

**MODELING OF CASSAVA-CASSAVA MOSAIC VIRUS
INTERACTIONS WITH COMPUTATIONAL BIOLOGY AND
BIOINFORMATICS APPROACH**

By

RAJANI K. R.

(2014-09-105)

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



**DEPARTMENT OF PLANT BIOTECHNOLOGY
COLLEGE OF AGRICULTURE
VELLAYANI, THIRUVANANTHAPURAM-695 522
KERALA, INDIA**

2019

DECLARATION

I, hereby declare that this thesis entitled “**Modeling of Cassava-Cassava Mosaic Virus interactions with computational biology and bioinformatics approach**” 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, associateship, fellowship or other similar title, of any other University or Society.

Place: Vellayani

Date: 29/11/2019



RAJANI K. R.

(2014-09-105)

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

(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 “**MODELING OF CASSAVA-CASSAVA MOSAIC VIRUS INTERACTIONS WITH COMPUTATIONAL BIOLOGY AND BIOINFORMATICS APPROACH**” is a record of research work done independently by **Ms. RAJANI K. R. (2014-09-105)** under my guidance and supervision and this has not previously formed the basis for the award of any degree, diploma, fellowship or associateship to her.

Place: Sreekariyam

Date: 29-11-2019

Dr. J. Sreekumar

(Chairman, Advisory Committee)

Principal Scientist (Agrl. Statistics),

Section of Extension and Social Sciences,

ICAR-CTCRI, Sreekariyam

Thiruvananthapuram-695 017

CERTIFICATE

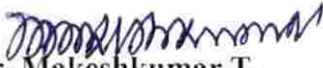
We, the undersigned members of the advisory committee of Ms. Rajani K. R. (2014-09-105) a candidate for the degree of B. Sc. - M. Sc. (Integrated) Biotechnology, agree that the thesis entitled **"MODELING OF CASSAVA-CASSAVA MOSAIC VIRUS INTERACTIONS WITH COMPUTATIONAL BIOLOGY AND BIOINFORMATICS APPROACH"** may be submitted by Ms. Rajani K. R. in partial fulfillment of the requirement for the degree.



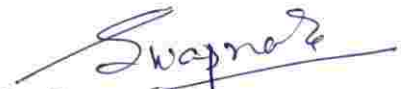
Dr. J. Sreekumar
Chairman, Advisory committee
Principal Scientist (Agrl. Statistics),
Section of Extension and
Social Sciences.
ICAR-CTCRI, Sreekariyam,
Thiruvananthapuram - 695 017



Dr. K. B. Soni
(Member, Advisory Committee)
Professor and Head
Department of Plant Biotechnology
College of Agriculture, Vellayani
Thiruvananthapuram- 695 552



Dr. Makesh Kumar T.
Principal Scientist (Plant Pathology)
Division of Crop Protection
ICAR- CTCRI
Sreekariyam,
Thiruvananthapuram- 695 017



Dr. Swapna Alex
(Member, Advisory Committee)
Professor and Course Director
B. Sc. - M. Sc. (Integrated)
Biotechnology Course
Department of Plant Biotechnology
College of Agriculture, Vellayani
Thiruvananthapuram- 695 552



Dr. M. K. Rajesh
(External Examiner)
Principal Scientist (Biotechnology)
Division of Crop improvement
ICAR- CPCRI
Kasaragod - 671124

ACKNOWLEDGEMENT

In the name of God, the Almighty, for his showers and blessings throughout my research work to complete my research successfully.

With boundless love and appreciation, I would like to extend my heartfelt gratitude and appreciation to the people who helped me in bringing this study into reality. It is with my heartfelt feelings I express my deepest gratitude to my beloved advisor Dr. J. Sreekumar, Principal Scientist (Section of Extension and Social Sciences, ICAR-CTCRI, Sreekariyam) for his guidance, patience, and encouragement over the last year. I am highly indebted for his valuable guidance and patience.

I express my deepest gratitude to the advisory committee members, Dr. K. B. Soni (Professor and Head, Dept. of Plant Biotechnology, COA), Dr. Swapna Alex (Professor and Course Director, B. Sc. – M. Sc. (Integrated) Biotechnology, Dept. of Plant Biotechnology, COA) and Dr. Makesh Kumar T. (Principal Scientist, Division of Crop Protection, ICAR-CTCRI).

I take immense pleasure to express my deep sense of gratitude to my college advisor Dr. Deepa S Nair (Asst. Professor, Dept. of Plantation, Crops and Spices, COA) and Dr. M. K. Rajesh (Principal Scientist, ICAR-CPCRI, Kasaragod).

I express my profound gratitude to Dr. A. Anilkumar (Dean, COA), Dr. Archana Mukherjee (Director, ICAR-CTCRI) for providing me all the facilities during the course of my work. I would like to express my sincere thanks and gratitude to Dr. M. N. Sheela, (Head Section of Extension and Social Sciences, ICAR-CTCRI) for permitting me and extending all the facilities to complete my work.

I would like to thank Jayakrishnan chettan, Merlin chechi, Divya chechi and other members of virology lab for helping me while doing the validation part of the work. I would like to put on record my sincere thanks to Sumayya chechi for her support and help during the final stages of research work.

I am pleased to place my etiquette to Dr. C Mohan (Principal scientist, Division of Crop Improvement, ICAR-CTCRI), Dr. Senthil Kumar K. M. (Scientist, Division of Crop

Improvement, ICAR-CTCRI), Prakash chettan, Ammu chechi and Tom chettan for helping me during the work.

This thesis becomes a reality with the support of my friends Jo, Reshu, Pattu, Paru, Hasmi, Adi, Vishnu, Alif, Amal and Rahul. I also express my thanks and appreciation to my beloved classmates who have willingly helped me out with their abilities

My special and wholehearted thanks to Athulettan for his exemplary support, monitoring and guidance during credit seminar and research work. I thank Jithu for supporting me during the work through enjoyable discussions.

I am deeply indebted to Ambu chettan, Sreenath chettan and all the scientists and staff members of ICAR- CTCRI, teachers in college, my seniors and juniors for their timely support. I acknowledge the favour of numerous persons who, though not been individually mentioned here, who have all directly or indirectly contributed to this work.

Last but not least, I can't forget the support, prayer and encouragement of my parents and my sister which inspires me all the way throughout my studies. I thank them for all the support and strength they gave me throughout my life.

Rajani K. R.

*DEDICATED TO MY
PARENTS*

TABLE OF CONTENTS

Sl. No.	Title	Page No.
	LIST OF TABLES	ii
	LIST OF FIGURES	iii
	LIST OF PLATES	v
	LIST OF APPENDICES	vi
	LIST OF ABBREVIATIONS	vii
1	INTRODUCTION	1-3
2	REVIEW OF LITERATURE	4-18
3	MATERIALS AND METHODS	19-35
4	RESULTS	36-72
5	DISCUSSION	73-76
6	SUMMARY	77-78
7	REFERENCES	79-91
8	APPENDICES	92-101
9	ABSTRACT	102

LIST OF TABLES

Table No.	Title	Page No.
1	RT-PCR reaction profile	35
2	Protein-protein interaction (PPI) in plant templates and cassava	37
3	Virus species interacting with <i>Arabidopsis thaliana</i>	40
4	Proteins in cassava predicted to interact with CMV	43
5	Predicted genes in CMV interacting with cassava	52
6	Identified subcellular locations of the predicted protein using Localizer.	60
7	Disease resistance protein and its corresponding genes in cassava	67
8	GO of predicted interacting proteins in Cassava mosaic virus	68
9	Quantification of RNA	70

LIST OF FIGURES

Figure No.	Title	Page No.
1	Production of cassava in world (FAOSTAT, 2019)	5
2	Production of cassava in India (FAOSTAT, 2019)	5
3	Work flow for the construction of cassava PPIN	23
4	Homologous PPI derived from interactions between homologs	24
5	Work flow for the prediction of Cassava-CMV PPI	30
6	Cassava PPI network derived by interolog-based method	38
7	A model of HPIDB Blast result	41
8	Predicted PPIN of Cassava-CMV	48
9	Predicted PPIN of Cassava-CMV	49
10	Predicted PPIN of Cassava-CMV	50
11	Merged Cassava-CMV PPIN	51
12	Blast2GO pipeline	53
13	Analysis progress of predicted cassava proteins	54

Figure No.	Title	Page No.
14	InterProScan families distribution of predicted cassava proteins	56
15	InterProScan domain distribution	57
16	Cellular component of the predicted proteins in Cassava	58
17	Relative gene expression of <i>AC2</i> and <i>CAT2</i> in healthy and infected cassava leaves	72

LIST OF PLATES

Plate No.	Title	Page No.
1	1.2% Et Br stained agarose gel showing RNA of two cassava leaf samples after electrophoresis	70

LIST OF APPENDICES

Sl. No.	Title	Page No.
1	Functional annotation result of the predicted PPIs in Cassava	92-101

LIST OF ABBREVIATIONS

APID	Agile Protein Interactome Dataserver
AtPIN	<i>Arabidopsis thaliana</i> Protein Interaction Network
BLAST	Basic Local Alignment Search Tool
BLASTp	Protein BLAST
BTV	<i>Brevipalpus-transmitted viruses</i>
cDNA	complementary DNA
CMD	Cassava Mosaic Disease
CMGs	Cassava Mosaic Geminiviruses
CTAB	Cetyl Trimethyl Ammonium Bromide
DEPC	Diethyl pyrocarbonate
FAO	Food and Agriculture Organization
FAOSTAT	Food and Agriculture Organisation Corporate Statistical Database
GCENs	Gene Co-expression Networks
GO	Gene Ontology
GOA	Gene Ontology Annotation
HPI	Host Pathogen Interaction
HPIDB	Host Pathogen Interaction Database
ICTV	International Committee on Taxonomy of Viruses
IntAct	Interaction database
MAMPs	Microbe Associated Molecular Patterns
MAPK	Microbe Associated Protein Kinases
MINT	Molecular Interaction Database
MS	Mass Spectroscopy
OD	Optical Density
PAIR	Predicted Rice Interactome Network
PCR	Polymerase Chain Reaction
PHI	Pathogen-Host Interaction

PPIs	Protein-Protein Interactions
PPIN	Protein-Protein Interaction Network
PR	Pathogenesis Related
PRIN	Predicted Rice Interactome Network
PRRs	Pathogen Recognition Receptors
PVI	Plant-Virus Interaction
q-PCR	Quantitative PCR
RT-PCR	Real Time-Polymerase Chain Reaction
SVM	Support Vector Machine
TAP	Tandem Affinity Purification
Y2H	Yeast 2 Hybrid

INTRODUCTION

1. INTRODUCTION

Cassava (*Manihot esculenta* Crantz) is a perennial shrub that belongs to Euphorbiaceae family. It is a native of South America and is believed to have been introduced by Portuguese traders in sub-Saharan Africa during the 16th century. Cassava is the third most important source of calories in the tropics after rice and maize (Food Safety Network, 2014).

According to FAO (Food and Agriculture Organization of the United Nations) Food Outlook Annual Report (2018), cassava plays a leading role in food security in India, especially in the major growing states of Kerala and Tamil Nadu. Jointly, both the states account for 98% of national output. Cassava production output is marginally down from 2017, with total production of about 4.1 million tonnes which is very less than half the record production of the crop that was harvested in 2014.

Cassava is vulnerable to a wide range of diseases caused by viruses. The virus is either seed transmitted or vector transmitted by whitefly (Macfadyen *et al.*, 2018). Among them, Cassava Mosaic Disease (CMD) is the most severe and widespread, thereby limiting production of the crop in cassava growing areas. CMD produces a variety of foliar symptoms such as mosaic, mottling, misshapen and twisted leaflets, and an overall reduction in size of leaves and plants. In India, CMD is caused by *Indian cassava mosaic virus* (ICMV) and *Sri Lankan cassava mosaic virus* (SLCMV). It has obtained considerable attention in the southern states of Kerala and Tamil Nadu, which are the main cassava growing areas of India.

Different interactions are generated between the plant (host) and the virus (pathogen) during each stage of the viral cycle. Host-pathogen interaction alters the host physiology. Hence, studies were undertaken to evaluate changes in physiology of healthy cassava plants as well as cassava mosaic virus infected cassava plants.

The pathogen–host interactions (PHIs) may be between proteins, nucleotide sequences, metabolites, and small ligands. The protein–protein interactions (PPIs) have been identified as the most relevant type in the functioning of PHI systems and therefore are the most studied type (Stebbins, 2005; Korkin *et al.*, 2011; Zoraghi and Reiner, 2013). A number of experimental methods have been applied to discover PPIs. Some traditional methods of determining PPIs are Yeast two Hybrid (Y2H) method, Tandem Affinity Purification (TAP) tagging, and Mass Spectroscopy (MS). The labour intensive experimental techniques for the detection of PPIs may not be generally applicable due to time constraints and high cost of experiments; therefore recently, computational techniques are preferred for the prediction of PPIs.

In 2002, Kitano mentioned that systems biology is an integrative research area in life science that mainly focuses on the study of non-linear interactions between biology entities through the integration and combination of bimolecular and medical sciences with mathematical, computational, and engineering disciplines. The different levels of omics data collected from pathogens and infected cells are critical components that drive bioinformatics analysis. This promotes the construction and analysis of infection specific gene-regulatory, metabolic, and protein–protein interaction networks (Westermann *et al.*, 2012 and Schulze *et al.*, 2015).

With an increasing amount of experimental PHI data, web-based databases were developed to derive and provide pathogen–host interactome data that mainly focuses on specific pathogens or hosts (Wattam *et al.*, 2013; Ako-Adjei *et al.*, 2014; Calderone *et al.*, 2014; Guirimand *et al.*, 2014).

Although the available databases are promising in data archiving, a large amount of PHI data is not stored in any of these databases, since these data are buried within the literature. Therefore, there is an urgent need for novel text mining methods specific for PHI data retrieval.

The current study focuses on the generation of Protein-Protein Interaction Network (PPIN) of cassava-Cassava Mosaic Virus (CMV). The objectives of the study includes data mining of plant-virus interaction through PPI networks, computational prediction of PPIs, construction of PPIN of predicted PPIs, analysis of predicted interactome and validation of predicted proteins.

*REVIEW OF
LITERATURE*

2. REVIEW OF LITERATURE

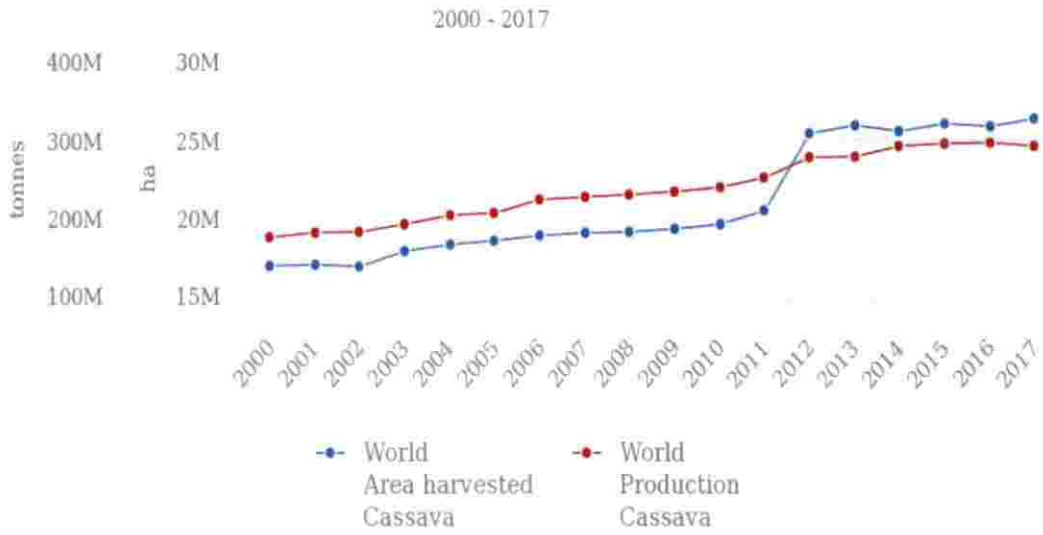
2.1 CASSAVA

Cassava (*Manihot esculenta* Crantz) is grown throughout tropical Africa, Asia and the America. Its large starchy roots and edible leaves provide food for 800 million people globally, many of whom partly relies on it because it is drought tolerant and requires little in the way of inputs. Due to the high starch content (20-40%) cassava is a desirable energy source both for human consumption and industrial biofuel applications (Ceballos *et al.*, 2010).

Sub-Saharan Africa (SSA) is the world's largest cassava growing region. According to FAOSTAT (2017), cassava production in SSA could reach a record of 161 million tonnes in 2018 that means around 3.3 million tonnes or 2% more than the level of 2017. In India, the cultivation of cassava is mainly done in Kerala, Tamil Nadu, Andhra Pradesh, Nagaland, Meghalaya and Assam. In Tamil Nadu and Andhra Pradesh, it is grown under open conditions whereas in Kerala, about 40% of cassava is raised as a mixed crop. The toughness of cassava enables it to grow profitably under a wide range of agro-ecological zones where cereals and other crops cannot thrive, making it a suitable crop for poor farmers to cultivate under marginal environments in Africa. The other interest for farmers to grow cassava is that it produces higher yields per unit of land than other crops such as yam, wheat, rice, and maize (Alabi *et al.*, 2011).

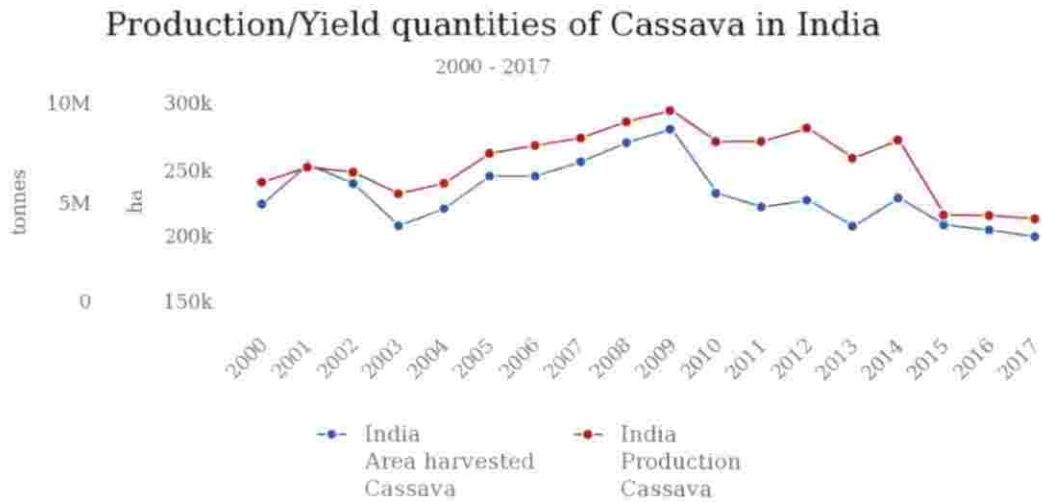
According to FAO classification, root and tuber crops form staple diet for 3% of the global population. Figure 1 & 2 represents FAOSTAT of cassava production in world and in India respectively. In the African continent and South America, cassava is mostly used for human consumption.

Production/Yield quantities of Cassava in World + (Total)



Source: FAOSTAT (Aug 27, 2019)

Figure 1. Production of cassava in world (FAOSTAT, 2019)



Source: FAOSTAT (Aug 27, 2019)

Figure 2. Production of cassava in India (FAOSTAT, 2019)

The roots of cassava are the major source of dietary starch. The tubers are eaten fresh and in various forms of processed food. Cassava leaves are also consumed as a green vegetable, especially in East Africa, to provide an important source of proteins, minerals, and vitamins. With increased possibility of starch from cassava as a source of ethanol for biofuels, its cultivation is transforming from subsistence to a more commercially-oriented farming enterprise (Nassar and Ortiz, 2010). Cassava is cultivated in about 13 states of India, and its major production is from the southern states of Kerala and Tamil Nadu.

2.2 CASSAVA MOSAIC DISEASE (CMD)

CMD is mainly caused by *Cassava Mosaic Virus* (CMV). They are members of the family *Geminiviridae* and the Genus *Begomovirus*. CMD produces different foliar symptoms like mosaic, mottling, misshapen and twisted leaflets. CMD-affected cassava plants produce few or no tubers, depending on the intensity of the disease and the age of the plant at the time of infection.

Nine distinct cassava mosaic viruses have been characterized worldwide from CMD-affected cassava plants and seven of them are from sub-Saharan Africa. Two other viruses, *Indian cassava mosaic virus* (ICMV) and *Sri Lankan cassava mosaic virus* (SLCMV), were reported from the Indian sub-continent. Currently, the International Committee on Taxonomy of Viruses (ICTV) has placed all of these viruses in the genus *Begomovirus*, the largest genus in the family *Geminiviridae*, and collectively, they are also called the Cassava Mosaic Begomoviruses (CMBs) or Cassava Mosaic Geminiviruses (CMGs) (Alabi *et al.*, 2011).

Both ICMV and SLCMV possess bipartite ssDNA genomes and are transmitted by whiteflies. Cassava is the primary host plant of ICMV and SLCMV but both viruses can experimentally infect *Nicotiana spp.* In addition, an infectious clone of SLCMV was infective in *Arabidopsis thaliana* inducing symptoms similar to those described on cassava including stunting, leaf deformation and developmental abnormalities (Mittal *et al.*, 2008).

CMBs (Cassava Mosaic Begomoviruses) induces several morphological and cytological modifications in cassava and the experimental host *Nicotiana benthamiana* (Atiri *et al.*, 2004). *Bemisia tabaci*, the whitefly vector, is mainly responsible for the secondary spread of CMBs, although other species of whitefly, such as *B. afer* can also transmit cassava mosaic disease (Dubern, 1994).

