15
cassava4.1_004309m	plant intracellular Ras-group-related LRR protein 6 [Manihot esculenta]	AT3G15410
cassava4.1_000927m	receptor-like protein kinase HSL1 [Manihot esculenta]	At1g28440
cassava4.1_023795m	LRR receptor-like serine/threonine-protein kinase ERECTA [Manihot esculenta]	TE1
cassava4.1_025780m	receptor-like protein 12 [Manihot esculenta]	RLP15
cassava4.1_000950m	LRR receptor-like serine/threonine-protein kinase ERECTA [Manihot esculenta]	TE1
cassava4.1_000470m	protein BRASSINOSTEROID INSENSITIVE 1-like [Manihot esculenta]	BR11
cassava4.1_004350m	uncharacterized protein LOC110608538 isoform X2 [Manihot esculenta]	AT4G23840
cassava4.1_005550m	uncharacterized protein LOC110608538 isoform X3 [Manihot esculenta]	AT4G23840
cassava4.1_025984m	hypothetical protein MANES_07G071400 [Manihot esculenta]	AT2G34930
cassava4.1_000469m	systemin receptor SR160-like [Manihot esculenta]	BR11
cassava4.1_029023m	DNA damage-repair/tolerance protein DRT100-like [Manihot esculenta]	AT1G33590
cassava4.1_001074m	probably inactive leucine-rich repeat receptor-like protein kinase At2g25790 [Manihot esculenta]	AT2G25790
cassava4.1_030753m	MDIS1-interacting receptor like kinase 1-like [Manihot esculenta]	AT4G28650
cassava4.1_004409m		AT1G78230
cassava4.1_004410m.		AT1G78230
cassava4.1_026529m	piriformospora indica-insensitive protein 2-like [Manihot esculenta]	RLP29
cassava4.1_000520m	LRR receptor-like serine/threonine-protein kinase FLS2 [Manihot esculenta]	MPL12.8
cassava4.1_001910m	hypothetical protein MANES_03G087100 [Manihot esculenta]	RLP46
cassava4.1_030307m	probable LRR receptor-like serine/threonine-protein kinase At4g36180 [Manihot esculenta]	RLP13
cassava4.1_022484m	hypothetical protein MANES_14G136500 [Manihot esculenta]	RLP14
cassava4.1_030243m	hypothetical protein MANES_14G136500 [Manihot esculenta]	RLP56

cassava4.1_021384m	hypothetical protein MANES_02G015800 [Manihot esculenta]	RLP15
cassava4.1_031679m		RLP45
cassava4.1_031485m	hypothetical protein MANES_14G136500 [Manihot esculenta]	RLP15
cassava4.1_025412m	phytosulfokine receptor 1-like [Manihot esculenta]	RLP1
cassava4.1_025658m	probable LRR receptor-like serine/threonine-protein kinase At4g36180 [Manihot esculenta]	RLP45
cassava4.1_027765m	DNA damage-repair/tolerance protein DRT100 [Manihot esculenta]	AT5G23400
cassava4.1_000644m	tyrosine-sulfated glycopeptide receptor 1 [Manihot esculenta]	PSY1R
cassava4.1_022271m	receptor-like protein 12 [Manihot esculenta]	AT2G34930
cassava4.1_001145m	LRR receptor-like serine/threonine-protein kinase HSL2 [Manihot esculenta]	HSL2
cassava4.1_026756m	receptor protein kinase CLAVATA1-like isoform X1 [Manihot esculenta]	CLV1
cassava4.1_027671m	probable LRR receptor-like serine/threonine-protein kinase At3g47570 [Manihot esculenta]	AT3G47570
cassava4.1_007587m	plant intracellular Ras-group-related LRR protein 9-like [Manihot esculenta]	PIRL9
cassava4.1_000708m	LRR receptor-like serine/threonine-protein kinase [Manihot esculenta]	AT4G26540
cassava4.1_022691m	LRR receptor-like serine/threonine-protein kinase GSO1 [Manihot esculenta]	AT2G34930
cassava4.1_028316m	leucine-rich repeat receptor-like serine/threonine-protein kinase BAM3 [Manihot esculenta]	MPA24.5
cassava4.1_032164m	hypothetical protein MANES_10G114900 [Manihot esculenta]	AT3G47570
cassava4.1_000962m	receptor-like protein kinase HSL1 [Manihot esculenta]	Atlg28440
cassava4.1_023351m	probable leucine-rich repeat receptor-like protein kinase At2g33170 [Manihot esculenta]	AT2G33170
cassava4.1_032910m	hypothetical protein MANES_S057200, partial [Manihot esculenta]	RLP45
cassava4.1_001882m	leucine-rich repeat receptor protein kinase EMS1-like isoform X2 [Manihot esculenta]	RLP15
cassava4.1_004421m		AT1G15740
cassava4.1_004417m		AT1G15740
cassava4.1_004418m		AT1G15740
cassava4.1_033500m	leucine-rich repeat receptor protein kinase EMS1-like [Manihot esculenta]	RLP15