CMV has two circular DNA molecules, designated DNA-A and DNA-B, of approximately 2.8 kb, both of which are required for systemic infection of plants. DNA-A encodes six genes whereas DNA-B encodes two genes. DNA-A viral strand encodes for the coat protein (CP) (AV1 ORF), and AV2 which functions as a suppressor of host RNA silencing, thereby modulating symptoms, or may also be involved in host specificity. The minus strand of DNA-A has four open reading frames (ORFs) that encode for the Rep associated protein (AC1), a transcriptional activator (TrAP/AC2), a replication enhancer (Ren/AC3), and the AC4 protein. The AC4 ORF is completely embedded within the coding region of the Rep protein, and it is the least conserved of all the geminiviral proteins, both in sequence and in function (Bisaro, 2006).

2.2.1 Impact of CMD on Cassava

Atiri *et al.*, 2004 reported that CMBs induce several morphological and cytological modifications in cassava and the experimental host *Nicotiana benthamiana*. The symptoms and accompanying cellular modifications depends on whether cassava is infected with a single virus, or if there is a concurrent infection of two or more CMBs resulting in synergistic interactions. The morphological alterations in cassava often result in loss of tuber and storage root yield that can occur even in resistant genotypes which shows only mild or no foliar symptoms. Overall, storage root yield loss across sub-Saharan Africa were estimated between 15-24% annually, which is equivalent to 12-23 million tons or an annual loss of US\$ 1.2-2.3 billion (Alabi *et al.*, 2011).

2.3 PATHOGENICITY IN CASSAVA

In response to pathogens, plants have developed a sophisticated mechanism of action, which depends on the ability to recognize pathogen-specific and foreign molecules for the plant, both in quantitative and qualitative resistance (Boller and He *et al.*, 2009, Vasquez *et al.*, 2018).

At the level of the plasma membrane, the recognition of pathogens depends on Pattern Recognition Receptors (PRRs) which recognize Microbe Associated Molecular Patterns (MAMPs). On the other hand, at the intracellular level, the recognition depends on the proteins encoded by the *R* (resistance) genes, which recognize effector proteins injected by the pathogens (Monaghan and Zipfel, 2012; Jones *et al.*, 2016). Once the recognition of the pathogen by the plant occurs, a series of defence responses is triggered. These defences include the strengthening of the cell wall through the synthesis of callose and lignin (Hauck *et al.*, 2003), the production of secondary antimicrobial metabolites such as phytoalexins (Almargo *et al.*, 2008) and the activation of the cascades of signalling by Mitogen-Activated Protein Kinases (MAPK) (Meng and Zhang, 2013)..

All these responses together with the induction of gene expression code for proteins related to pathogenicity (PR) (van Loon *et al.*, 2006).

2.4 PLANT-VIRUS (PLANT-PATHOGEN) INTERACTIONS

Plant viruses are obligate intracellular parasites that are infectious, which mostly consist of positive ssRNA (single-stranded ribonucleic acid) and only in a few cases by single-stranded or double-stranded deoxyribonucleic acid.

Viruses can enter the plant cell passively only through wounds caused by physical injuries due to environmental factors or by vectors. Among vectors, several species of insects, mites, nematodes and some soil inhabitant fungi can transmit specific viruses. Viruses use energy and proteins from the host cell to perform its processes.

Different interactions are generated between the plant and the virus during each stage of the viral cycle. If the viral particle is not recognized by the host plant, a compatible interaction between the plant and the virus is established. This interaction may be favourable for the virus. However, if the plant recognizes the viral particle, an incompatible interaction that is unfavourable for the virus is established. It is known that plants can recognize the virus, limiting it to the site of the infection. A series of complex cascade defence reactions can be induced thereby limiting virus replication and virus movement within the host plant (Hammond-Kosack and Jones, 2000).

Flor in 1971 described that plants have developed defence mechanism at the molecular level based on the gene for gene theory. This model is defined by the expression of a resistance gene (*R*) in the plant, which can bind directly or indirectly to the product of the avirulence gene (*avr*) of the pathogen (Bent, 1996; Ellis *et al.*, 2000b).

For over 20 years, *Arabidopsis thaliana* has been developed as a model organism for molecular plant genetics. *Arabidopsis* is widely used as a model for the study of Plant-Pathogen co-evolution (Pagan *et al.*, 2010). In 2017, Arena *et al.* evaluated *A. thaliana* as an alternative host for *Brevipalpus-transmitted viruses* (BTV). They reported that CiLV-C (*Citrus leprosis virus*) is able to infect *Arabidopsis* inducing localized chlorotic symptoms upon infestation with *Brevipalpus viruliferous* mites. Interaction between *A. thaliana* NAC domain protein ATAF2 (AT5g08790) and *Tobacco Mosaic Virus* (TMV) replicase protein is reported by Wang *et al.* (2009). Sahu *et al.* (2014) predicted the interactions between *Arabidopsis* and *Pseudomonas syringae* pathovar tomato strain DC3000 (PstDC3000) in genome scale. *Pseudomonas syringae*, a major bacterial leaf pathogen is asserted to infect the plant host *Arabidopsis thaliana* and has been accepted as a model system for experimental characterization of the molecular dynamics of plant-pathogen interactions. They predicted 868645 Protein-Protein Interactions (PPIs) between 14043 *Arabidopsis* proteins and 1337 *P. syringae* proteins. PPI prediction between *R. solanacearum* and *Arabidopsis thaliana* was

done by Li *et al.* (2011). They predicted 3,074 potential PPIs between 119 *R. solanacearum* and 1,442 *A. thaliana* proteins.

2.5 PROTEIN-PROTEIN INTERACTION

Protein-Protein Interaction (PPI) refers to physical contacts build between two or more proteins resulting from the biochemical events or electrostatic forces. Therefore, PPIs and their associated networks are essential for the understanding of cellular processes, such as enzymatic activity, immunological recognition, DNA repair, network pathway, signalling cascades and transcription control. A study of protein interaction networks is important not only from a theoretical way but also in terms of potential practical applications.

For the identification of protein interactions, many experimental methods have been developed. Some of the experimental methods allow screening of a large number of proteins in a cell. Such methods include yeast two-hybrid (Y2H), Tandem Affinity Purification (TAP), Mass Spectroscopy (MS). Other methods focus on examining and characterizing specific biochemical and physiochemical properties of a protein complex. Despite this, a complete interaction network for many organisms is not available. Due to the low interaction coverage, experimental biases toward certain protein types and cellular localizations reported by most experimental techniques, there is a need for the development of computational methods to predict whether two proteins interact.

Recently, a number of compatible computational approaches have been developed for the large-scale prediction of protein-protein interactions based on protein sequence, structure and evolutionary relationships in complete genomes (Shoemaker and Panchenko A.R., 2007).

2.6 COMPUTATIONAL APPROACHES FOR PREDICTING PPI (PROTEIN-PROTEIN INTERACTION)

Computational methods provide equivalent approach for detecting protein-protein interactions. Indeed, the broad availability of experimental data has declined the development of numerous computational methods over the past few years.

In general, all computational approaches to PPI prediction attempt to leverage knowledge of experimentally determined previously known interactions in order to predict new PPIs. These methods enable one to discover novel putative interactions and often provide information required for designing new experiments for specific protein sets (Pitre *et al.*, 2008). Methods specific for intra-species interactions are usually used in PPI prediction studies (Nourani *et al.*, 2015). On the other hand, concentrating on the interactions among different organisms is a young branch of this field.

2.6.1 Machine learning and data mining based approach

Machine learning techniques (supervised and semi supervised) have been applied intensively for interspecies PPI predictions. However, these methods require template PPI data sets associated with appropriate biological and biochemical properties as features for training and testing purposes.

Baldi and Brunak, (2001) applied machine learning techniques to bioinformatics and is a well-accepted idea, which includes early efforts for PPI predictions. These methods utilize accessible PPI data as features for training and classifying interacting and non-interacting protein pairs. Support Vector Machine (SVM) based approaches are successfully applied in PHI prediction studies (Kshirsagar *et al.*, 2013a; Mei, 2013). Cui *et al.* (2012) presents a SVM based approach, which uses a fixed length feature vector, indicating relative frequency of consecutive amino acids in the protein sequence.

Machine learning based methods which compose PPI prediction as a classification task use both interacting and non-interacting protein pairs as positive and negative classes, respectively.

2.6.2 Homology based approaches

The logic behind this type of methods is the assumption of conserved interactions between a pair of proteins which have interacting homologs in another species. The conserved interaction is called as “Interolog”. The simple method of identifying interologs is as follows:

For example, consider a template PPI pair (a, b) in a source species, find the homolog ‘a’ in the host and the homolog ‘b’ in the pathogen, conclude that (a, b) interact. Simplicity and clear biological basis are the main advantages of these methods. However, homology to known interactions is not sufficient for assessing the biological evidence of the predicted results. Different filtering techniques should be considered for evaluating the feasibility of the interactions under an in vivo condition and hence decreasing the false positives. A homology detection method using template PPI databases, DIP (Salwinski *et al.*, 2004) and iPfam (Finn *et al.*, 2013), is published in Krishnadev and Srinivasan (2008) for the prediction of PHI pairs. Searching the sequences of host and pathogen proteins within two template databases are conducted to find a superset of all interactions which are physically and structurally compatible. These potential interactions are refined within two additional filtering steps, for the detection of biologically feasible interactions including integration of expression and sub-cellular localization data (Tyagi *et al.*, 2009).

In 2011 Krishnadev and Srinivasan have applied the same procedure for different pathogens in their subsequent works. Another study was done with the same approach by using sequence similarity enhanced with domain-domain interaction detection (Schleker *et al.*, 2012a). They have two compressive reviews of the computational approaches predicting *Salmonella*-Host interactions

(Schleker *et al.*, 2012b, 2015), which include comparing *Salmonella*-Human and *Salmonella*-Plant interaction predictions.

Homolog knowledge can be used indirectly as a remedy for data scarcity and data unavailability by homolog knowledge transfer. Homolog information (features) can be used when the information of a protein is unavailable. Mei (2013) has designed different experiments to show the performance of substituting homology features. Pessimistic experiment, which uses only homology features for train and test without incorporating any base proteins (target), has promising results, indicating that using homolog information is an effective substitute for the target information to tackle the problem of data unavailability.

Lee *et al.* (2008) uses high confidence intra-species PPIs to detect interologs using ortholog information. The hypothesis is that when two orthologous groups are shared between more than two species, there will be a possible interolog between those orthologous groups. The possible interactions are filtered using gene ontology annotations followed by pathogen sequence filtering based on the presence or absence of translocational signals to clarify the predictions. The notable point is slight intersection of the predicted interactions with those of the reported predictions in Dyer *et al.* (2007) due to applying different techniques and datasets for same pathogen-host system. Zhou *et al.* (2014) introduces the “stringent homology” which does not rely only on intra-species template PPIs to discover interologs and make use of two different organisms as the source of template PPIs to predict PHIs. They also claim that it is not only for the targeted host proteins which tend to be hub in their own PPI network and this is also true about targeting pathogen proteins. The most important obstacle for using homology based methods is scarcity of available homolog information.

2.6.3 Structure based approaches

A number of studies are based on structural similarities and use template PPIs to detect similar interacting pairs within host and pathogen proteins. Primary

ideas presented in Davis *et al.* (2007) called comparative modelling and was based on their prior work (Davis *et al.*, 2006). Their method starts with a set of host and pathogen proteins and then sequence matching procedures are used to decide the similarities between the host or pathogen proteins with known structure or known interaction protein partners. Sequence similarity score is only used when structure information is unavailable as a statistical potential evaluation, to predict interacting partners. The main disadvantage of this method is that finding high similarity between pathogen proteins and proteins with known structure is not guaranteed for all pathogen proteins.

Therefore, lack of the spatial structural information would restrict the applicability of this method. Furthermore, they have only the ability to collect limited number of standard PPIs from literature to evaluate their prediction performance.

2.6.4 Domain and motif based approach

Wojcik and Schächter, (2001) and Pagel *et al.* (2004) introduced the idea of utilizing domains as building blocks of proteins for predicting PPIs is well-studied for single organisms concerning the fact that domains are the mediators of interactions. The approach presented in Dyer *et al.* (2007) is one of the pioneer published researches for predicting PHIs. However, small list of interaction is presented and their biological importance is not strongly evaluated.

To predict interactions between host and pathogen proteins, they present an algorithm that links protein domain profiles with interactions between proteins from the same organism. For every pair of functional domains (d, e) which is present in protein pair (g, h) respectively, the probability of interacting (g, h) is assessed using Bayesian statistics. To apply this idea to a pathogen-host system, they identify domains in every host and pathogen proteins and determine the interaction probability for each pair of host and pathogen proteins that contain at least one domain. A similar knowledge source is chosen in Kim *et al.* (2007) which make use of domain information from InterProScan (Quevillon *et al.*,

2005). They predict PPIs using PreID (Kim *et al.*, 2002) and PreSPI (Han *et al.*, 2004) algorithms based on domain information.

2.7 COMPUTATIONAL METHOD FOR INTER-SPECIES PPI PREDICTION

Many computational methods have been developed to predict PPIs, but most of them are intended for PPIs within a species rather than PPIs across different species such as PPIs between virus and host. Methods for predicting intra-species PPIs do not distinguish interactions between proteins of the same species from those of different species, and thus are not appropriate for predicting inter-species PPIs. The knowledge of host pathogen PPIs is crucial for understanding the pathogenesis of the relevant disease. However, experimental resources for studying interactions between host and pathogen proteins are scarce. Several computational methods for predicting interspecies PPIs have been developed, including methods based on interolog, interacting domain/motif, structure, and even machine learning (Zhou *et al.*, 2012).

2.7.1 Interolog Based Approach

Interolog based methods composed of the conventional way of predicting host-pathogen interactions. The methods are based on the hypothesis that pairs of interacting proteins in one species are expected to be conserved in related species. The idea behind this approach is that if two proteins interact in one organism, their interolog in another organism have a higher chance of interacting. This is based on the hypothesis that sequence and structural similarities between gene products suggest functional similarities. Sahu *et al.* (2014) predicted the interactions between *Arabidopsis* and *Pseudomonas syringae* pathovar tomato strain DC3000 (PstDC3000) using interolog-based method and domain based method. The interolog-based method predicted ~0.79M PPIs involving around 7700 *Arabidopsis* and 1068 *Pseudomonas* proteins in the full genome.

2.8 PROTEIN-PROTEIN INTERACTION NETWORK

A protein-protein interaction network (PPIN) is a collection of PPIs, deposited in online databases. PPINs may contribute other datasets, such as protein structural information, which may lead to understanding the different subparts that contribute to the function of a whole biological system.

A major issue in using PPINs in practice involves handling with errors in the form of missing interactions and false signals. In a PPI network, proteins are represented as nodes. Some nodes interact with many more partners than average; these proteins are called hubs (Albert, 2005). Barabasi & Oltvai (2004) reported that loss of hubs may cause the breakdown of the PPIN into isolated clusters.

Protein-Protein Interactions (PPIs) are of interest in biology because they regulate roughly all cellular processes, including metabolic cycles, DNA transcription and replication, different signalling cascades and many additional processes. Proteins carry out their cellular functions through cooperative interactions with other proteins, so it is important to know the specific nature of these relationships. Indeed, the importance of understanding these interactions has prompted the development of various experimental methods used in measuring them. While the amount of genomic sequence information continues to increase exponentially, the annotation of protein sequences appears to be somewhat lagging behind, both in terms of quality and quantity.

Multi-branched, high-throughput functional genomics approaches are needed to bridge the gap between raw sequence information and the appropriate biochemical and medical information. Therefore, computational methods are required for discovering interactions that are not accessible to high throughput methods. These computational predictions can then be checked by using more labour-intensive methods. A number of computational approaches for protein interaction discovery have been developed over recent years. These methods differ in feature information used for protein interaction prediction. Many studies

have proved that knowing the tools and being familiar with the databases is important for new research in protein-protein interaction.

2.9 THE STRUCTURE OF PROTEIN NETWORKS

The structure of protein interaction networks has been examined by recent studies in several species. These studies have discovered that regardless of species, the known protein networks are scale-free. It means that some hub proteins have a huge proportion of the interactions while most proteins (are not hub and) only contain a small fraction of ones. It is an obvious fact that understanding the structure of a species' protein interaction network only provides one dimension of the biochemical machinery controlling a cell's behaviour. Thus, several groups have integrated dynamics of gene expression with protein interaction networks in order to uncover how these networks change in different biological states.

Network topology is also introduced to characterize a network structure. There are four higher-level topological indices including average degree (K), clustering coefficient (C), average path length (L), and diameter (D). It is possible to calculate four topological distributions such as degree distribution $P(k)$, degree distribution of cluster coefficients $C(k)$, shortest path distribution $SP(i)$, and topological coefficient distribution $TC(k)$, which take more attentions and are comprehensively used in cellular networks, such as PPI networks, MNs (Metabolic Networks), gene co-expression networks (GCEN), and domain interaction networks. The topological features of cellular networks are efficiently explained by these criteria which also provide vast insights into cellular evolution, molecular function, network stability, and dynamic responses.

2.10 FUTURE DIRECTION AND CONCERNS: EVOLUTION OF PROTEIN-PROTEIN INTERACTION NETWORKS

Protein-protein interaction network is highly dynamic and studying the evolution of protein-protein interaction networks is one of the central problems of systems biology, the results of such researches are crucial for a better

understanding of the evolution of living systems and could be used for protein interaction and function prediction.

2.11 PERFORMANCE EVALUATION

The lack of gold standard PHI data and the complexity of PHI mechanisms lead to a hard assessment phase, in a way that predicted interactions are rarely supported by a biological basis. Some studies validate their results by measuring the shared interactions with other published materials (Mukhopadhyay *et al.*, 2012, 2014; Segura-Cabrera *et al.*, 2013).

2.12 VALIDATION USING q-PCR

Genes specific for interacting pairs of proteins possessing specific functions are selected for validation. Validation of identified genes is essential for further analysis. Different methodologies are available like qPCR, micro-array analysis for detection and quantitation. Due to high sensitivity and efficiency of qPCR, it is widely adopted for expression analysis.

Real-time PCR (RT-PCR) is also known as quantitative PCR or qPCR. In qPCR amplification, cDNA is detected in real time as PCR is in progress by the use of fluorescent reporter for RNA expression studies. Fluorescent probes mostly used are sequence-specific TaqMan probe and generic non-sequence-specific double-stranded DNA binding dye such as SYBR green. The principle behind this technique is that the intensity of fluorescence emitted by the probe at each cycle is directly proportional to the template quantity.

*MATERIALS AND
METHODS*

3. MATERIALS AND METHODS

The study entitled “Modeling of Cassava-Cassava Mosaic Virus interactions with computational biology and bioinformatics approach” was carried out at the Section of Extension and Social Sciences, ICAR-Central Tuber Crops Research Institute, Sreehariyam, Thiruvananthapuram during 2018-2019. In this chapter, details regarding experimental materials and methodology used in the study are elaborated.

3.1 DATA SOURCES, DATABASES AND VISUALIZATION TOOLS USED IN PPI PREDICTION STUDY

The protein sequence data were obtained from a partially inbred line-AM560-2. The whole genome assembly (approx.221.2 MB arranged on 18 chromosomes) and whole genome annotation (33,033 genes) of AM560-2 genotype of *Manihot esculenta* v6.1 (cassava) were downloaded from Phytozome, the Plant Comparative Genomics Portal of the Department of Energy’s Joint Genome Institute. (www.phytozome.jgi.doe.gov) (Bredeson *et al.*, 2016). *Cassava Mosaic Virus* (CMV) proteome were downloaded from UniProt database (www.uniprot.org).

3.1.1 STRING

The STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) is specific to functional associations (stable physical associations, transient binding, substrate chaining, and information relay) between proteins, on a global scale (Szklarczyk *et al.*, 2014). Singh and Singh (2019) constructed interologous PPI network of Tea (*Camellia sinensis*) leaf from RNA-Seq datasets using STRING database. In this, a total of 11,208 nodes with 1,97,820 interactions were successfully predicted using this interolog based approach. The Database URL: (<http://string-db.org>).

3.1.1.1 *STRING viruses*

'STRING viruses' is an expanded form of STRING, to include intra-virus and virus–host PPIs. The STRING viruses database provides a single unified interface to virus–virus and host–virus PPIs from text mining and many experimental sources. Furthermore, the data can also be directly imported into Cytoscape (Shannon *et al.*, 2003) using the STRING Cytoscape app (Szaklarczyk *et al.*, 2016).

3.1.2 APID

APID (Agile Protein Interactomes Data Server) is a bioinformatics web server developed to provide protein interactomes at different quality levels and allowing their analysis and visualization as networks. APID contains binary interactions for 807 organisms, including 19 species with at least 500 reported binary interactions (Alonso-López *et al.*, 2019). Database URL: <http://apid.dep.usal.es>.

3.1.3 HPIDB

HPIDB 3.0 generates a comprehensive set of Host Pathogen Interaction (HPI) by (i) in-house manual curation of published, experimental HPI data and (ii) bringing in external HPI data provided by previously known molecular interaction resources (Ammari *et al.*, 2016). Sahu *et al.* (2014) used HPIDB for the prediction of *Arabidopsis-Pseudomonas syringae* interactome. In this, each protein in *Arabidopsis* and *Pseudomonas* is BLASTed against all the protein sequences in HPIDB database to identify the homologs with E-value, sequence identity and aligned sequence length coverage.

3.1.4 Prediction tool: VirusHostPPI

Amino acid sequence similarity between different types of viruses or hosts is relatively low, therefore sequence-based prediction of virus-host PPIs for new viruses or hosts is quite challenging. Zhou *et al.* (2018) developed a new

prediction method of virus-host PPIs which is applicable to new viruses or hosts. The prediction tool is based on SVM (Support Vector Machine) method.

3.1.5 Cytoscape

Cytoscape is a free software package, which is one of the most popular protein-protein interaction visualization and data integration tools. Cytoscape is a general purpose modelling environment for integrating biomolecular interaction networks and states. Cytoscape is available at (<http://www.cytoscape.org/>). Cytoscape is a Java application verified to run on Windows, Mac OS X and Linux. Steps for installation:

- Install Java 8

Cytoscape version 3.2 and later requires Java 8. A 64 bit Java Runtime Environment is necessary (JRE). Using a 64 bit java allows the largest network to be loaded and enables the fastest network processing.

- Download Cytoscape v.3.7.1 from <http://cytoscape.org>
- Install Cytoscape

(Automatic installation packages exist for windows, Mac OS X, and Linux platforms. Cytoscape can be installed from a compressed archive distribution and also it can be built from the source code).

- Unpack it
- Launch the application:

Cytoscape supports the import of networks from delimited text files and excel workbooks. It also allows importing of networks from public databases. Cytoscape can read network/pathway files written in Simple Interaction File (SIF or .sif format), Nested Network Format (NNF or .nnf format), Graph Markup Language (GML or .gml format), GMLL (extensible graph markup and modelling language), SBML, BioPAX, PSI-MI Level 1 and 2.5, Cytoscape.js

JSON, Cytoscape CX, GraphML, Delimited text and Excel Workbook (.xls, .xlsx) format.

3.1.6 Blast2GO

Blast2GO (Conesa *et al.*, 2005) is a comprehensive bioinformatics tool for the functional annotation and analysis of genome scale sequence datasets (Götz *et al.*, 2008). A typical basic use case of Blast2GO consists of 5 steps: BLASTing, mapping, annotation, statistics analysis and visualization.

3.1.7 QuickGO

QuickGO was developed by the GOA (Gene Ontology Annotation) group in August 2001 as a fast, web-based browser for GO term information (Huntley *et al.*, 2009). All GO annotations were assigned to UniProt Knowledgebase (UniProtKB) accessions. Using QuickGO, it is very easy to start browsing the GO and its associated annotations. Database URL: <http://www.ebi.ac.uk/QuickGO>.

3.2 COLLECTION OF DATA FROM LITERATURES FOR CASSAVA PROTEIN-PROTEIN INTERACTION

Work flow for the prediction of cassava PPIN is depicted in Figure 3. The procedure is based on the logic underlying interolog based method (shown in Figure 4), which implies two proteins (A and B) are predicted to interact if their relative homologs (A' and B') interacts. The interolog method is inspired by the hypothesis that the function of protein is retained and passed through their orthologs in evolution-related organisms.

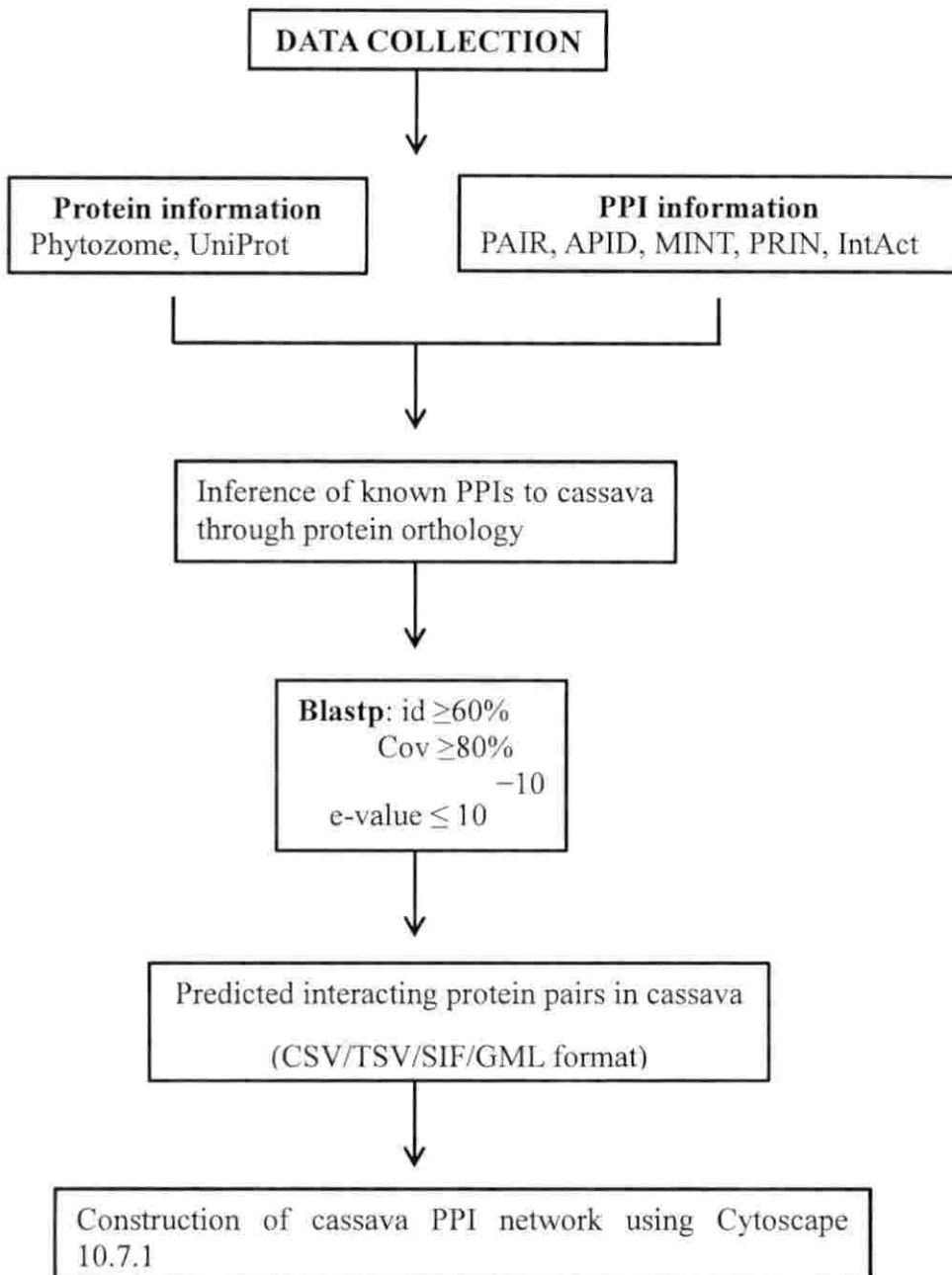


Figure 3. Work flow for the construction of cassava PPIN

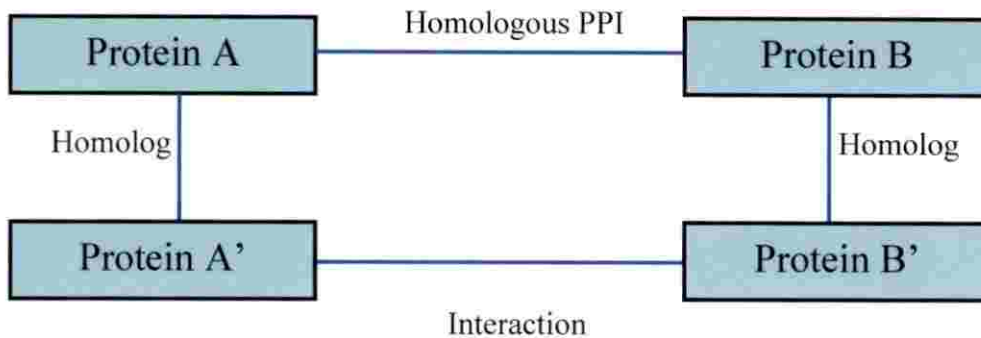


Figure 4. Homologous PPI derived from interactions between homologs: Protein A' and B' are the proteins which have direct interactions, while Protein A and B are their homologs, respectively. The interaction between A and B is called homologous protein-protein interaction (Thanasomboon *et al.*, 2017)

The whole proteome of *Indian Cassava Mosaic Virus* (ICMV) is downloaded from UniProt database (<http://www.uniprot.org/>) which contains 53 protein sequences. Similarly, the entire proteome of *Manihot esculenta* containing 34,468 protein sequences is extracted from the Phytozome v12 database (<http://www.phytozome.jgido.gov>).

3.2.1 Collection of template interaction data

The interolog method is generally based on the evidence of PPI information known to exist in other organisms. In this study, template plant species, whose PPI information was known, were selected based on one of these criteria:

- (1) Having a close evolution with cassava: *Ricinus communis* (castor bean), *Populus trichocarpa* (poplar) and *Glycine max* (soybean).
- (2) Being recognized as a starch-storing plant: *Solanum tuberosum* (potato), *Zea mays* (maize) and *Oryza sativa* (rice).
- (3) Having abundant PPI information: *Arabidopsis thaliana*.

The protein information of these template plants were obtained from Phytozome v9 and UniProt databases, and the protein interaction information was collected from five databases: IntAct, MINT, AtPIN, PAIR, and PRIN.

3.2.2 Inference of known PPIs to cassava through protein orthologous

To find protein orthologs in cassava, BLASTp search is performed against the cassava genome sequence. The cassava orthologous proteins were identified if the identity percentage ≥ 60 , coverage percentage $\geq 80\%$ and e-value $\leq 10^{-10}$.

3.2.3 Construction of cassava PPI network

Complete cassava PPI network of protein orthologs in cassava is constructed using Cytoscape v3.7.1. Cytoscape can generate publication quality images from network views. The network view can be exported in the JPG, PNG, PS (Post Script), SVG and PDF format.

3.3 DATA MINING OF PLANT-VIRUS INTERACTIONS FOR THE PREDICTION OF CASSAVA-CMV PPI

Template plant used in the study is *Arabidopsis thaliana*. *A. thaliana* is having abundant PPI information. The main databases containing *Arabidopsis* datasets are AtPIN, AtPID, PAIR.

A. thaliana is infected by a vast variety of viruses. Viruses that infect *Arabidopsis* are selected on the basis that the infecting virus is closely related to *Cassava Mosaic Virus* i.e., with reference to ICTV (International Committee on Taxonomy of Viruses). The viruses selected for the study are: *Cauliflower mosaic virus* (strain Strasbourg), *Cucumber mosaic virus* (strain FNY), *Cabbage leaf Curl virus*, *Tobacco mosaic virus* and *Tomato golden mosaic virus*, *Bean golden yellow mosaic virus*. The PPI between *Arabidopsis* and some of the viruses are obtained

from APID. The interactome data is manually searched for the corresponding PPI pair (HPI).

In this study, the probability of interaction between cassava and *Cassava Mosaic Virus* (CMV) protein is inferred from interolog-based approach. To infer the prediction from the interolog, three types of datasets are used in the study: STRINGviruses consortium 2018 dataset, HPIDB and APID dataset. The prediction framework is shown in Figure 4.

3.3.1 Interaction data of template Plant-Virus PPIN from Viruses.STRING

Viruses.STRING consortium (2018) is a protein–protein interaction database specifically catering to virus–virus and virus–host interactions. This database combines evidence from experimental and text-mining channels to provide combined probabilities for interactions between viral and host proteins. As of Jan 2019, the database contains 177,425 interactions between 239 viruses and 319 hosts. The database is publicly available at (viruses.string-db.org). The Viruses.STRING database interaction data can also be queried from the Cytoscape STRING app. This requires version 3.6 of Cytoscape or greater and version 1.4 of the STRING app or greater, which is available for free in the Cytoscape app store.

The PPIs reported by STRING represent functional associations between proteins. Experimental data for virus–virus and virus–host PPIs was imported from BioGrid, MintAct, DIP, HPIDB and VirusMentha.

For the prediction of CMV (virus)–cassava (host) interaction, interaction between host and virus proteins was manually curated from Viruses.STRING by searching name of template virus and host. For example, in STRINGviruses database the name for host name is given as *Arabidopsis* (template host) and the name of virus name is selected as *Cauliflower Mosaic Virus* (CaMV) (template

virus). PPI network for the template host and its corresponding infecting virus is analysed in Virus.STRING consortium.

The resulting interacting protein pairs are searched in UniProtKB for retrieving similar protein in cassava and *Cassava Mosaic Virus*.

For example, *Bean Golden Yellow Mosaic Virus (BGYMV)*, which belongs to begomovirus genus interacts with *Arabidopsis thaliana*. Nuclear Shuttle Protein (NSP) of *Bean Golden Yellow Mosaic Virus (BGYMV)* interacts with Mitogen Activated Protein Kinase4 (MAPK4) of *Arabidopsis thaliana*. Likewise, name of template virus and host is given separately. The interacting pairs of proteins between template virus and host were blasted (Blastp) against cassava proteome and CMV proteome respectively. Resulting highly identical proteins in cassava and CMV were searched in UniProtKB for its UniProt ID, UniProt name and gene name.

3.3.2 Interaction data of template homologous PPI dataset from HPIDB

Host Pathogen Interaction Database (HPIDB) 3.0 is a resource for HPI data. As of 2019 Jan, HPIDB contains 69,787 unique protein interactions between 66 host and 668 pathogen species.

Each protein in cassava and CMV is BLASTed against all the protein sequences in the HPIDB database to identify the homologs with E-value, sequence identity and aligned sequence length coverage of 1.0E-4, 50 and 80% respectively. Each protein pair between CMV and cassava is predicted to interact if an experimentally verified interaction exists between their respective homologous proteins in HPIDB database. Resulting highly identical proteins in cassava and CMV were searched in UniProtKB for its UniProt ID, UniProt name and gene name.

3.3.3 Template Plant-Virus Interactome dataset from APID

APID includes a comprehensive collection of protein interactomes for more than 400 organisms (25 of which include more than 500 interactions) produced by integration of only experimentally validated protein-protein physical interactions. The interactome data for the target organisms can be downloaded from APID. The APID search allows two categories of interactomes:

- 1) Organisms with more than 500 interactome (eg: *Arabidopsis*).
- 2) Organisms with less than 500 interactions.

As of Jan 2019, APID includes a comprehensive compendium of 90,379 distinct proteins and 678,441 singular interactions. The whole interactome data of an organism (given in the list of APID) can be downloaded from APID in delimited text format.

The interacting pairs of proteins between template virus and host were blasted (Blastp) against cassava proteome and CMV proteome respectively. Resulting highly identical proteins in cassava and CMV were searched in UniProtKB for its UniProt ID, UniProt name and gene name.

3.4 PPI PREDICTION TOOL – VirusHostPPI

VirusHostPPI employs a new prediction method for virus-host PPIs which is applicable to new viruses or hosts. The tool contains virus-host PPIs from four databases, APID, IntAct, Mentha and UniProt, which use same protein identifiers. The sequences of the proteins involved in any of the PPIs were obtained from the UniProt database. As of December 2016, there are a total of 12,157 PPIs between 29 hosts and 332 viruses (Zhou *et al.*, 2018). VirusHostPPI uses Support Vector Machine (SVM) model to predict the interactions between virus and host. Support Vector Machine (SVM) has been applied to several biological problems such as prediction of protein-protein interactions, homology detection, and analysis of

gene expression data (Cui *et al.*, 2012). Information on the viruses involved in the virus-host PPIs is available at: <http://bclab.inha.ac.kr/VirusHostPPI>.

3.5 CONSTRUCTION OF PROTEIN-PROTEIN INTERACTION NETWORK (PPIN) OF PREDICTED PROTEIN PAIRS INVOLVED IN CASSAVA-CMV INTERACTION

Work flow for the construction of Cassava-CMV is depicted in Figure 5.

3.5.1 Predicted PPI dataset formulation

The predicted protein-pairs of cassava-CMV are formulated into delimited text and excel format such that the dataset can be imported into Cytoscape. The host (cassava) is assigned with UniProt id A, UniProt name A and Gene name A. The pathogen (CMV) is assigned with UniProt id B, UniProt name B and Gene name B.

3.5.2 Cassava-CMV PPI network construction

The predicted PPI network is constructed using cytoscape version 3.7.1. The created dataset in text format is imported to Cytoscape. Before completely importing, source and target column should be selected. The source column is the gene name of predicted cassava protein and target column is the corresponding gene name of the predicted CMV protein. The imported interacting pairs of proteins can be clearly visualized. Cytoscape provides option for merging the networks, such that two networks can be visualised in a single window.

3.6 FUNCTIONAL ANNOTATION OF PREDICTED PROTEIN PAIRS

Functional annotation is an important assessment for elucidating the functional relevance of the host and pathogen proteins involved in the PPIs. Gene ontology (GO) is a comprehensive functional system to annotate the gene products. The two annotation tools used in this study are: QuickGO and Blast2GO.

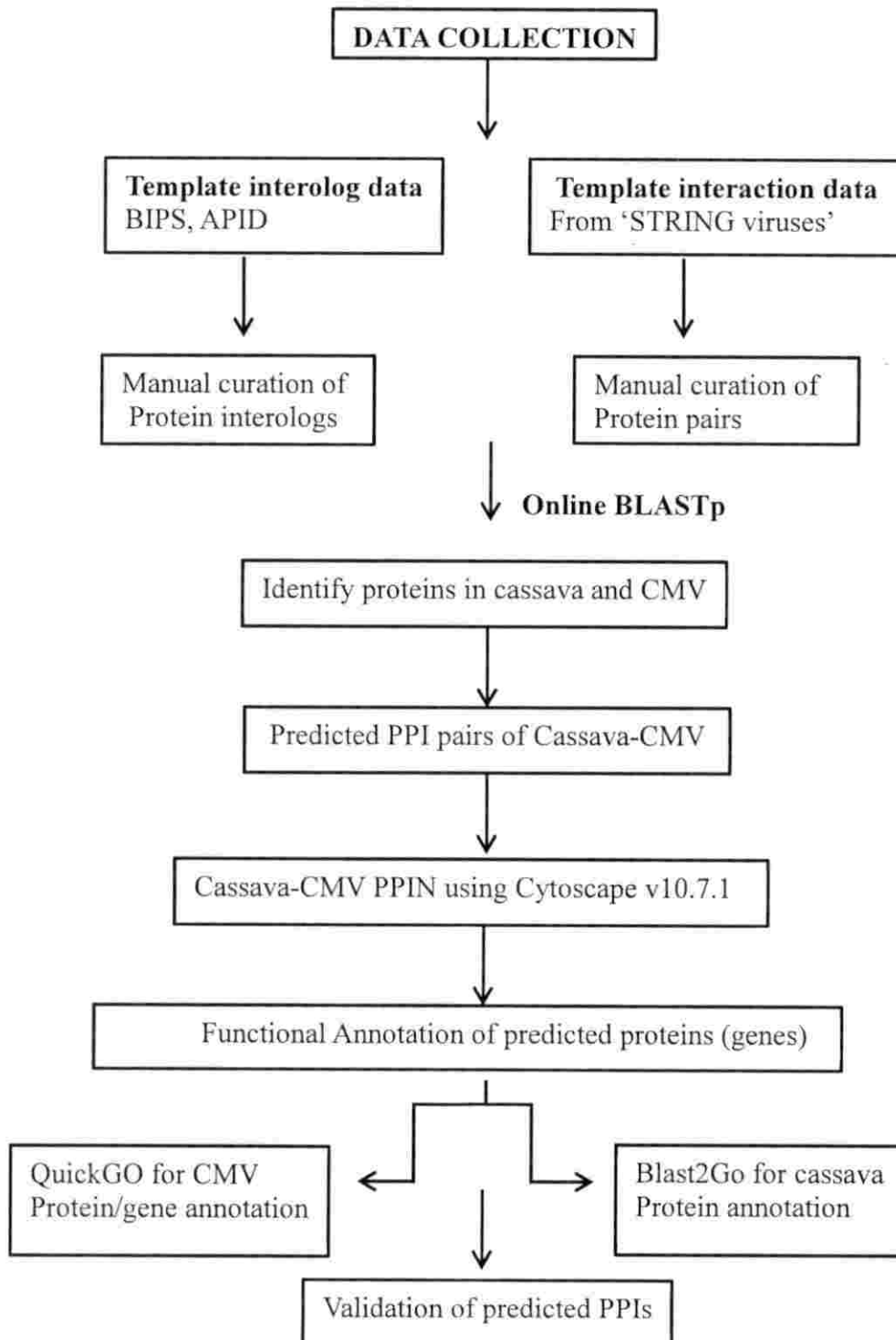


Figure 5. Work flow for the construction of Cassava-CMV PPI

3.6.1 Functional annotation using QuickGO

Functional annotation of interacting virus proteins was done using QuickGO. QuickGO is a web-based tool that allows easy browsing of the Gene Ontology (GO) and all associated electronic and manual GO annotations provided by the GO consortium annotation groups. QuickGO users can view and search information provided for GO terms (identifiers, words/phrases in the title or definition, cross-references and synonyms), as well as protein data from UniProtKB (accession numbers, names and gene symbols). Results are ranked so that terms most closely matching the query are returned first. Individual words and combinations of words are scored according to the field in which they occur and their frequency within GO. QuickGO URL: <http://www.ebi.ac.uk/QuickGO>.

3.6.2 Functional annotation using OmicsBox/Blast2GO

OmicsBox/Blast2GO is a high-quality functional annotation work station and it is a platform for analysis of genomic datasets. One can design the required custom annotation style through the many configurable parameters. Statistical charts are available to guide users in the annotation process. Blast2GO is designed for experimentalists and is user friendly.

OmicsBox/Blast2GO offers two different features to retrieve the gene/protein sequences as well as corresponding annotations from a list of identifiers within Blast2GoPRO. Both features can be found under: File > Load > Load Annotations. The expected input file is a text file with the identifiers in a single column without a header. Annotation pipeline: Blast, Interproscan, Mapping and Annotation. It can be queried online at <http://www.biobam.com>.

3.7 EXPERIMENTAL VALIDATION

The experimental validation of computationally predicted interacting protein pairs were conducted by randomly choosing a pair of interacting proteins (Catalase and Transcription activator protein). RT-PCR was performed as

described below using total RNA isolation from leaf samples of two different varieties of cassava available at ICAR-CTCRI.