cassava4.1_034276m	LRR receptor-like serine/threonine-protein kinase GSO2, partial [Manihot esculenta]	AT2G34930
cassava4.1_031066m	leucine-rich repeat receptor-like protein kinase PXC2 [Manihot esculenta]	AT5G66330
cassava4.1_025415m	receptor-like protein 12 [Manihot esculenta]	RLP33
cassava4.1_034154m	probable LRR receptor-like serine/threonine-protein kinase IRK [Manihot esculenta]	AT2G15320
cassava4.1_026317m	probable leucine-rich repeat receptor-like protein kinase At5g63930 [Manihot esculenta]	AT2G34930
cassava4.1_000577m	leucine-rich repeat receptor-like protein kinase PEPR1 [Manihot esculenta]	PEPR1
cassava4.1_033635m	probable LRR receptor-like serine/threonine-protein kinase At1g12460 [Manihot esculenta]	AT1G12460
cassava4.1_000930m	receptor-like protein kinase HSL1 [Manihot esculenta]	At1g28440
cassava4.1_000947m	receptor-like protein kinase HAIKU2 [Manihot esculenta]	At1g09970
cassava4.1_024080m	hypothetical protein MANES_12G064900 [Manihot esculenta]	RLP6
cassava4.1_021358m	hypothetical protein MANES_S040700, partial [Manihot esculenta]	RLP56
cassava4.1_001024m	probably inactive leucine-rich repeat receptor-like protein kinase At2g25790 isoform X1 [Manihot esculenta]	AT2G25790
cassava4.1_011734m	probable leucine-rich repeat receptor-like protein kinase At1g35710 [Manihot esculenta]	AT5G61240
cassava4.1_025489m	hypothetical protein MANES_S053800 [Manihot esculenta]	RLP9
cassava4.1_004516m	plant intracellular Ras-group-related LRR protein 4-like [Manihot esculenta]	PIRL4
cassava4.1_023757m	hypothetical protein MANES_01G142200 [Manihot esculenta]	AT2G33170
cassava4.1_001224m	receptor-like kinase TMK4 [Manihot esculenta]	AT3G23750
cassava4.1_001398m	probably inactive leucine-rich repeat receptor-like protein kinase At5g06940 [Manihot esculenta]	AT5G06940
cassava4.1_025435m	hypothetical protein MANES_14G142300 [Manihot esculenta]	AT2G24130
cassava4.1_001118m	leucine-rich repeat receptor-like protein kinase TDR [Manihot esculenta]	MOL1
cassava4.1_034145m	probable LRR receptor-like serine/threonine-protein kinase At2g24230 [Manihot esculenta]	AT2G24230