3.7.1 Selected varieties of cassava

- H165: Healthy leaf sample
- H165: leaf sample showing CMV infection symptoms

3.7.2 RNA Isolation

RNA was extracted from fresh tender leaves of healthy cassava plant (variety: H165) and CMD infected cassava plant (variety: H165) using Qiagen RNeasy Plant Mini Kit, TRIzol method and CTAB method.

100 mg leaf tissue was pulverised in pre-chilled mortar and pestle using liquid nitrogen and was transferred to a 2 ml sterile tube. 1 ml of CTAB buffer (pre-warmed at 65°C for 10 min) was added followed by centrifugation at 15000 rpm for 15 minutes. Supernatant is transferred to a fresh tube and equal volume of chloroform isoamyl alcohol (24:1) is added and centrifuged at 20000 g for 10 minutes at 4°C. After centrifugation, supernatant is transferred to a fresh tube and 0.25 volume ice cold 10 M Lithium chloride is added and thoroughly mixed. This was kept for overnight incubation at 20°C. After centrifugation at 30,000 g for 30 minutes at 4°C, the pellet is washed with 75% ethanol by centrifugation at 10,000 g at 4°C. Washing step is repeated and RNA pellet was air dried at 37°C for 30 min and dissolved in 50µl DEPC water. After incubation at 37°C for 1 hour and tapping intermittently, RNA is stored at -80°C.

3.7.3 Agarose gel electrophoresis

1.2% agarose gel was used to check the quality and integrity of the extracted RNA. 1.2% agarose solution was prepared by weighing out 1.2 g agarose in a conical flask and dissolving it using 100 ml 1X TBE buffer. Every reagent was prepared in DEPC treated water. Agarose was dissolved by heating and after that the flask was allowed to cool and when the temperature of the flask

decreases, about 0.9 μl (10 mg/ml) of EtBr was added directly to the gel and gentle mixing was done.

Casting tray was prepared with combs to which gel was poured and allowed to solidify. 4 μl of isolated RNA sample mixed with 2 μl of 1X loading dye was loaded into the wells of prepared gel. Horizontal gel electrophoresis unit was used to run the gel. The gel was run for about 30 min at 110V. The run was terminated after the dye front reached $3/4^{\text{th}}$ of the gel. Then it was visualized in UV light using a gel documentation system.

3.7.4 RNA quantification

The concentration of RNA was determined using a Nano-drop (using 1 OD₂₆₀=40 μg RNA). A₂₆₀/A₂₈₀ ratios were also calculated for each sample.

3.7.5 cDNA synthesis

cDNA from the isolated RNA was prepared using Revert Aid First strand c-DNA synthesis kit. The preparation was in accordance with manufacture's protocol.

3.7.6 Primer designing for Predicted PPI pair - Primer3Plus

A primer is a short strand of RNA or DNA which generally have a size about 18-22 bases, that serves as a starting point for DNA synthesis. Primer pairs are designed to amplify the genomic region around each discovered gene. Sequences are selected for primer designing based on the experimental result of the predicted PPI pair. Primer pairs are designed using Primer 3 plus tool.

Primer3Plus is a widely used programme for designing PCR primers. PCR (Polymerase Chain Reaction) is an essential and ubiquitous tool in genetics and molecular biology. Primer3 can also design hybridization probes and sequencing primers. Primer3 picks primers for PCR reactions, considering certain important criteria such as oligonucleotide melting temperature (T_m), size, GC content,

primer-dimer possibilities, PCR product size, positional constraints within the source/template sequence, possibilities for ectopic priming (amplifying the wrong sequence) and many other constraints. Good primer design is essential for successful reactions. The parameters considered in primer designing are described below:

3.7.6.1 Primer Length

It is generally accepted that the original length of the PCR primers is 18-22 bp. This is long enough for adequate specificity and short enough for primers to bind easily to the template at the annealing temperature

3.7.6.2 Primer Melting Temperature

Primer melting temperature (T_m) is the temperature at which one half of the DNA duplex will dissociate to become single stranded and indicated the duplex stability. Primers with melting temperature in the range of 52-28°C generally produce the best results.

3.7.6.3 GC content

The GC content (the number of G's and C's in the primer as a percentage of the total bases) of primer should be 40-60%.

3.7.6.4 GC Clamp

The presence of G or C bases within the last five bases from the 3' end of primers (GC clamp) helps promote specific binding at the 3' end due to the stronger bonding of G and C bases. More than 3 G's or C's should be avoided in the last 5 bases at the 3' end of the primer.

3.7.7 RT-qPCR validation

Real Time quantitative Polymerase Chain Reaction (RT-qPCR) is a tool used for gene expression studies. The qPCR reaction was performed with forward and reverse primers (specific to the predicted protein catalase in cassava and transcription activator protein in *Cassava mosaic virus*). qPCR analysis for the

samples were done at Rajiv Gandhi Centre for Biotechnology (RGCB) Bio Innovation Centre, Trivandrum. The reaction profile is depicted in Table 1.

Table 1. RT-qPCR reaction profile

Components	Volume (μ l)
Diluted cDNA	1.5
Forward primer	1
Reverse primer	1
DyNAmo Flash SYBR Green qPCR master mix	5
Double distilled water	1.5

3.7.7.1 Thermal Profile

Initial denaturation: 95°C 5min

Denaturation: 95° 10s

Annealing: 55°C

Extension: 72°C 15-30s

Number of cycles: 35-45 cycles, step 2-4

After the completion of the real time reactions, the threshold cycle (C_T) was recorded and gene expression level was calculated using comparative C_T method. The gene expression level of two proteins in Cassava leaves are represented as $2^{-\Delta\Delta C_t}$.

$$\Delta C_t = C_t (\text{target gene}) - C_t (\text{reference gene})$$

$$\Delta\Delta C_t = \Delta C_t (\text{sample}) - \Delta C_t (\text{control}).$$

RESULTS

4. RESULT

The results of the study “Modeling of Cassava-Cassava Mosaic Virus interactions with computational biology and bioinformatics approach” carried out at the Section of Extension and Social Sciences, ICAR-Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram during 2018-2019 are presented in this chapter.

The study focuses on the prediction of protein-protein interaction between cassava and CMV. For this, firstly, Protein-Protein Interaction (PPI) in cassava is predicted. Secondly, PPI between Cassava-CMV is predicted and the predicted protein pairs are functionally annotated. The predicted protein pairs in Cassava-CMV interaction is analysed for the presence of virus resistance proteins. In both the prediction, interolog-based approach is used.

4.1 COLLECTION OF DATA FROM LITERATURES FOR CASSAVA PROTEIN-PROTEIN INTERACTION PREDICTION

4.1.1 Construction of cassava PPI network using interolog-based method

The proteomic dataset used for the study was generated using the interolog-based method. Interolog method, relies on existing data, is adopted for PPI prediction. Upon the homology-based principle of this method, seven plant species were selected as templates, based on three criteria: Most abundant PPI information (model plant *Arabidopsis*); starch-storing crops (potato, rice and maize); closely related to cassava (castor bean, poplar and soybean). According to PPI information from various databases:

Arabidopsis thaliana contains: 235,215 interactions of 17,962 proteins.

Oryza sativa (rice) contains: 76,829 interactions of 5,219 proteins

Solanum tuberosum (potato) contains: 42 interactions of 48 proteins

Zea mays (maize) contains: 25 interactions of 29 proteins

Glycine max (soybean) contains: 10 interactions of 12 proteins

Ricinus communis (castor bean) contains: 10 interactions of 10 proteins

Populus trichocarpa (poplar) contains: 8 interactions of 10 proteins.

To infer PPI information for cassava from each template plant, BLASTp search of the cassava genome sequence database was carried out. The cassava orthologous proteins that showed identity percentage ≥ 60 , coverage percentage $\geq 80\%$ and e-value $\leq 10^{-10}$ were identified. Interactions were inferred as orthologous PPIs in cassava if those orthologous proteins matched the proteins of template plants that had previously been identified to have protein-protein interaction. Based on the results obtained, majority of the PPIs were from *Arabidopsis thaliana*. Protein-Protein Interactions (PPIs) in plant templates and cassava is shown in Table 2.

Table 2. Protein-Protein Interactions (PPIs) in plant templates and cassava

Template Plants	Genome Information		PPI information		cassava interactome	
	No. of genes	No. of proteins	No. of PPI	No. of proteins	Inferred PPIs in Cassava	Orthologs in Cassava
<i>Arabidopsis</i>	27,416	35,386	235,215	17,962	90,069	7,193
Rice	55,986	154,310	76,829	5,219	212	84
Potato	35,119	59,699	42	48	19	15
Maize	32,540	88,383	25	29	5	8
Soybean	54,175	83,795	10	12	7	7
Poplar	41,335	83,796	8	10	5	7
Castor bean	25,878	31,576	10	10	2	2
					90,173	7,209

The resulting interolog-based PPI network of cassava consists of 90,173 interactions interconnecting 7,209 proteins, which accounted for about 21% of proteins in the whole genome. The predicted interacting pairs of proteins are represented in the form of a network (interactome). The network is generated using Cytoscape v3.7.1. Protein-Protein Interaction Network (PPIN) of cassava is shown in Figure 6.

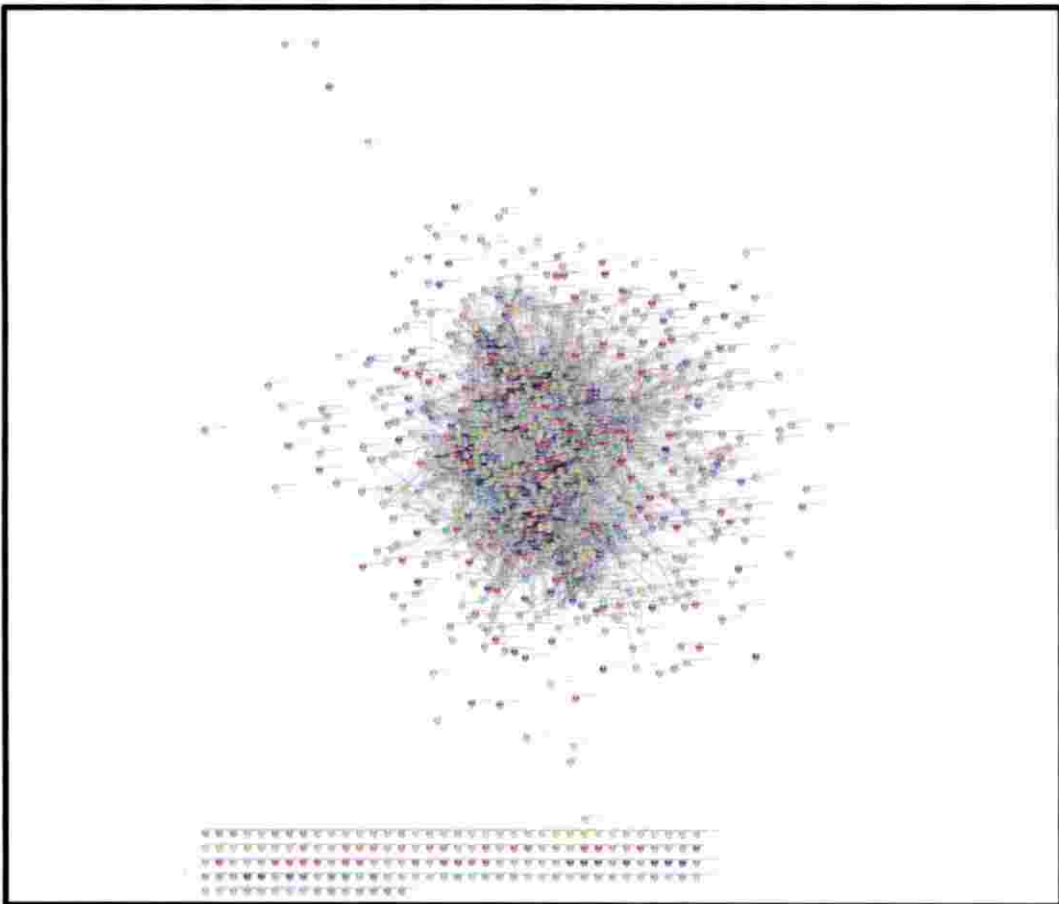


Figure 6. Cassava PPI network derived by interolog based method. The network is generated using the Cytoscape tool and STRING app.

4.2 DATA MINING OF PLANT-VIRUS INTERACTIONS FOR THE PREDICTION OF CASSAVA-CMV PPI

In this study the template plant used is *Arabidopsis thaliana* and its corresponding infecting viruses that are similar in taxonomy with CMV is selected. For the prediction of interacting protein pairs between cassava and *Cassava Mosaic Virus* (CMV) three datasets were used; STRINGviruses consortium 2018, Host Pathogen Interaction Database (HPIDB) and Agile Protein Interactome DataServer (APID).

STRING Viruses consortium 2018 is employed for the analysis of template Plant-Virus Interaction (PVI) network. In this study the PVI network of model plant (template plant) *A. thaliana* is taken.

Input

Template virus and plant species from which the user wants to predict putative binding partners can be selected from the list of name given in the dataset. In this study, *Arabidopsis thaliana* is selected as the model plant and corresponding plant virus species are selected. 25 virus species and its different strain were searched against *Arabidopsis thaliana*. Out of 25 viruses selected, 7 virus species showed interaction with *A. thaliana*. The list of the virus species showing interaction with *A.thaliana* are given in Table 3.

Output

The output is a network containing protein-protein interaction between virus and host that can be viewed or downloaded. The user can browse the data associated with the partner proteins, which redirects to UniProt.

Table 3. Virus species interacting with *Arabidopsis thaliana* (obtained from STRING viruses consortium, 2018)

NCBI Taxon Id	STRING type	STRING name
10840	Core	<i>Beet curly top virus</i> (strain California/logan)
12216	Core	<i>Potato virus Y</i>
37128	Core	<i>Potato mop-top v irus</i>
12167	Core	<i>Potato virus M</i>
12305	Core	<i>Cucumber mosaic virus</i>
10641	Core	<i>Cauliflower mosaic virus</i>
220340	Core	<i>Bean golden yellow mosaic virus</i>

HPIDB is a database employed for homolog PPI identification. Cassava and CMV proteins are BLASTed against plant proteins in HPIDB database. From this, blast hits for *A. thaliana* were selected. A model of the blast result of cassava proteins are depicted in Figure 7.

Search by Sequence Results

Download the host-pathogen interaction data from the zip folder

A description of results is here [README](#)

Utilizing Blast Version: BLASTP 2.5.0+

Database: blast_data/total.fasta

Qry ID	Hit ID	E-Value	Pct ID	Qry Cov	Bit Score	A
tr A0A2C9U1Q2 A0A2C9U1Q2_MANES	UNIPROT_AC:Q38997	5.81e-90	50.190	63	281	26
tr A0A2C9U1Q2 A0A2C9U1Q2_MANES	UNIPROT_AC:P92958	2.56e-87	50.394	61	274	25
tr A0A2C9VKE1 A0A2C9VKE1_MANES	UNIPROT_AC:Q9FR53	0.0	81.000	91	1474	90
tr A0A2C9VX86 A0A2C9VX86_MANES	UNIPROT_AC:Q42547	0.0	78.252	100	830	45
tr A0A2C9VX86 A0A2C9VX86_MANES	UNIPROT_AC:A0A1J9W864	4.42e-175	52.643	96	500	47
tr A0A2C9VX86 A0A2C9VX86_MANES	UNIPROT_AC:Q81UM1	1.09e-136	51.958	77	409	38
tr A0A2C9VVT7 A0A2C9VVT7_MANES	UNIPROT_AC:Q42547	0.0	78.742	100	788	44
tr A0A2C9VVT7 A0A2C9VVT7_MANES	UNIPROT_AC:A0A1J9W864	3.80e-169	52.402	99	484	45
tr A0A2C9VVT7 A0A2C9VVT7_MANES	UNIPROT_AC:Q81UM1	4.11e-134	52.575	79	400	36
tr A0A2C9VVU3 A0A2C9VVU3_MANES	UNIPROT_AC:Q42547	0.0	76.220	100	811	44
tr A0A2C9VVU3 A0A2C9VVU3_MANES	UNIPROT_AC:A0A1J9W864	1.92e-169	51.050	96	486	47
tr Q9SW99 Q9SW99_MANES	UNIPROT_AC:Q42547	0.0	77.033	100	822	45
tr Q9SW99 Q9SW99_MANES	UNIPROT_AC:A0A1J9W864	1.71e-168	51.055	96	484	47
tr A0A2C9WMD1 A0A2C9WMD1_MANES	UNIPROT_AC:Q42547	0.0	74.419	100	565	34

Figure 7. A model of HPIDB blast result

The dataset can be downloaded in zip file format. Zip folder contents:

- 1) The blast tab delimited results file.
- 2) The homologous HPI results tab delimited file. Records contain the query id, e-value, query coverage, percentage identity, HPIDB homologous pathogen hit and the HPIDB interacting partners.
- 3) Unique interacting protein results tab delimited file contains unique HPIDB host and HPIDB pathogen proteins.

The third database used is Agile Protein Interactome Dataserver (APID). Interactions were obtained for *A. thaliana* with *Brome Mosaic Virus* (BMV), *Cauliflower Mosaic Virus* (CaMV), *Rice Dwarf Virus* (RDV), *Tobacco Mosaic Virus* (TMV), *Tomato Yellow Leaf Curl Virus* (TYLCV), *Tomato Golden Mosaic Virus* (TGMV), and *Tomato Mosaic Virus* (ToMV). From APID, 19 interactions were predicted for cassava-CMV interaction.

Combining interaction data from three databases, 351 proteins in cassava is predicted to interact with 11 proteins in CMV.

4.3 PPI PREDICTION TOOL – VirusHostPPI

Through the interolog-based method, 351 interacting protein pairs were obtained for 11 CMV. VirusHostPPI prediction tool enables the confirmation of the predicted interacting protein pairs between cassava and CMV.

If the host protein sequence and the virus protein sequences are given as input, the tool detects whether the proteins are interacting or not. The protein pairs predicted through interolog-based method were filtered through VirusHostPPI prediction tool and it is found that 114 proteins of cassava are interacting with 10 proteins of CMV. The predicted protein pairs between cassava and *Cassava mosaic virus* are shown in Table 4.

Table 4. Proteins in cassava predicted to interact with CMV

Sl No	Cassava UniProt Name A	Gene name A	CMV UniProt Name B	Gene name B
1	A0A2C9U8S7	MANES_16G041200	H8WR50	AC1
2	A0A2C9U1Q2	MANES_18G055300	H8WR50	AC1
3	A0A2C9UYQ8	MANES_11G066400	H8WR50	AC1
4	A0A2C9U4V6	MANES_18G137500	H8WR50	AC1
5	A0A2C9UCD6	MANES_16G128800	H8WR50	AC1
6	A0A2C9VKE1	MANES_07G055800	H8WR50	AC1
7	A0A2C9U3K5	MANES_18G144600	H8WR50	AC1
8	A0A2C9U398	MANES_18G108400	H8WR50	AC1
9	A0A2C9UNM2	MANES_13G002600	H8WR50	AC1
10	A0A2C9V0P3	MANES_11G119700	H8WR50	AC1
11	A0A2C9V0Q1	MANES_11G119700	H8WR50	AC1
12	A0A2C9V4R5	MANES_10G087600	H8WR50	AC1
13	A0A251LK88	MANES_02G205900	H8WR50	AC1
14	A0A251LK92	MANES_02G205900	H8WR50	AC1
15	A0A2C9WFM4	MANES_02G199900	H8WR50	AC1
16	A0A251LKM5	MANES_02G199800	H8WR50	AC1
17	A0A2C9VX86	MANES_05G130700	TRAP_ICMV	AC2, AL2
18	A0A2C9VVT7	MANES_05G130700	TRAP_ICMV	AC2, AL2
19	Q9SW99	MANES_18G004500	TRAP_ICMV	AC2, AL2
20	A0A2C9VVU3	MANES_05G130500	TRAP_ICMV	AC2, AL2
21	A9YME8	CAT2	TRAP_ICMV	AC2, AL2
22	A0A2C9WMD1	MANES_01G154400	TRAP_ICMV	AC2, AL2
23	A0A2C9TZH3	MANES_18G004400	TRAP_ICMV	AC2, AL2
24	A0A2C9WLH4	MANES_01G165500	H8WR48	AC3
25	A0A2C9WD70	MANES_02G123000	H8WR48	AC3
26	A0A2C9W3T7	MANES_03G016800	H8WR51_9GEMI	AC4
27	A0A2C9VUS4	MANES_05G096100	H8WR51_9GEMI	AC4
28	A0A2C9WC92	MANES_02G041800	H8WR51_9GEMI	AC4

29	A0A2C9VPI5	MANES_06G096600	H8WR51_9GEMI	AC4
30	A0A2C9WNB8	MANES_01G187900	H8WR51_9GEMI	AC4
31	A0A2C9VUV8	MANES_05G096100	H8WR51_9GEMI	AC4
32	A0A2C9W164	MANES_04G017500	H8WR51_9GEMI	AC4
33	A0A2C9U0M9	MANES_18G021400	H8WR51_9GEMI	AC4
34	A0A2C9W3M2	MANES_03G013700	H8WR51_9GEMI	AC4
35	A0A2C9WCA3	MANES_02G089000	CAPSD_ICMV	AR1, AV1
36	A0A2C9UNR1	MANES_13G041100	CAPSD_ICMV	AR1, AV1
37	A0A2C9U746	MANES_17G111400	CAPSD_ICMV	AR1, AV1
38	A0A2C9V1V9	MANES_11G163400	CAPSD_ICMV	AR1, AV1
39	A0A2C9VPE6	MANES_06G015000	CAPSD_ICMV	AR1, AV1
40	A0A2C9U5Z3	MANES_17G050100	CAPSD_ICMV	AR1, AV1
41	A0A2C9VIE7	MANES_07G042200	CAPSD_ICMV	AR1, AV1
42	A0A2C9V824	MANES_09G052800	CAPSD_ICMV	AR1, AV1
43	A0A2C9VS27	MANES_05G005300	CAPSD_ICMV	AR1, AV1
44	A0A2C9VU82	MANES_05G005300	CAPSD_ICMV	AR1, AV1
45	A0A2C9VS73	MANES_05G005300	CAPSD_ICMV	AR1, AV1
46	A0A2C9UAD5	MANES_16G106800	CAPSD_ICMV	AR1, AV1
47	A0A251L698	MANES_03G028900	CAPSD_ICMV	AR1, AV1
48	A0A251L6B6	MANES_03G028900	CAPSD_ICMV	AR1, AV1
49	A0A2C9VNM8	MANES_06G072300	CAPSD_ICMV	AR1, AV1
50	A0A2C9UP13	MANES_13G051400	CAPSD_ICMV	AR1, AV1
51	A0A2C9VTU5	MANES_06G163000	CAPSD_ICMV	AR1, AV1
52	A0A2C9VRR0	MANES_06G123900	CAPSD_ICMV	AR1, AV1
53	A0A251KRR0	MANES_05G014800	CAPSD_ICMV	AR1, AV1
54	A0A2C9W4N0	MANES_03G002200	CAPSD_ICMV	AR1, AV1
55	A0A2C9WKT2	MANES_01G143300	CAPSD_ICMV	AR1, AV1
56	A0A2C9WN37	MANES_01G143400	CAPSD_ICMV	AR1, AV1
57	A0A2C9WM25	MANES_01G143800	CAPSD_ICMV	AR1, AV1
58	A0A2C9WHR8	MANES_01G043900	CAPSD_ICMV	AR1, AV1
59	A0A2C9WC46	MANES_02G001600	CAPSD_ICMV	AR1, AV1

60	A0A2C9W5R6	MANES_03G002200	CAPSD_ICMV	ARI, AV1
61	A0A2C9W041	MANES_05G196000	CAPSD_ICMV	ARI, AV1
62	A0A2C9VRP8	MANES_06G172900	CAPSD_ICMV	ARI, AV1
63	A0A2C9W4Z3	MANES_04G154400	CAPSD_ICMV	ARI, AV1
64	A0A2C9VIE3	MANES_07G042400	CAPSD_ICMV	ARI, AV1
65	A0A2C9VE39	MANES_08G065000	CAPSD_ICMV	ARI, AV1
66	A0A2C9VUQ0	MANES_05G091800	CAPSD_ICMV	ARI, AV1
67	A0A2C9UKB8	MANES_14G097400	CAPSD_ICMV	ARI, AV1
68	A0A199UA28	MANES_S084200	CAPSD_ICMV	ARI, AV1
69	A0A2C9VSE5	MANES_05G017100	CAPSD_ICMV	ARI, AV1
70	A0A2C9UFZ4	MANES_15G151500	CAPSD_ICMV	ARI, AV1
71	A0A2C9U024	MANES_18G002500	CAPSD_ICMV	ARI, AV1
72	A0A199UA28	MANES_S084200	CAPSD_ICMV	ARI, AV1
73	A0A2C9WR64	MANES_01G238100	CAPSD_ICMV	ARI, AV1
74	A0A2C9W4Q2	MANES_03G002200	CAPSD_ICMV	ARI, AV1
75	A0A2C9URD4	MANES_13G083400	CAPSD_ICMV	ARI, AV1
76	A0A2C9UQA8	MANES_13G051400	CAPSD_ICMV	ARI, AV1
77	A0A251LEQ3	MANES_02G035400	CAPSD_ICMV	ARI, AV1
78	A0A2C9U2C3	MANES_18G079600	CAPSD_ICMV	ARI, AV1
79	A0A251LCR3	MANES_03G201400	O90282	AV1
80	A0A2C9VFA0	MANES_08G110100	O90282	AV1
81	A0A2C9VBR2	MANES_09G175000	O90282	AV1
82	A0A2C9VTE1	MANES_05G001300	O90282	AV1
83	A0A2C9VBQ4	MANES_09G178100	O90282	AV1
84	A0A2C9VFC8	MANES_08G113100	O90282	AV1
85	A0A2C9VT85	MANES_05G043200	H8WR53	BC1
86	A0A2C9VW90	MANES_05G145400	H8WR53	BC1
87	A0A2C9W0I7	MANES_04G075000	H8WR53	BC1
88	A0A2C9V4J0	MANES_10G047400	H8WR53	BC1
89	A0A2C9W634	MANES_03G013200	H8WR53	BC1
90	A0A2C9W0F0	MANES_04G075000	H8WR53	BC1

91	A0A2C9WG65	MANES_02G219700	Q65975	ORF2
92	A0A199UAY8	PCaP1	Q89703	ORF3
93	A0A2C9UXZ5	MANES_11G028400	IBMP_CSVMV	ORF4
94	A0A2C9W286	MANES_04G137900	IBMP_CSVMV	ORF4
95	A0A2C9W3E8	MANES_03G006600	IBMP_CSVMV	ORF4
96	A0A2C9UCF2	MANES_16G132500	IBMP_CSVMV	ORF4
97	A0A251K2X1	MANES_09G040700	IBMP_CSVMV	ORF4
98	A0A2C9VFH4	MANES_08G039600	IBMP_CSVMV	ORF4
99	A0A251K7S8	MANES_09G174300	IBMP_CSVMV	ORF4
100	A0A2C9VFE3	MANES_08G114400	IBMP_CSVMV	ORF4
101	A0A2C9VKP1	MANES_07G120100	IBMP_CSVMV	ORF4
102	A0A2C9V4Q7	MANES_10G025200	IBMP_CSVMV	ORF4
103	A0A2C9UAB5	MANES_16G090500	IBMP_CSVMV	ORF4
104	A0A2C9WK52	MANES_01G119800	IBMP_CSVMV	ORF4
105	A0A2C9VM01	MANES_06G016100	IBMP_CSVMV	ORF4
106	A0A2C9VMX9	MANES_07G126300	IBMP_CSVMV	ORF4
107	A0A2C9VKV5	MANES_07G126300	IBMP_CSVMV	ORF4
108	A0A2C9VMB7	MANES_07G126500	IBMP_CSVMV	ORF4
109	A0A2C9UGV6	MANES_15G140500	IBMP_CSVMV	ORF4
110	A0A2C9U7A7	MANES_17G091600	IBMP_CSVMV	ORF4
111	A0A2C9WCR1	MANES_02G056600	IBMP_CSVMV	ORF4
112	A0A2C9VV50	MANES_05G107000	IBMP_CSVMV	ORF4
113	A0A2C9VP44	MANES_06G088600	IBMP_CSVMV	ORF4
114	A0A2C9VMR2	MANES_06G036000	IBMP_CSVMV	ORF4

4.4 CONSTRUCTION OF PROTEIN-PROTEIN INTERACTION NETWORK (PPIN) OF PREDICTED PROTEIN PAIRS INVOLVED IN CASSAVA-CMV INTERACTION

To predict the genome wide interactions, all proteins of cassava and CMV are paired up, which constitute 351 protein pairs. A total of 351 probable protein pairs were predicted from interolog based method. After filtering of 351 protein pairs in VirusHostPPI prediction tool, 114 protein pairs were found to be interacting which includes 114 cassava proteins and 10 CMV proteins. Cytoscape is employed for the construction of PPIN. The interaction network of the predicted PPI is shown in Figure 8, 9, 10& 11. On an average, one CMV protein has at least one cassava interacting partner. Predicted genes in CMV are shown in Table 5.

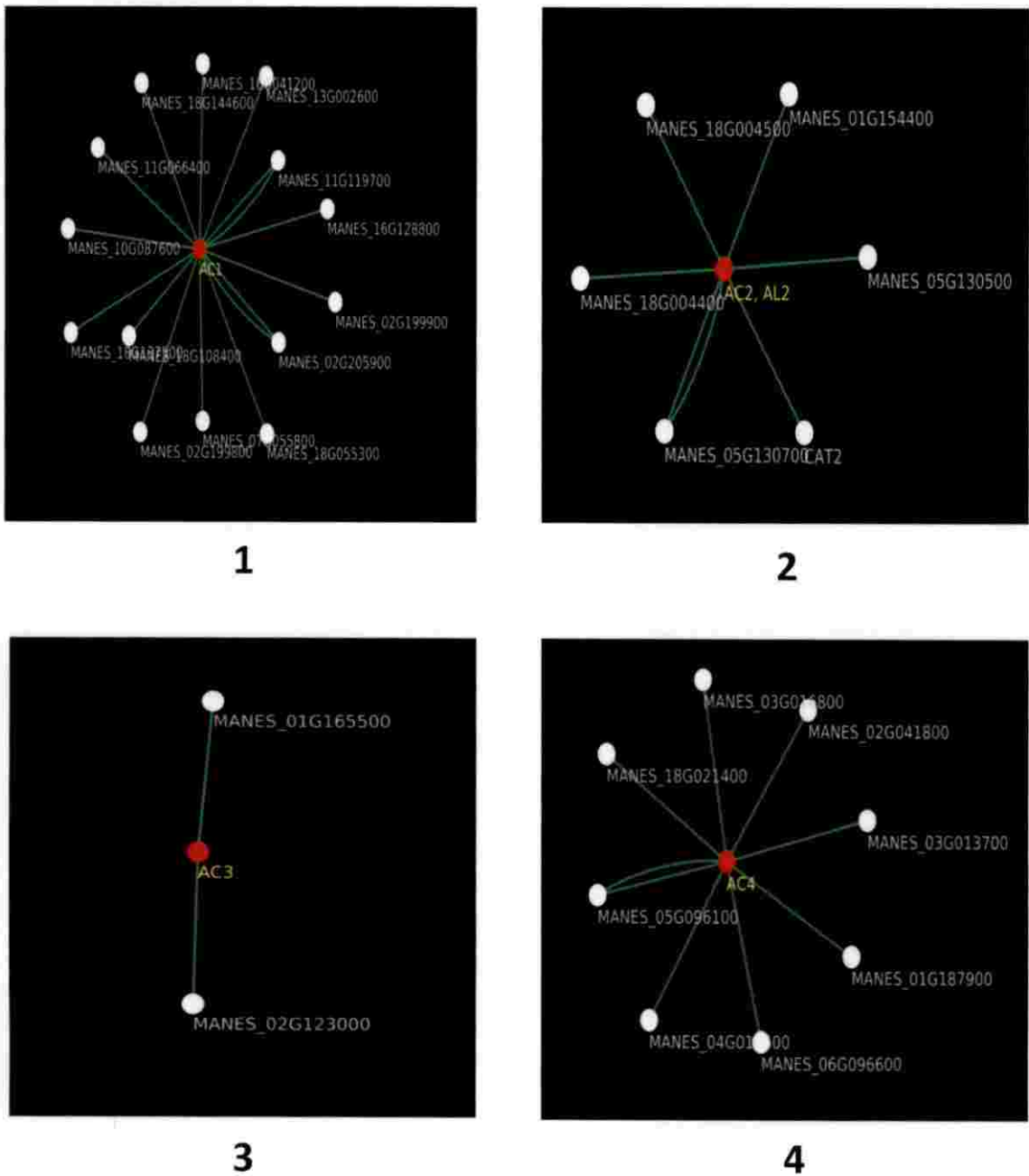
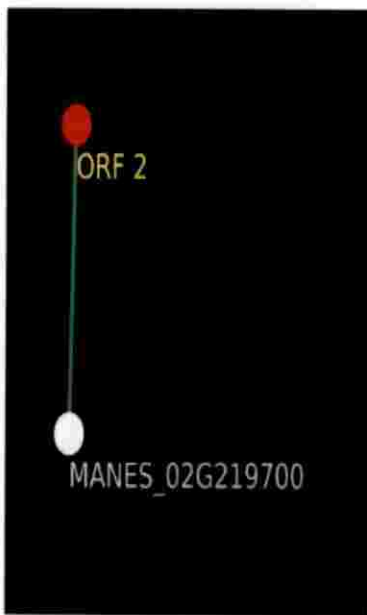
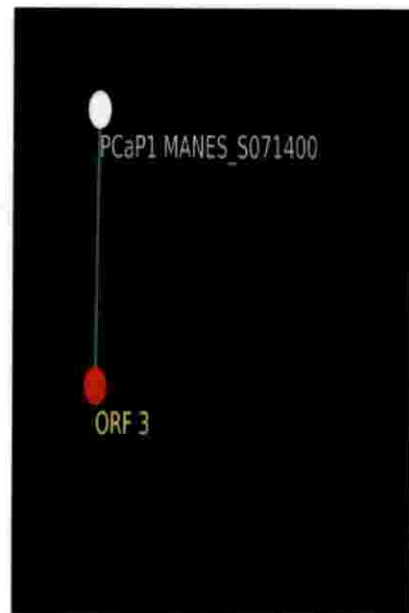


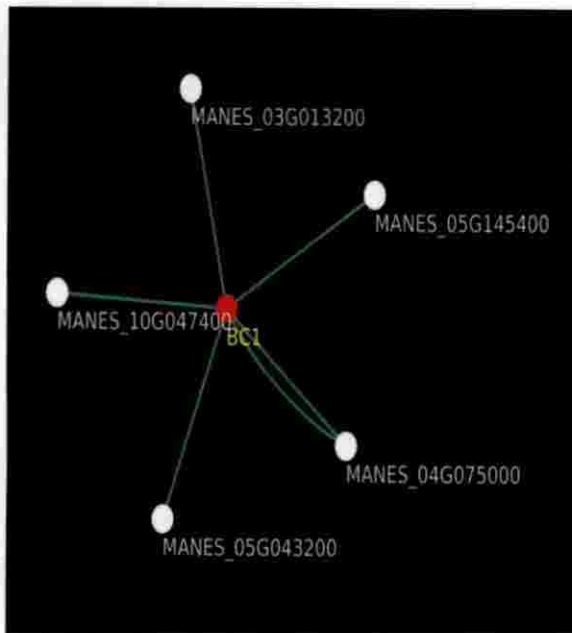
Figure 8. Predicted PPIN of Cassava-CMV. White nodes denotes gene name of cassava and red node denotes gene name of CMV. Red node represents AC1, AC2, AC3 and AC4 respectively.



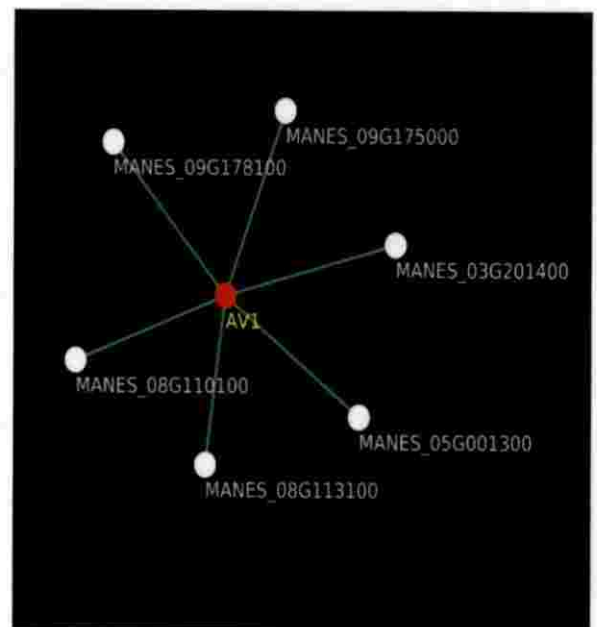
5



6

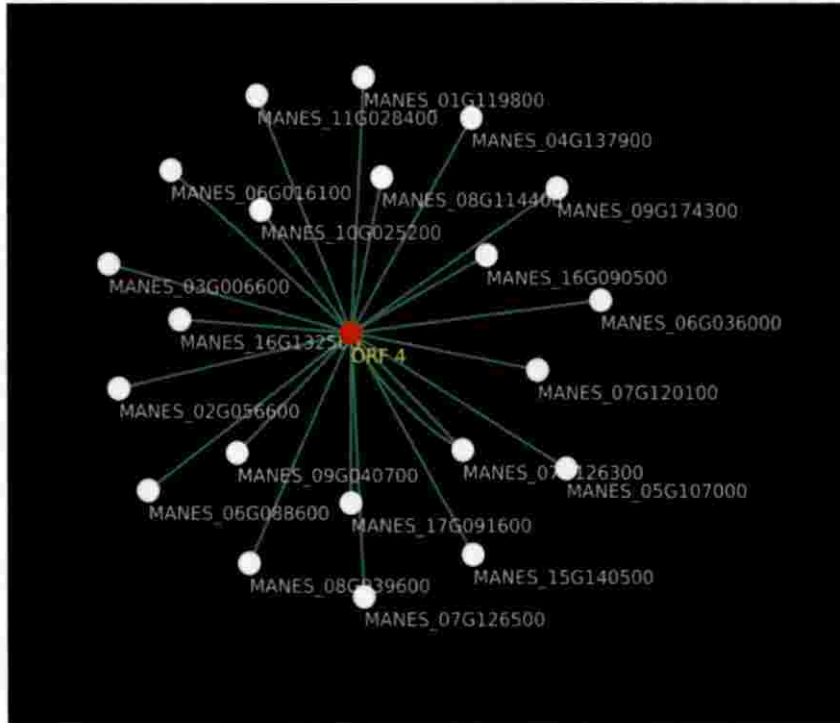


7

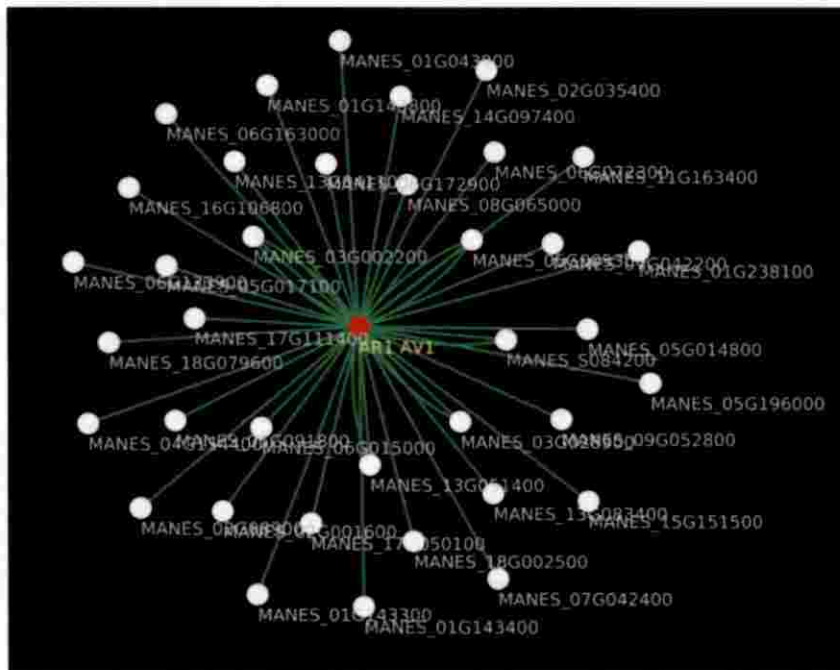


8

Figure 9. Predicted PPIN of Cassava-CMV. Red node represents ORF2, ORF3, BC1, AV1 respectively



9



10

Figure 10. Predicted PPIN of Cassava-CMV. Red node represents ORF4 and AR1

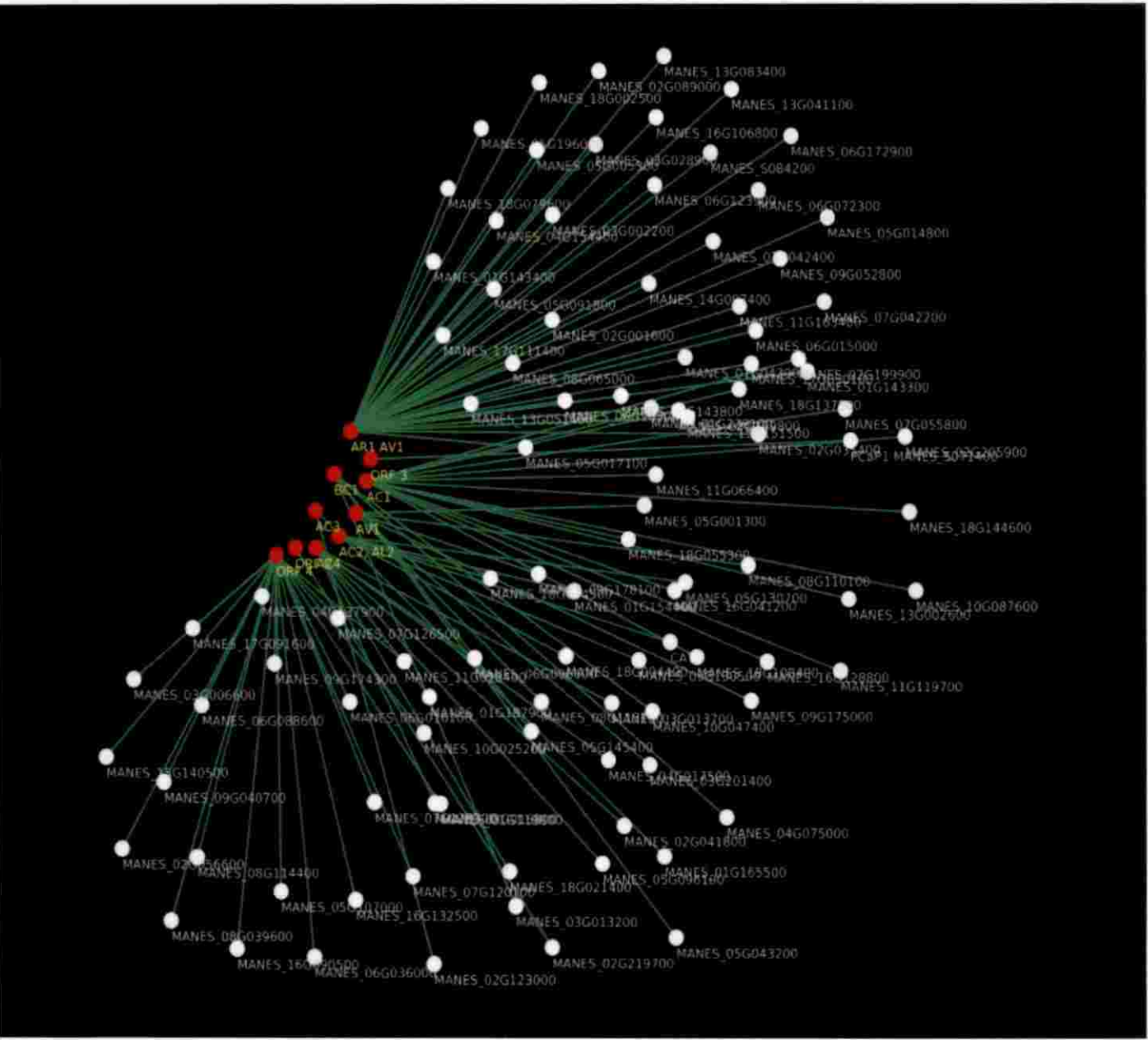


Figure 11. Merged Cassava-CMV PPIN

Table 5. Predicted genes in CMV interacting with cassava

Virus gene name	No. of interaction	No. of host genes	No. of host proteins
AR1, AV1	44	37	44
ORF4	22	21	22
AC1	16	14	16
AC4	9	8	9
AC2, AL2	7	6	7
AV1	6	6	6
BC1	6	5	6
AC3	2	2	2
ORF2	1	1	1
ORF3	1	1	1
Total	114	101	114

Predicted effector hubs:

The effectors of CMV with highest number of edges (hubs) are AR1, ORF4 and AC1. These effectors have more than 10 PPIs in the Cassava-CMV interactome. There are effectors with less than 10 predicted PPIs. These are AC4, AC2, AV1, BC1, AC3, ORF2 and ORF3. These hub proteins play important role in pathogenesis, hence can be further investigated for deciphering virulence mechanism.