cassava4.1_030915m	receptor-like protein 12 [Manihot esculenta]	RLP33
cassava4.1_025627m	receptor-like kinase TMK4 [Manihot esculenta]	AT3G23750
cassava4.1_001694m	putative leucine-rich repeat receptor-like serine/threonine-protein kinase At2g24130 isoform X1 [Manihot esculenta]	AT2G24130
cassava4.1_027458m	LRR receptor-like serine/threonine-protein kinase FLS2 [Manihot esculenta]	MPL12.8
cassava4.1_009924m	LRR receptor-like serine/threonine-protein kinase ERL2 [Manihot esculenta]	AT1G03440
cassava4.1_031410m	probable inactive leucine-rich repeat receptor kinase XIAO [Hevea brasiliensis]	AT2G34930
cassava4.1_028695m	probable LRR receptor-like serine/threonine-protein kinase At3g47570 [Manihot esculenta]	AT5G63930
cassava4.1_000584m	hypothetical protein MANES_08G142700 [Manihot esculenta]	AT5G63930
cassava4.1_031530m	hypothetical protein MANES_10G083000 [Manihot esculenta]	AT4G08850

125

**COMPARATIVE EVALUATION OF TOOLS FOR GENE REGULATORY
NETWORK PREDICTION AND NETWORK RECONSTRUCTION
USING GENOMIC DATA**

By

RESHMA BHASKER T.

(2013-09-103)

Abstract of Thesis

Submitted in partial fulfilment of the

Requirement for the degree of

B. Sc. – M. Sc. (INTEGRATED) BIOTECHNOLOGY

Faculty of Agriculture

Kerala Agricultural University, Thrissur



B. Sc. – M. Sc. (INTEGRATED) BIOTECHNOLOGY

DEPARTMENT OF PLANT BIOTECHNOLOGY

COLLEGE OF AGRICULTURE

VELLAYANI, THIRUVANANTHAPURAM-695522

KERALA, INDIA

2018

9. ABSTRACT

Developing regulatory network of genes controlling traits which are of importance economically and commercially are gaining much significance in present times. GRN's provide an insight into the transcriptional mechanisms that regulate the robust and stochastic gene expression and their relationship with the phenotypic variability that can be utilized for better crop improvement strategies. The former approaches for Gene Regulatory Network construction mainly rely on using gene expression data as input, but the time consumption and high cost of expression analysis paved way for developing new methodologies that make GRN prediction easier.

The integration of genomic information along with gene expression data, could make the process of Gene Regulatory Network (GRN) construction more reliable than using expression data alone as input source. Using this approach, we have tried to develop the regulatory network of genes controlling immunity in cassava with special context to Bacterial blight resistance. Initially the immunity related genes in cassava were identified by protein domain search and analysis using HMMER. Cassava specific genes were further filtered for high competency, mapped and annotated to determine its biological role and function. A set of 1919 immunity related genes in cassava were identified, out of which 22 of them were specifically conferring virus resistance, 727 of them were screened for bacterial blight resistance by microarray data integration and a network was created using the predicted interactions identified from 324 genes using STRING. The network obtained was visualized using Cytoscape and cross validated with simulated dataset generated from SynTReN. The generated network of immunity related genes in cassava could give more insight into the defence mechanism in cassava that can help in adapting better crop improvement and management strategies.

A comparison of various approaches used for GRN prediction like probabilistic method, mutual information-based method, correlation-based approaches etc was also done and various tools like ARACNE, WGCNA etc were

127

evaluated. Networks with different sizes, 50, 100 and 150 was generated and the network parameters like clustering coefficient, network density etc were compared. Clustering coefficient does not seem to vary with increase in network size but network heterogeneity and density were observed to increase. The statistical analysis of the performance of different methods resulted into a conclusion that the mutual information-based approaches are better tools for Gene Regulatory Network construction than the other methods and it performed with a specificity of 75.7% and a sensitivity of 79.4%.

174552