4.5 FUNCTIONAL ANNOTATION OF PREDICTED PROTEIN PAIRS

The presence of annotated functional categories that are closely related to host defence and pathogen infection support the validity of the predicted PPIs of the prediction models. This study used the biological process, molecular process and cellular components (GO term) to see the relevance of the predicted proteins.

Functional annotation of predicted proteins in cassava was obtained from Omicxbox Blast2GO (<http://www.biobam.com>). Pipeline of Blast2Go is depicted in Figure 12.

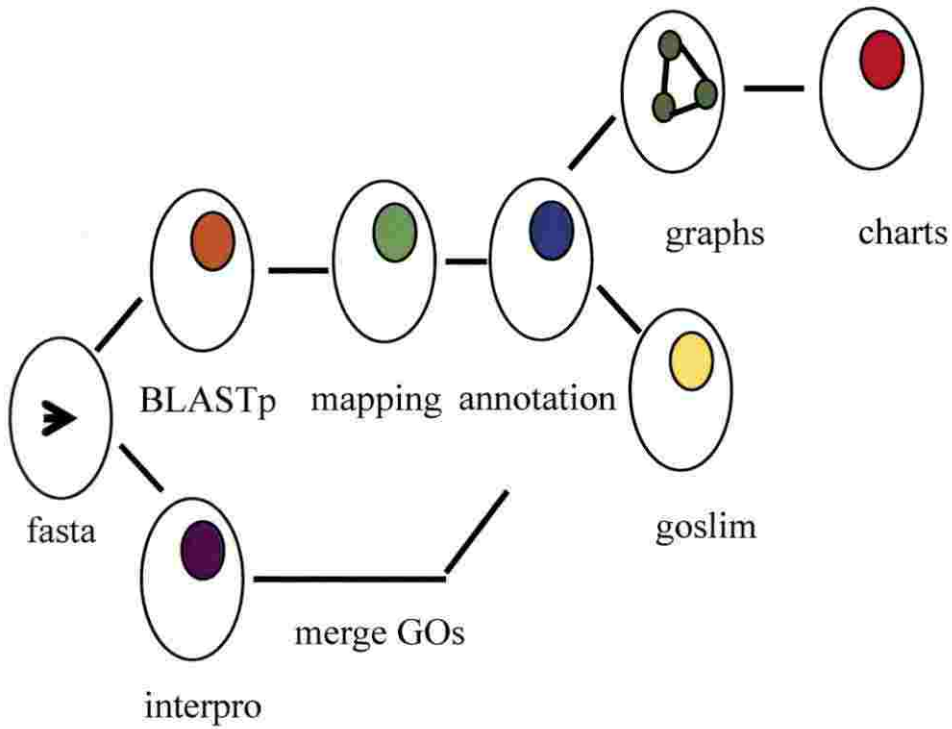


Figure 12. Blast2GO pipeline

GO annotations of 99% of the *Manihot esculenta* proteins were obtained from Blast2GO. Out of 114 proteins, 113 proteins sequences showed blast result, interProScan results, mapping and annotation. Analysis progress of 114 predicted cassava proteins is shown below (Figure 13).

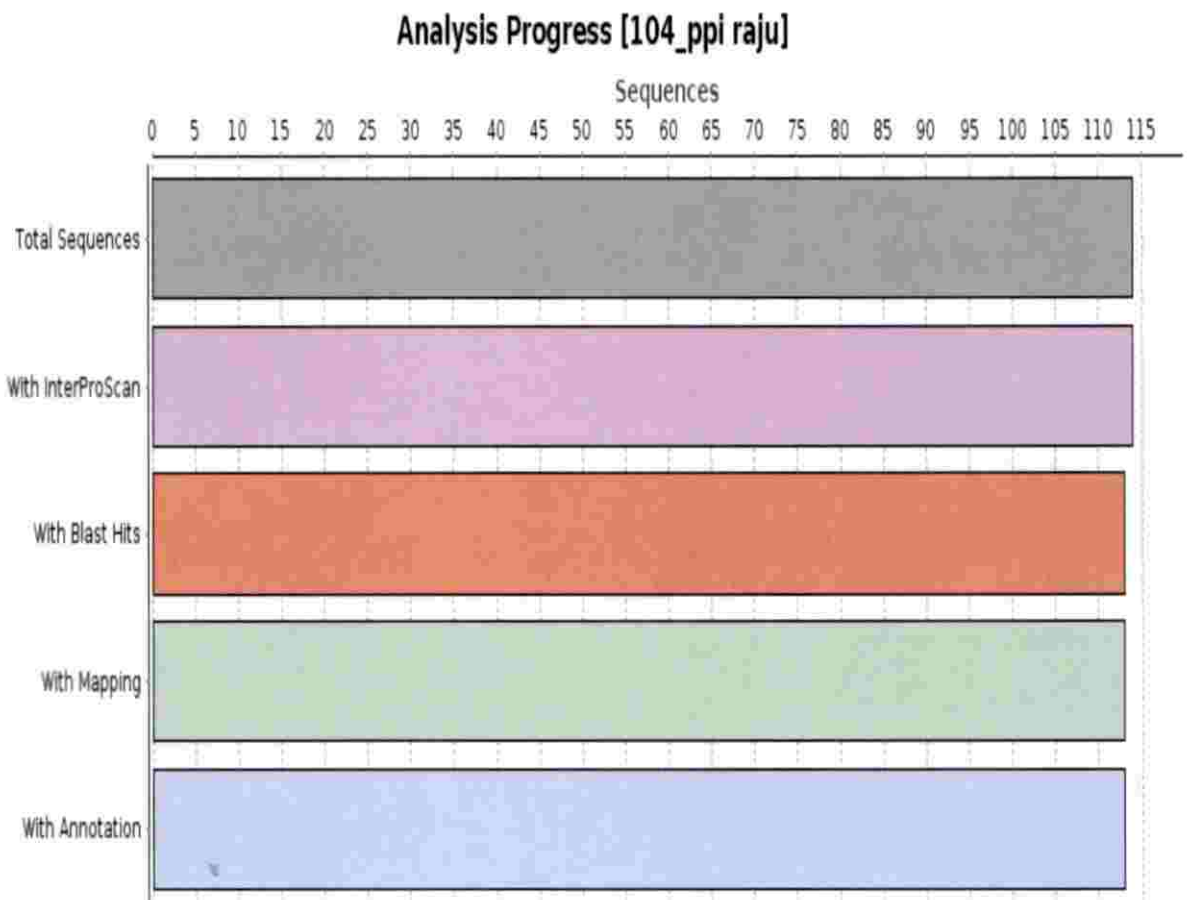


Figure 13. Analysis progress of predicted cassava proteins

InterProScan family distribution results were obtained from the second step in Blast2GO, which showed that majority of the proteins comes under NAC domain superfamily (IPR036093). The NAC domain is an N-terminal module of nearly 160 amino acids, which is found in proteins of the NAC family of plant-specific transcriptional regulators. NAC proteins are involved in developmental processes, including formation of the shoot apical meristem, floral organs and lateral shoots, as well as in plant hormonal control and defence. The NAC domain has been shown to be a DNA-binding domain (DBD) and a dimerization domain. InterPoScan family distribution and domain distribution are depicted in Figure 14 & 15 respectively. A graph level 5 pie chart showing cellular component of the predicted proteins in cassava is shown in Figure 16.

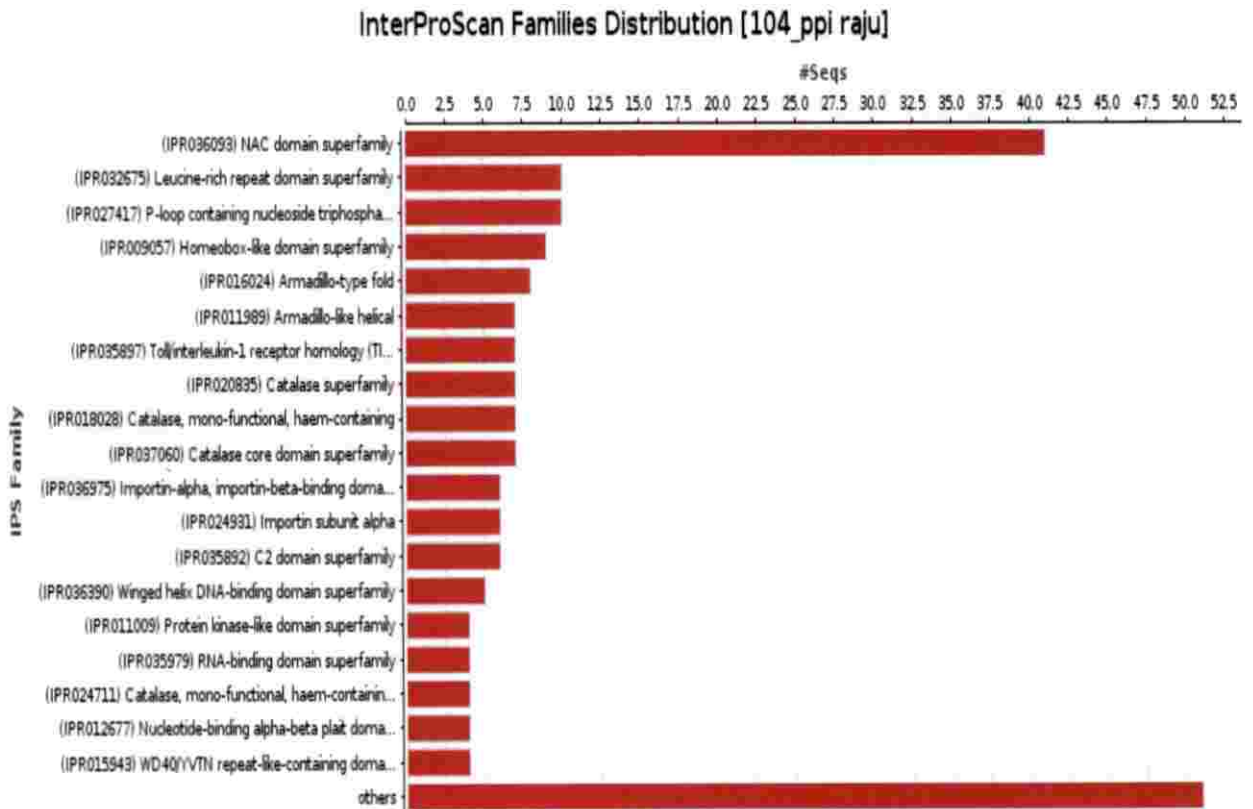


Figure 14. InterProScan families distribution of predicted cassava proteins

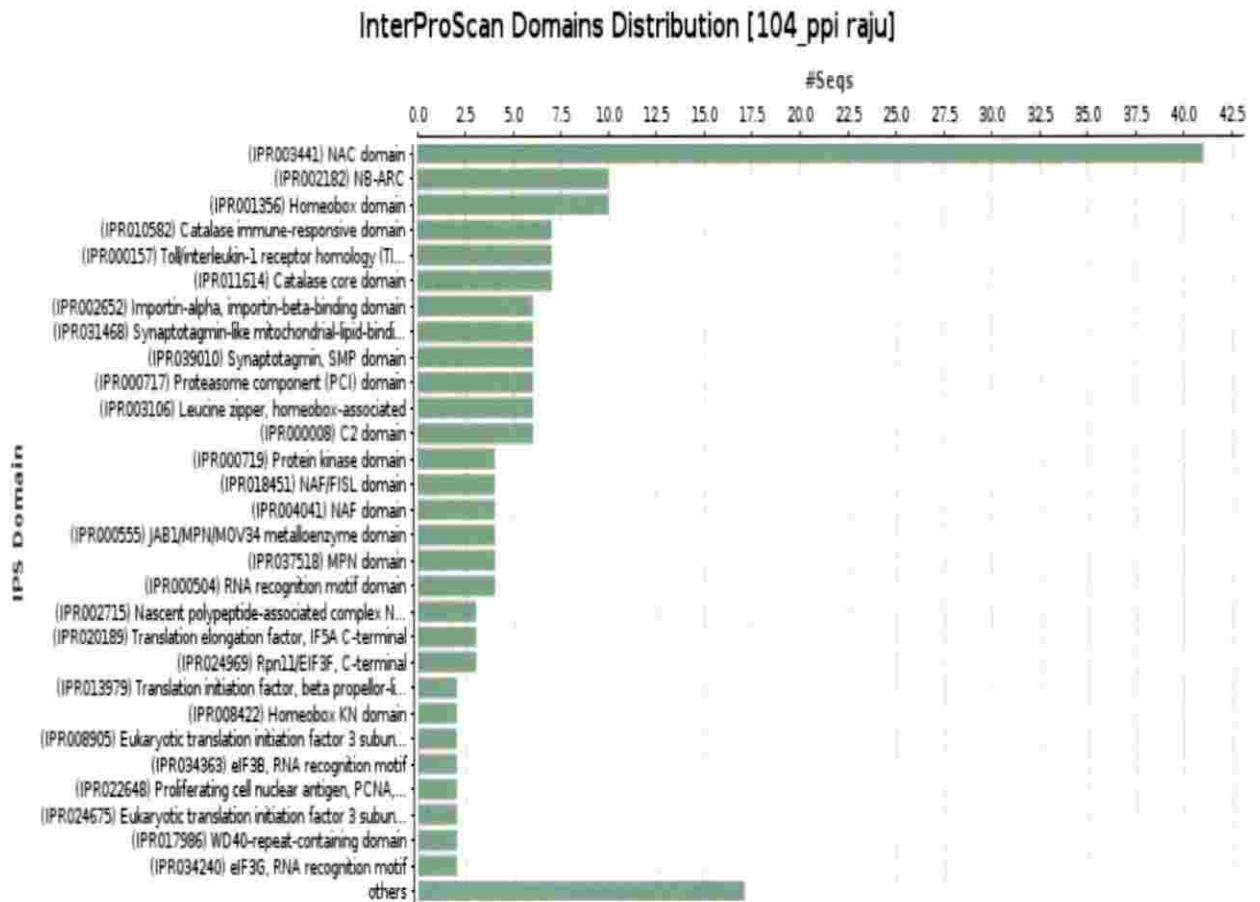


Figure 15. InterProScan domains distribution

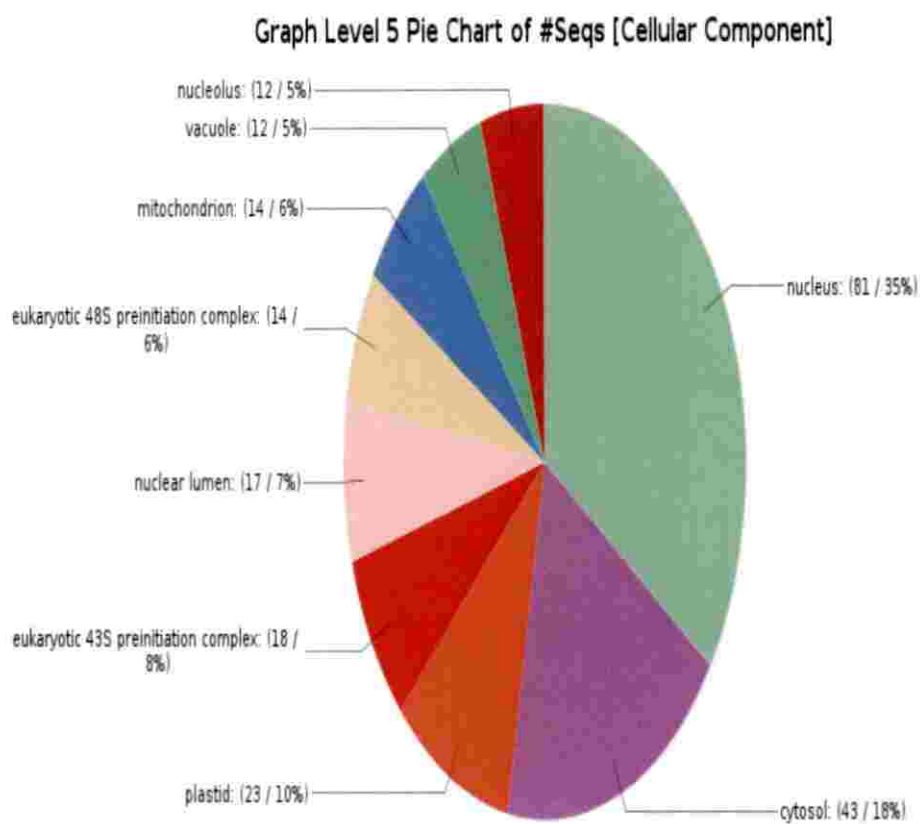


Figure 16. Cellular component of the predicted proteins in cassava

5.1 Subcellular localization of cassava proteins targeted by the CMV proteins

Pathogens suppress host immunity by directing a range of secreted proteins or effectors, to the cytoplasm of host cells. Once these effector proteins traversed the host plasma-membrane, are transported to many subcellular locations where they subvert the host immune system to enable pathogen growth and reproduction. The knowledge of cellular compartments of the cassava proteins targeted by the predicted CMV will be helpful in deciphering the mechanism of host-pathogen interactions. If the targeted cassava proteins are located in cellular compartments that are very relevant to the pathogen's infection or very likely to be involved in interactions with the pathogen, then the prediction result supports the host-pathogen predictions.

In this study, to have a clear understanding about the location of the interactions in host, the subcellular localization of the predicted cassava proteins were extracted using Localizer tool (<http://localizer.sciro.au/>). The subcellular locations of all predicted cassava proteins are listed in Table 6. It is found that 57.9% host proteins are localized in nucleus, 4.4% in chloroplast, and 0.9% in mitochondrion. It reveals that majority of the interaction occurs in nucleus, and chloroplast region.

Table 6. Identified subcellular locations of predicted proteins using Localizer.

('Y' represents yes and '-' represents null)

SI No	Identifier	Chloroplast	Mitochondria	Nucleus
1	A0A2C9U8S7_MANES	-	-	-
2	A0A2C9U1Q2_MANES	-	-	-
3	A0A2C9U4V6_MANES	-	-	-
4	A0A2C9UYQ8_MANES	-	-	Y (FKVK)
5	A0A2C9VKE1_MANES	-	-	Y (KKLELWRGILKKKGF R)
6	A0A2C9VX86_MANES	-	-	-
7	A0A2C9VVT7_MANES	-	-	-
8	A0A2C9VVU3_MANES	-	-	-
9	Q9SW99_MANES	-	-	-
10	A0A2C9WMD1_MANES	-	-	-
11	A0A2C9TZH3_MANES	-	-	-
12	A9YME8_MANES	-	-	-
13	A0A2C9WLH4_MANES	-	-	-
14	A0A2C9WD70_MANES	-	-	-
15	A0A2C9W3T7_MANES	Y (0.996 1-24)	Y (0.875 1-22)	Y (RRLARALKNGRRKTS, KKVHVATLERYRRT KRP,RRRHQNSAISSS SKKKKK,RRHQNSAISS SSSKKKKK,KRLKKV HVATLERYRRTK)
16	A0A2C9VUS4_MANES	-	-	Y (PGKKRRL)
17	A0A2C9WC92_MANES	-	-	Y (KRCCENLTEENRRLQ K)
18	A0A2C9VPI5_MANES	-	-	Y

				(PEKKRRL)
19	A0A2C9WNB8_MANES	-	-	-
20	A0A2C9VUV8_MANES	-	-	Y (PGKKRRL)
21	A0A2C9W164_MANES	-	-	Y (KRSR)
22	A0A2C9U0M9_MANES	-	-	Y (KKIKLLSMLDEVDRR YKE)
23	A0A2C9V8V6_MANES	-	-	Y (KKRKKD, KKQANKLT KFKRKETR, KRCCQTL TEENRRLQK)
24	A0A2C9W3M2_MANES	-	-	Y (KRKR, RKRK, RLKAK)
25	A0A2C9VFA0_MANES	-	-	Y (RKNKREDSLLKKRRE G, RRRREDNLVEIRKN KRE, RKKAYKTGVDA DEARRRRE, KKAYKTG VDADEARRRRED)
26	A0A2C9VFC8_MANES	-	-	Y (RKNRREESLQKKRRE G, RRRREDNMVEIRKN RRE, RRNRYKVAVDAE EGRRRRE)
27	A0A2C9VBQ4_MANES	-	-	Y (RKNKREDNLLKKRRE G, RRRREDNLVEIRKN KRE, RKKAYKTGVDA DEARRRRE, KKAYKTG VDADEARRRRED)
28	A0A2C9VBR2_MANES	-	-	Y (RKNRREESLQKKRRE G, RRRREDNMVEIRKN RRE, RRNRYKVAVDAE EGRRRRE)
29	A0A2C9VTE1_MANES	-	-	Y (RKNKREESLQKKRRE G, RRRREDNMVEIRKN KRE, RRNRYKVAVDAE EGRRRRE)
30	A0A251LCR3_MANES	-	-	Y (RKSKREESLQKKRRE

				G,RRRREDNMVEIRKS KRE,RRNRYKVAVDA DEGRRRRE)
31	A0A2C9VT85_MANES	-	-	Y (FKVK)
32	A0A2C9VW90_MANES	-	-	Y (KRPVGILSVKVL RAM KLKK)
33	A0A2C9V4J0_MANES	-	-	Y (KKKTKMIRK)
34	A0A2C9W0I7_MANES	-	-	Y (RKT KHIKK, KKPVGIL SVKVL RALKLKK)
35	A0A2C9W634_MANES	-	-	-
36	A0A2C9W0F0_MANES	-	-	Y (RKT KHIKK, KKPVGIL SVKVL RALKLKK)
37	A0A2C9WCA3_MANES	-	-	-
38	A0A2C9UNR1_MANES	-	-	Y (LGEV)
39	A0A2C9U746_MANES	-	-	Y (RRKRRK, KKIMVLYKS SKKGTK)
40	A0A2C9V1V9_MANES	-	-	Y (KKT MVFYKGKAPAG RKTKW)
41	A0A2C9VPE6_MANES	-	-	-
42	A0A2C9U5Z3_MANES	-	-	-
43	A0A2C9V824_MANES	-	-	-
44	A0A2C9VIE7_MANES	-	-	-
45	A0A2C9VU82_MANES	-	-	Y (KRKR, KISKNKKKASK KDEKAEPDSKKTRPNK KSRK)
46	A0A2C9VS27_MANES	-	-	Y (KRKR, KISKNKKKASK KDEKAEPDSKKTRPNK KSRK)
47	A0A2C9UAD5_MANES	-	-	Y (RKRRK, KKILVLYTNF

				GKNRKPEK)
48	A0A2C9VS73_MANES	-	-	Y (KRKR,KISKNKKKASK KDEKAEPDSKKTRPNK KSRK)
49	A0A2C9UQA8_MANES	-	-	Y (KRKR,KISKNKKKASK KDEKAEPDSKKTRPNK KSRK)
50	A0A2C9UP13_MANES	-	-	Y (KRKR,KISKNKKKASK KDEKAEPDSKKTRPNK KSRK)
51	A0A2C9VNM8_MANES	-	-	Y (RKRRK)
52	A0A251L6B6_MANES	-	-	Y (RKRRK,KKILVLYTNF GKNRKPEK)
53	A0A251L698_MANES	-	-	Y (RKRRK,KKILVLYTNF GKNRKPEK)
54	A0A2C9VTU5_MANES	Y (0.838 1-61)	-	Y (RPRR)
55	A0A2C9WM25_MANES	-	-	-
56	A0A2C9WN37_MANES	-	-	-
57	A0A2C9WKT2_MANES	-	-	-
58	A0A2C9W4N0_MANES	-	-	-
59	A0A2C9VRR0_MANES	-	-	-
60	A0A2C9W5R6_MANES	-	-	-
61	A0A2C9WC46_MANES	-	-	-
62	A0A251KRR0_MANES	-	-	Y (KPTGKPRKVKGIGTK KPIGTKRT)
63	A0A2C9WHR8_MANES	-	-	Y (KKSrk)
64	A0A2C9W4Z3_MANES	-	-	Y (KQSRSEKSRKAMLK LGMK)

65	A0A2C9W041_MANES	-	Y (0.99 1-21)	-
66	A0A2C9VRP8_MANES	-	-	Y (PKPS)
67	A0A2C9VUQ0_MANES	-	-	Y (KKRK)
68	A0A2C9VE39_MANES	-	-	Y (PSQKRNR)
69	A0A2C9VIE3_MANES	-	-	-
70	A0A2C9VSE5_MANES	-	-	-
71	A0A199UA28_MANES	-	-	-
72	A0A2C9UKB8_MANES	-	-	Y (RKRRK)
73	A0A2C9U024_MANES	-	-	-
74	A0A2C9UFZ4_MANES	-	-	Y (RKRRK, KKIMVLYKN TKKGSK)
75	A0A2C9URD4_MANES	-	-	-
76	A0A2C9W4Q2_MANES	-	-	-
77	A0A2C9WR64_MANES	-	-	Y (KRKR, RKRR)
78	A0A2C9U2C3_MANES	-	-	-
79	A0A251LEQ3_MANES	-	-	-
80	A0A251L5A8_MANES	-	-	-
81	A0A2C9WG65_MANES	-	-	Y (RKRR)
82	A0A199UAY8_MANES	-	-	-
83	A0A2C9UXZ5_MANES	-	-	Y (RKREAEKERARRDRL, KPPRPKFGPKWRFNQ HRPQLPQRRDEEVEAR KREAEKERARR)
84	A0A2C9W286_MANES	-	-	Y (TRKREAEKERARR, RK REAEKERARRDRL, RR DEEVATRKREAEKERA R)

85	A0A251K2X1_MANES	-	-	Y (KKMSSSNAKALNSMK QKLG)
86	A0A2C9VFH4_MANES	-	-	Y (KKMSSSNAKALNSMK QKLG)
87	A0A251K7S8_MANES	-	-	Y (PKPS,REAKRKK,KRLL ARKSIIKRKEE,KRKKI FYVRTEERLRKL,KRP EDLMLSYVSGEKGKD R,KKLQKLAKTMDYLE RAKRE,RLRKLHEEEE ARKHEEAERRRKEEAE RKAKLDEIAEKQRQRE RELEEKEK)
88	A0A2C9VFE3_MANES	-	-	Y (PKPS,KRLLARKSIIEK RKEE,KRPEDLMLSYV TGEKGKDR,KKLQKLA KTMDYLERAKRE,RKQ EREAKRKKIFYVRSEE ERLRKLHEEEEARKRE EAERRRKEEAERKAKL DEIAEKQRQRELEE KERLR)
89	A0A2C9W3E8_MANES	-	-	Y (KRTTYTGFELFRIKER)
90	A0A2C9UCF2_MANES	-	-	Y (KRLHEEEKLERQKLR, KRTTYTGFELFRIKER)
91	A0A2C9VKP1_MANES	-	-	Y (RKLAKARLSKKA)
92	A0A2C9V4Q7_MANES	-	-	Y (RKLAKARLSKRA)
93	A0A2C9UAB5_MANES	-	-	-
94	A0A2C9WK52_MANES	-	-	Y (KRRS)
95	A0A2C9VM01_MANES	-	-	Y (AEKEANSRKKKTGGKK K)
96	A0A2C9VMX9_MANES	Y (0.971 1-25)	-	-

97	A0A2C9VKV5_MANES	Y (0.971 1-25)	-	-
98	A0A2C9VMB7_MANES	Y (0.971 1-25)	-	-
99	A0A2C9UGV6_MANES	-	-	-
100	A0A2C9U7A7_MANES	-	-	-
101	A0A2C9WCR1_MANES	-	-	-
102	A0A2C9VV50_MANES	-	-	-
103	A0A2C9VP44_MANES	-	-	-
104	A0A2C9VMR2_MANES	-	-	-
105	A0A2C9U3K5_MANES	-	-	Y (KRSR,KKGVDQAEKE ERRRRRTEK)
106	A0A2C9U398_MANES	-	-	Y (KRSR,RSSKRIR,KRKI NTWTFNANFNVIKRR, RKINTWTFNANFNVIK RRL)
107	A0A2C9V0P3_MANES	-	-	Y (KRSR)
108	A0A2C9V0Q1_MANES	-	-	Y (KRSR)
109	A0A2C9V4R5_MANES	-	-	Y (KRRR,KVKK,RRDISTE EYMKLSKR)
110	A0A251LK88_MANES	-	-	Y (LGEV)
111	A0A251LK92_MANES	-	-	Y (LGEV,RDPKRMK)
112	A0A2C9WFM4_MANES	-	-	Y (KKTCLIDYLSKRV)
113	A0A251LKM5_MANES	-	-	Y (PFPKRLK)
114	A0A2C9UNM2_MANES	-	-	-

Functional annotation of predicted 144 proteins of cassava is shown in Appendices. Sequence name, description, SIM mean are listed. (Tags represents: Interpro [I], Blast [B] Mapping [M] and Annotation [A]).

Among the 114 predicted protein pairs, 10 proteins come under disease resistance protein family (TIR-NBS-LRR class) in cassava. The resistance proteins in cassava are listed in Table 7.

Table 7. Disease resistance proteins and its corresponding genes in cassava

Sl no.	Seq name	Phytozome protein id	UniProt Gene name	Length
1	A0A2C9U3K5	cassava4.1_000798m	MANES_18G144600	1029
2	A0A2C9U398	Unknown	MANES_18G108400	1187
3	A0A2C9V0P3	cassava4.1_032695m	MANES_11G119700	1135
4	A0A2C9V4R5	cassava4.1_023065m	MANES_10G087600	1239
5	A0A251LK88	cassava4.1_023606m	MANES_02G205900	1158
6	A0A251LK92	cassava4.1_023606m	MANES_02G205900	1284
7	A0A2C9WFM4	cassava4.1_000798m	MANES_02G199900	1133
8	A0A2C9UNM2	cassava4.1_033689m	MANES_03G013700	1100
9	A0A2C9V0Q1	cassava4.1_032695m	MANES_11G119700	967
10	A0A251LKM5	cassava4.1_028330m	MANES_02G199800	771

All the gene products of CMV were annotated. GO annotations of the CMV genome were obtained from QuickGO (<http://www.ebi.ac.uk/QuickGO>). GO of predicted interacting proteins of *Cassava Mosaic Virus* is shown below (Table 8).

Table 8. GO of predicted interacting proteins in *Cassava Mosaic Virus*

UniProt id	Gene name	Function	Reference
Q66284	<i>ORF 4</i>	Part of host cell cytoplasm, Involved in regulation of translation	GO_REF:0000037 GO_REF:0000039
Q08589	<i>AC2, AL2</i>	Enables in structural molecule activity, DNA binding and metal ion binding. Part of viral capsid, host cell cytoplasm, host cell nucleus. Involved in viral process.	GO_REF:0000002 GO_REF:0000037 GO_REF:0000039
O90282	<i>AV1</i>	Enables in structural molecule activity, DNA binding, metal ion binding. Part of viral capsid, host cell nucleus, virion.	GO_REF:0000002 GO_REF:0000038 GO_REF:0000040
H8WR53	<i>BC1</i>	Enables in DNA binding. Part of host cell membrane, integral component of membrane. Involved in transport of virus in host, cell to cell.	GO_REF:0000002 GO_REF:0000038
H8WR50	<i>AC1</i>	Involved in nucleic acid phosphodiester bond hydrolysis and metabolic process. Enables in structural molecule activity , DNA replication, endodeoxyribonuclease activity, producing 5'-phosphomonoesters , hydrolase activity metal ion binding, nucleotide binding catalytic activity , nucleotidyl transferase activity helicase activity, DNA binding transferase activity, ATP binding , nuclease activity and endonuclease activity. Part of host cell nucleus.	GO_REF:0000108 GO_REF:0000002 GO_REF:0000038 GO_REF:0000040
H8WR51	<i>AC4</i>	Protein A4. Enables in DNA binding and metal ion binding. Involved in viral process.	-
Q89703	<i>ORF 3</i>	Involved in RNA-dependent DNA biosynthetic process, nucleic acid phosphodiester bond hydrolysis, RNA phosphodiester bond hydrolysis, endonucleolytic proteolysis, DNA recombination and metabolic process. Enables in RNA-DNA hybrid ribonuclease activity, DNA binding, nucleotidyl transferase activity , DNA-directed DNA polymerase activity, peptidase activity, RNA-directed DNA polymerase activity, catalytic activity, transferase activity, metal ion binding , RNA binding, aspartic-type endopeptidase activity, hydrolase activity, endonuclease activity and nuclease activity.	GO_REF:0000108 GO_REF:0000002 GO_REF:0000003 GO_REF:0000037
H8WR48	<i>AC3</i>	Involved in viral process.	GO_REF:0000002 GO_REF:0000038

Q65975	ORF2	Enables in ATP binding.	GO_REF:0000002
H8WR46	AV2	Part of host cell cytoplasm and host cell perinuclear region of cytoplasm. Involved in negative regulation of gene silencing by RNA and viral process.	GO_REF:0000002 GO_REF:0000038

The interacting proteins of cassava in Cassava-CMV showed further interaction with predicted cassava protein, i.e., intraspecies (PPI) interaction. The first step in this study was to predict PPIs in cassava (inter-species interaction) and the second step was to predict PPIs between Cassava-CMV (Intra-species interaction). From the results obtained, it is found that the predicted cassava proteins in cassava-CMV interaction interact with the predicted proteins in cassava interactome. The results were obtained using STRING.

4.6 EXPERIMENTAL VALIDATION

The *in-silico* predicted proteins were validated using the designed primers against healthy and infected varieties of cassava.

4.6.1 Isolation of RNA

RNA isolation of 2 cassava leaf samples were done using CTAB method and were stored at -20°C.

4.6.2 Analysis of RNA

The RNA samples isolated using the CTAB method were analysed using 1.2% agarose gel electrophoresis (Plate 1). Distinct two bands were observed which shows no apparent RNA degradation.

4.6.3 Quantification of RNA

Quantification of RNA was done using NanoDrop® ND-100. The concentration of RNA (ng/μl), A260/230, A260/280, obtained are shown below (Table 9).

Table 9. Quantification of RNA

Sample	RNA yield (ng/ μ l)	A260/280	A260/230
H165 (healthy)	1894	2.1	2.05
H165 (infected)	1925	2.08	1.97



Plate 1: 1.2% EtBr stained agarose gel showing RNA of 2 Cassava leaf samples after electrophoresis (5 μ l RNA sample + 1 μ l 1 X loading dye).

Lane 1 & 2: H165 (healthy)

Lane 3 & 4: H165 (infected)

The relative gene expression of predicted proteins (catalase and Transcription Activator Protein) in healthy cassava variety (H165) and CMV infected cassava variety (H165) were studied using SYBR green PCR assay. *Cat2* gene that codes for catalase is selected in the study because catalase activity is high in infected leaf samples as compared to healthy leaf samples (Duraismy *et al.*, 2017). *AC2* gene is a viral protein that codes for Transcription Activator Protein (TrAP). It is predicted that the two proteins catalase and TrAP interacts with each other during a viral infection. Gene expression pattern of comparative C_t method showed the up-regulation of *AC2* gene in CMV infected leaf sample. Relative gene expression of *AC2* and *Cat2* in healthy and CMV infected cassava leaves are shown in Figure 17.

4.6.1 Designed Primer

Primer sets were designed for *Cat2* gene and *AC2* gene

- Cassava *Cat2* gene (catalase) : Product size-95bp

Forward Primer: 5'CAGCGTGTGTGCCATGCTAG3'

Reverse Primer: 5'CATGAATAACAGTGGAGAAACGGAC3'

- CMV *AC2* gene (TrAP: Transcription Activator Protein): Product size-95bp

Forward Primer: 5'CCCAAAGCCAACAGAGAGA3'

Reverse Primer: 5'CATCACCGAGTCCAACACAAT3

Reference gene: Actin (Product size-95bp)

Forward Primer: CCCAAAAGCCAACAGAGAGA

Reverse Primer: CATCACCGAGTCCAACACAAT

4.6.2 EXPRESSION STUDY OF PREDICTED PROTEINS IN CASSAVA

The predicted interacting proteins were present in healthy and susceptible (infected) variety were targeted using designed specific primers and the SYBR green PCR assay was used for studying gene expression. The relative gene expression of healthy and susceptible varieties is studied using $2^{-\Delta\Delta C_t}$ method.

Actin (primers ACT F and ACT R) was used as the reference gene for the expression study.

The standard fluorescent amplification representing exponential growth of PCR products was observed in each cycle, yielding threshold cycle (C_t) values that ranged from 15-28 for the target and reference (ACT F and ACT R) primers. The C_t (Cycle threshold) value is given in the logarithmic scale and inversely proportional to the quantity of cDNA. Thus the highly expressed gene has low ΔC_t values and low expressed gene have high ΔC_t values. The fold change ($-\Delta\Delta C_t$) can be calculated by comparing the normalized expression (ΔC_t) of the two conditions. The fold change, viz. the expression ratio, indicated the up regulation and down regulation of the gene when it was positive and negative respectively.

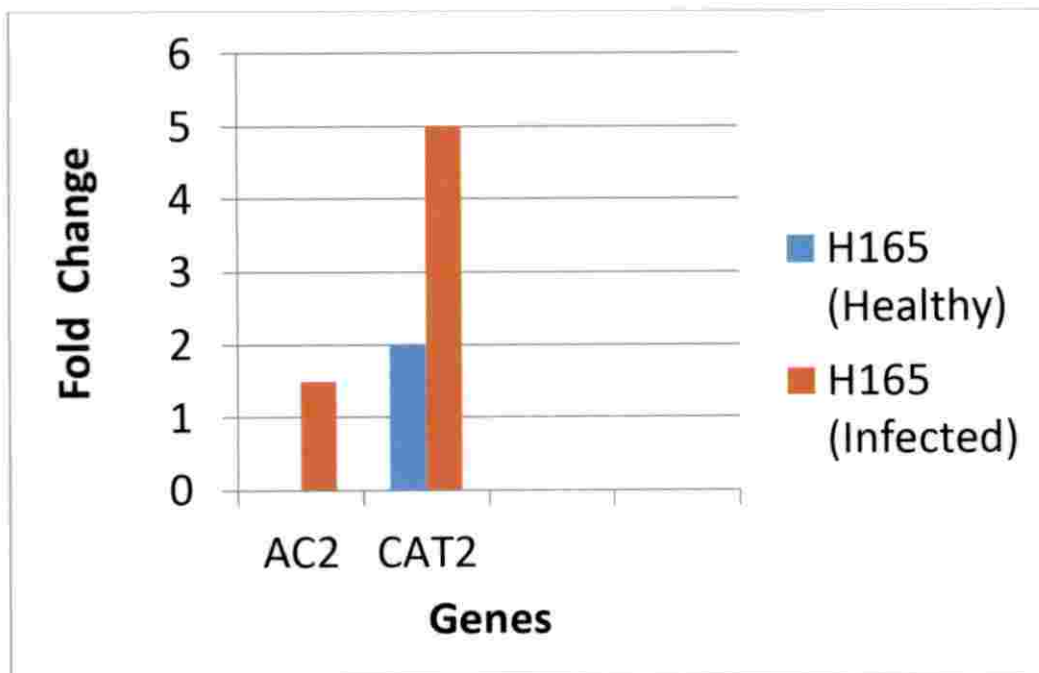


Figure 17. Relative gene expression of *AC2* and *CAT2* in healthy and CMV infected cassava leaves (Variety: H165).

DISCUSSION

5. DISCUSSION

The study entitled “Modeling of Cassava-Cassava Mosaic Virus interaction with computational biology and bioinformatics approach” was conducted to predict interacting pair of proteins between cassava and *Cassava Mosaic Virus* (CMV) based on interolog method using genomic data of template plant. The study also includes confirmation of the predicted interacting pairs using prediction tool, PPI network construction and functional annotation of the predicted protein for better understanding of pathogenesis mechanism of the crop. The results of this study presented in chapter 4 are discussed here.

Cassava Mosaic Virus (CMV) is a ssDNA virus causing economically important disease in *Manihot esculenta* thereby leading to severe agricultural losses in Asian and African countries. 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 (FAOSTAT, 2017). Similarly, the appearance of Cassava Mosaic Disease (major pathogen involved is CMV) seems to significantly constrain its productivity. Viral-host protein-protein interaction plays a vital role in pathogenesis, since it defines viral infection of the host and regulation of the host proteins. PPIs are essential process in all living cells and play a crucial role in the infection process, and initiating a defence response. In this context, understanding the PPI network (interactome) between plant proteins and pathogen proteins is a critical step for studying the molecular basis of pathogenesis (Pinzón *et al.*, 2010 and Kim *et al.*, 2008). In particular, computational approaches ameliorate the study of host-pathogen protein interactions in a genome-wide range.

Many computational methods have been developed to predict PPIs, but most of them are intended for PPIs within same species rather than for PPIs across different species. Methods for predicting intra-species PPIs do not distinguish interactions between proteins of the same species from those of different species, and thus are not appropriate for predicting inter-species PPIs. Motivated by a

recent increase in data of virus-host PPIs, a few computational methods have been developed to predict virus-host PPIs using machine learning methods. Zhou *et al.* (2018) developed a prediction tool (VirusHostPPI) of virus-host PPIs, which is applicable to new viruses and hosts. The predicted PPIs using interolog based method are confirmed by the prediction tool (VirusHostPPI), which identifies whether a protein pair interacts or not (<http://bclab.inha.ac.kr/VirusHostPPI>). The prediction tool works on the principle of SVM based approach.

In this study, a systematic attempt has been made to predict cassava-CMV PPIs by interolog-based method. From the proteomic datasets used for the study, 351 cassava proteins and 11 CMV proteins were predicted to interact by a simple and effective method: interolog based approach. After filtering of the predicted protein pairs using VirusHostPPI (prediction tool), 114 cassava proteins were found to be interacting with 10 CMV proteins. The reported results are coherent with the previous studies in which it is demonstrated that a few pathogen proteins involved in interaction with the host interactome (Kim *et al.*, 2008). Li *et al.* (2012) predicted protein-protein interactions between *Ralstonia solanacearum* and *Arabidopsis thaliana*. They predicted 3,074 potential PPIs between 119 *R. solanacearum* and 1,442 *A. thaliana* proteins. Sahu *et al.*, 2014 used two different methods for the prediction of PPI (Interolog based method and domain based method) between *Arabidopsi thaliana* and *Psuedomonas syringae* pathovar tomato strain DC3000 (PstDC3000). They reported that interolog-based method predicted nearly 0.79 Million PPIs involving around 7700 *Arabidopsis* and 1068 *Pseudomonas* proteins in the full genome while the domain-based method predicted 85650 PPIs comprising 11432 *Arabidopsis* and 887 *Pseudomonas* proteins.

The predicted cassava proteins in Cassava-CMV interaction were combined for functional annotation using Blast2GO. Effective annotation obtained from Blast2GO could provide several valuable data regarding the

identified interacting proteins. Among the total (114) proteins identified, 113 proteins showed blast hit (with *Arabidopsis thaliana*), interProScan results, mapping and annotation. InterProScan result showed that majority of proteins comes under NAC containing domain protein superfamily. NAC TFs are one of one of the largest families of transcription factors (TFs) in plants and they play vital roles in regulating plant growth and development processes including abiotic stress responses. Hu *et al.* (2015) reported 96 NAC genes in cassava. In their study, 96 predicted NAC proteins ranged from 82 to 656 amino acid residues with an average of 342 amino acid. They also studied the evolutionary relationships between cassava NAC proteins and known NACs from *Arabidopsis*.

Subcellular locations of the predicted proteins were found using localizer. It is found that 57.9% host proteins are localized in nucleus, 4.4% in chloroplast, and 0.9% in mitochondrion. It reveals that major of the interactions occur in nucleus, and chloroplast region. Also the localizations for a large number of proteins are still unknown which need a special attention for experimental characterization.

Functional annotation revealed the presence of 10 disease resistance proteins in the predicted Cassava-CMV interaction proteins. In 2015, Lozano *et al.* identified 228 NBS-LRR type genes and 99 partial NBS genes among the 30,666 annotated protein-coding genes. They reported that these represent almost 1% of the total predicted genes and show high sequence similarity to proteins from other plant species.

Understanding the Protein-Protein Interaction (PPI) network (i.e., interactome) between plant proteins and pathogen proteins is a critical step for studying the molecular basis of pathogenesis (Pinzon *et al.*, 2010; He *et al.*, 2008; Kim *et al.*, 2008). However, it is still a challenging task to identify the plant proteins targeted by a pathogen protein through existing experimental techniques.

Currently, only a few pairs of such interactions have been identified, which is far from being enough to systematically decipher the molecular mechanism of pathogenicity. Due to internal limitations of the computational methods, the predicted data may still suffer from two drawbacks. First, the predicted PPI network is still far from complete. Second, the predicted data may inevitably contain a lot of false positives. To quantitatively assess the reliability of the predicted PPIs, experimentally determined PPI data are required. Even so, the predicted PPI data have allowed us to catch a glimpse of the overall picture of the PPI network between CMV and cassava (*Manihot esculenta*). We hope that the current work can shed light for further research into the molecular pathogenesis of CMV. For instance, the predicted data may inspire a path to the discovery of new anti-viral drug targets.

It has been established that a pathogen mutates its genes extensively to infect a host, whereas a plant defends the attacks by expanding its gene families (Stahl and Bishop, 2000). Therefore, to some extent, the ratio of proteins involved in the predicted PPI network may reflect the plant–pathogen arms race at the molecular level.

SUMMARY

6. SUMMARY

The study entitled “Modeling of Cassava-Cassava Mosaic Virus interaction with computational biology and bioinformatics approach” was carried out at the Section of Extension and Social Sciences, ICAR-Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram during 2018-2019. The objectives of the study were to predict interacting pairs of proteins between Cassava and CMV, construction of Protein-Protein Interaction Network (PPIN) and validation of predicted protein pairs.

The study had mainly three objectives, PPI prediction between cassava and Cassava Mosaic Virus, predicted PPI network construction and validation of the predicted pairs. PPI prediction was done using interolog-based method and the template plant used is *Arabidopsis thaliana*. The preliminary datasets of PPIs for the prediction of cassava-CMV interaction were obtained mainly from three databases (STRING Viruses, APID, HPIDB). A total of 351 PPIs between 351 proteins in cassava and 11 proteins in CMV were predicted. These proteins were filtered using VirusHostPPI prediction tool. After filtering 114 PPIs between 114 cassava proteins and 10 CMV proteins were obtained. Using Functional annotation tools, the predicted proteins were functionally annotated. Predicted cassava proteins were annotated using Blast2Go and CMV proteins were annotated using QuickGO. The results showed the presence of 10 disease resistance proteins in predicted cassava proteins. These disease resistance proteins were predicted to interact with *ACI* gene of CMV which codes for replication associated proteins in CMV. Moreover, InterProScan results showed that majority of the proteins comes under NAC containing domain superfamily. NAC TFs are one of the largest families of transcription factors (TFs) in plants and they play vital roles in regulating plant growth and development processes including abiotic stress responses.

From the predicted PPI pair, one pair (*Cat2* gene of cassava and *AC2* gene of CMV) of interacting proteins of cassava and CMV interaction is validated using q-PCR. Primers were designed for both the proteins. These primers were validated using a healthy and CMV infected varieties.

6.1 SCOPE OF FUTURE WORK

As the resources were limited, only one predicted PPI pair was validated for differentiating expression of genes in healthy and infected cassava varieties. Further study can be done for the identification of interaction between predicted Cassava-CMV proteins and intraspecies PPI in cassava.

REFERENCES

7. REFERENCES

- Ako-Adjei, D., Fu, W., Wallin, C., Katz, K.S., Song, G., Darji, D., Brister, J.R., Ptak, R.G. and Pruitt, K.D. 2014. HIV-1, human interaction database: current status and new features. *Nucleic acids res.* 43(D1): 566-570.
- Alabi, O.J., Kumar, P.L. and Naidu, R.A. 2011. Cassava mosaic disease: a curse to food security in subSaharan Africa.
- Albert, R. 2005. Scale-free networks in cell biology. *J. of cell sci.* 118(21): 4947-4957.
- Almagro, L., Gómez Ros, L.V., Belchi-Navarro, S., Bru, R., Ros Barceló, A. and Pedreno, M.A. 2008. Class III peroxidases in plant defense reactions. *J. of Exp. Bot.* 60(2): 377-390.
- Alonso-López, D., Campos-Laborie, F.J., Gutiérrez, M.A., Lambourne, L., Calderwood, M.A., Vidal, M. and De Las Rivas, J. 2019. APID database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database*, 2019.
- Ammari, M.G., Gresham, C.R., McCarthy, F.M. and Nanduri, B. 2016. HPIDB 2.0: a curated database for host–pathogen interactions. *Database*, 2016.
- Arena, G.D., Ramos-González, P.L., Nunes, M.A., Jesus, C.C., Calegario, R.F., Kitajima, E.W., Novelli, V.M. and Freitas-Astúa, J. 2017. Arabidopsis thaliana as a model host for Brevipalpus mite-transmitted viruses. *Scientia Agricola.* 74(1): 85-89.

- Atiri, G.I., Ogbe, F.O., Dixon, A.G.O., Winter, S. and Ariyo, O., 2004. Status of cassava mosaic virus diseases and cassava begomoviruses in sub-Saharan Africa. *J. of Sust Agric.* 24(3): 5-35.
- Baldi, P., Brunak, S. and Bach, F. 2001. *Bioinformatics: the machine learning approach*. MIT press.
- Barabasi, A.L. and Oltvai, Z.N. 2004. Network biology: understanding the cell's functional organization. *Nat. rev. genet.* 5(2): 101.
- Bent, A.F. 1996. Plant disease resistance genes: Function meets structure. *The Plant Cell* 8(10): 1757p.
- Bisaro, D.M. 2006. Silencing suppression by geminivirus proteins. *Virology* 344(1): 158-168.
- Boller, T. and He, S.Y. 2009. Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Sci.* 324(5928): 742-744.
- 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. biotechnolo.* 34(5): 562p.
- Calderone, A., Licata, L. and Cesareni, G. 2014. VirusMentha: a new resource for virus-host protein interactions. *Nucleic acids res.* 43(D1): 588-592.

- Cassava plays a leading role in food security in India, especially in the major growing states of Tamil Nadu and Kerala. *Annual Report*. FAO(2018).
- Ceballos, H., Okogbenin, E., Pérez, J.C., López-Valle, L.A.B. and Debouck, D. 2010. Cassava. In *Root and tuber crops* pp. 53-96
- Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A. and Stark, C., 2017. The BioGRID interaction database: 2017 update. *Nucleic acids res.* 45(D1): 369-379.
- Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinforma.* 21(18): 3674-3676.
- Cui, G., Fang, C. and Han, K. 2012, December. Prediction of protein-protein interactions between viruses and human by an SVM model. In *BMC bioinforma.* (Vol. 13, No. 7, p. S5). BioMed Central.
- Davis, F.P., Barkan, D.T., Eswar, N., McKerrow, J.H. and Sali, A. 2007. Host-pathogen protein interactions predicted by comparative modeling. *Protein Sci.* 16(12): 2585-2596.
- Davis, F.P., Braberg, H., Shen, M.Y., Pieper, U., Sali, A. and Madhusudhan, M.S. 2006. Protein complex compositions predicted by structural similarity. *Nucleic acids res.* 34(10): 2943-2952.
- Dubern, J. 1994. Transmission of African cassava mosaic geminivirus by the whitefly (*Bemisia tabaci*). *Trop. Sci.* 34(1): 82-91.

- Duraisamy, R., Arumugam, C., Natesan, S., Muthurajan, R., Gandhi, K., Lakshmanan, P., Janavi, G.J., Karuppusamy, N. and Chokkappan, M. 2017. Host-Pathogen Interaction of Cassava (*Manihot esculenta* Crantz) and Cassava Mosaic Viruses (ICMV and SLCMV). *Int. J. Curr. Microbiol. App. Sci*, 6(7): 1305-1317.
- Dyer, M.D., Murali, T.M. and Sobral, B.W. 2007. Computational prediction of host-pathogen protein-protein interactions. *Bioinforma*. 23(13): 159-166.
- Edison, S., 2000, February. Present situation and future potential of cassava in India. In *Cassava's Potential in Asia in the 21st Century: Present Situation and Future Research and Development Needs. Proc. 6th Regional Workshop, held in Ho Chi Minh city, Vietnam* (pp. 61-70).
- Ellis, J., Dodds, P. and Pryor, T. 2000. The generation of plant disease resistance gene specificities. *Trends in plant sci*. 5(9): 373-379.
- FAOSTAT 2009. FAOSTAT. Available at <http://faostat.fao.org> (accessed 26 May 2009; verified 24 May 2011). Food and Agriculture Organization (FAO) of the United Nations, Rome, Italy.
- Finn, R.D., Miller, B.L., Clements, J. and Bateman, A. 2013. iPFam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic acids res*. 42(D1): 364-373.
- Flor, H.H., 1971. Current status of the gene-for-gene concept. *Ann. Rev. of phytopathol*. 9(1): 275-296.
- Food Safety Network. (2014). Cassava Nutritional Network. 1-866-50-FSNET: University of Guelph, March 14, 2005; 2 p.

- Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J. and Conesa, A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids res.* 36(10): 3420-3435.
- Guirimand, T., Delmotte, S. and Navratil, V. 2014. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic acids res.* 43(D1): 583-587.
- Hammond-Kosack, K.E., 2000. Responses to Plant Pathogens In “Biochemistry and Molecular Biology of Plants” p 1102-1156 Ed BB Buchanan, W Gruissem and RL Jones. *7 Am. Soc. of Plant Physio.* P136.
- Han, D.S., Kim, H.S., Jang, W.H., Lee, S.D. and Suh, J.K. 2004. PreSPI: a domain combination based prediction system for protein–protein interaction. *Nucleic acids res.* 32(21): 6312-6320.
- Hauck, P., Thilmony, R. and He, S.Y. 2003. A *Pseudomonas syringae* type III effector suppresses cell wall-based extracellular defense in susceptible *Arabidopsis* plants. *Proc. of the Natl. Acad. of Sci.* 100(14): 8577-8582.
- He, F., Zhang, Y., Chen, H., Zhang, Z. and Peng, Y.L. 2008. The prediction of protein-protein interaction networks in rice blast fungus. *BMC genomics*, 9(1): 519.
- Hu, W., Wei, Y., Xia, Z., Yan, Y., Hou, X., Zou, M., Lu, C., Wang, W. and Peng, M. 2015. Genome-wide identification and expression analysis of the NAC transcription factor family in cassava. *PLoS One.* 10(8): p.e0136993.

- Huntley, R.P., Binns, D., Dimmer, E., Barrell, D., O'Donovan, C. and Apweiler, R. 2009. QuickGO: a user tutorial for the web-based Gene Ontology browser. *Database*, 2009.
- Jones, J.D., Vance, R.E. and Dangl, J.L. 2016. Intracellular innate immune surveillance devices in plants and animals. *Sci.* 354(6316): p.aaf6395.
- Kim, W.K., Park, J. and Suh, J.K. 2002. Database of interacting proteins large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. In *Genome Inform* (13): 42-50.
- Kim, J.G., Park, D., Kim, B.C., Cho, S.W., Kim, Y.T., Park, Y.J., Cho, H.J., Park, H., Kim, K.B., Yoon, K.O. and Park, S.J. 2008. Predicting the interactome of *Xanthomonas oryzae* pathovar *oryzae* for target selection and DB service. *BMC bioinforma.* 9(1): 41p.
- Kim, W.K., Kim, K., Lee, E., Marcotte, E.M., Kim, H. and Suh, J. 2007. Identification of disease specific protein interactions between the gastric cancer causing pathogen, *H. pylori*, and Human Hosts using protein network modeling and gene chip analysis. *Gastric Cancer*, 1: 179-187.
- Kitano, H. 2002. Systems biology: a brief overview. *Sci.* 295(5560): 1662-1664.
- Korkin, D., Thieu, T., Joshi, S. and Warren, S. 2011. Mining hostpathogen interactions. *Syst. and Computational Biol.–Mol. and Cell. Exp. Sys.* pp.163-184.

- Krishnadev, O. and Srinivasan, N. 2008. A data integration approach to predict host-pathogen protein-protein interactions: application to recognize protein interactions between human and a malarial parasite. *In silico Biol.* 8(3, 4): 235-250.
- Krishnadev, O. and Srinivasan, N. 2011. Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int. j. of biol macromolecules.* 48(4): 613-619.
- Kshirsagar, M., Carbonell, J. and Klein-Seetharaman, J. 2013. Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks. In *NIPS Workshop on Machine Learning for Computational Biol.* (Vol. 2012).
- Lee, S.A., Chan, C.H., Tsai, C.H., Lai, J.M., Wang, F.S., Kao, C.Y. and Huang, C.Y.F. 2008. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC bioinforma.* 9(12): S11.
- Li, Z.G., He, F., Zhang, Z. and Peng, Y.L. 2012. Prediction of protein-protein interactions between *Ralstonia solanacearum* and *Arabidopsisthaliana*. *Amino Acids*, 42(6): 2363-2371.
- 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.
- Macfadyen, S., Paull, C., Boykin, L.M., De Barro, P., Maruthi, M.N., Otim, M., Kalyebi, A., Vassão, D.G., Sseruwagi, P., Tay, W.T. and Delatte, H. 2018. Cassava whitefly, *Bemisia tabaci* (Gennadius)(Hemiptera:

Aleyrodidae) in East African farming landscapes: a review of the factors determining abundance. *Bulletin of Entomol Res.* 108(5): 565-582.

Mei, S. 2013. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. *PLoS One*, 8(11): 79606.

Meng, X. and Zhang, S. 2013. MAPK cascades in plant disease resistance signaling. *Annu. Rev. of phytopathol.* 51: 245-266.

Mittal, D., Borah, B.K. and Dasgupta, I. 2008. Agroinfection of cloned *Sri Lankan cassava mosaic virus* DNA to *Arabidopsis thaliana*, *Nicotiana tabacum* and cassava. *Arch. of viro.* 153(11): 2149-2155.

Monaghan, J. and Zipfel, C. 2012. Plant pattern recognition receptor complexes at the plasma membrane. *Curr. opinion in plant boil.* 15(4): 349-357.

Mukhopadhyay, A. and Maulik, U. 2014. Network-based study reveals potential infection pathways of hepatitis-C leading to various diseases. *PloS one*, 9(4): p.94029.

Mukhopadhyay, A., Maulik, U. and Bandyopadhyay, S. 2012. A novel biclustering approach to association rule mining for predicting HIV-1–human protein interactions. *PLoS One*, 7(4): p.e32289.

Nassar, N. and Ortiz, R. 2010. Breeding cassava to feed the poor. *Sci. American*, 302(5): 78-85.

- Nourani, E., Khunjush, F. and Durmuş, S. 2015. Computational approaches for prediction of pathogen-host protein-protein interactions. *Frontiers in microbial.* 6: 94p.
- Pagan, I., Fraile, A., Fernandez-Fueyo, E., Montes, N., Alonso-Blanco, C. and García-Arenal, F. 2010. Arabidopsis thaliana as a model for the study of plant-virus co-evolution. *Philos. Trans. of the R. Soc. B: Biolo. Sci.* 365(1548): 1983-1995.
- Pagel, P., Wong, P. and Frishman, D. 2004. A domain interaction map based on phylogenetic profiling. *J. of mol. Boil.* 344(5): 1331-1346.
- Pinzón, A., Rodriguez-R, L.M., González, A., Bernal, A. and Restrepo, S. 2010. Targeted metabolic reconstruction: a novel approach for the characterization of plant-pathogen interactions. *Briefings in bioinforma.* 12(2): 151-162.
- Pitre, S., Alamgir, M., Green, J.R., Dumontier, M., Dehne, F. and Golshani, A. 2008. Computational methods for predicting protein-protein interactions. In *Protein-Protein Interaction* (pp. 247-267). Springer, Berlin, Heidelberg.
- Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. 2005. InterProScan: protein domains identifier. *Nucleic acids res.* 33(suppl_2): 116-120.
- Sahu, S.S., Weirick, T. and Kaundal, R. 2014, December. Predicting genome-scale Arabidopsis-Pseudomonas syringae interactome using domain and interolog-based approaches. In *BMC bioinforma.* (Vol. 15, No. 11, p. S13). BioMed Central.

- Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. 2004. The database of interacting proteins: 2004 update. *Nucleic acids res.* 32(suppl_1): 449-451.
- Schleker, S., Garcia-Garcia, J., Klein-Seetharaman, J. and Oliva, B. 2012. Prediction and Comparison of Salmonella–Human and Salmonella–Arabidopsis Interactomes. *Chem. & biodivers.* 9(5): 991-1018.
- Schleker, S., Kshirsagar, M. and Klein-Seetharaman, J. 2015. Comparing human–Salmonella with plant–Salmonella protein–protein interaction predictions. *Frontiers in microbial.* 6: 45p.
- Schulze, S., Henkel, S.G., Driesch, D., Guthke, R. and Linde, J. 2015. Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. *Frontiers in microbial.* 6: 65p.
- Segura-Cabrera, A., García-Pérez, C.A., Guo, X. and Rodríguez-Pérez, M.A. 2013. A viral-human interactome based on structural motif-domain interactions captures the human infectome. *PloS one*, 8(8): p.e71526.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome res.* 13(11): 2498-2504.
- Shoemaker, B.A. and Panchenko, A.R. 2007. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS computational boil.* 3(4): 43.

- Singh, G. and Singh, V. 2019. Construction and analysis of an interologous protein–protein interaction network of *Camellia sinensis* leaf (TeaLIPIN) from RNA–Seq data sets. *Plant cell rep.* :1-14.
- Stahl, E.A. and Bishop, J.G. 2000. Plant–pathogen arms races at the molecular level. *Curr. opinion in plant boil.* 3(4): 299-304.
- Stebbins, C.E. 2005. Structural microbiology at the pathogen–host interface. *Cellular microbial.* 7(9): 1227-1236.
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P. and Kuhn, M., 2014. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids res.* 43(D1): D447-D452.
- Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P. and Jensen, L.J. 2016. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids res.*: 937.
- Thanasomboon, R., Kalapanulak, S., Netrphan, S. and Saithong, T. 2017. Prediction of cassava protein interactome based on interolog method. *Sci. rep.* 7(1): 17206.
- Tyagi, N., Krishnadev, O. and Srinivasan, N. 2009. Prediction of protein–protein interactions between *Helicobacter pylori* and a human host. *Mol. bioSys.* 5(12): 1630-1635.

- Van Loon, L.C. and Van Strien, E.A. 1999. The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins. *Physiolo. and molecular plant pathol.* 55(2): 85-97.
- Van Loon, L.C., Rep, M. and Pieterse, C.M. 2006. Significance of inducible defense-related proteins in infected plants. *Annu. Rev. Phytopathol.*, 44: 135-162.
- Vásquez, A.X., Soto Sedano, J.C. and López Carrascal, C.E. 2018. Unraveling the molecules hidden in the gray shadows of quantitative disease resistance to pathogens. *Acta Biológica Colombiana*, 23(1): 5-16.
- Wang, X., Goregaoker, S.P. and Culver, J.N. 2009. Interaction of the Tobacco mosaic virus replicase protein with a NAC domain transcription factor is associated with the suppression of systemic host defenses. *J. of virol.* 83(19): 9720-9730.
- Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R. and Machi, D. 2013. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic acids res.* 42(D1): D581-D591.
- Westermann, A.J., Gorski, S.A. and Vogel, J., 2012. Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* 10(9): 618.
- Wojcik, J. and Schächter, V. 2001. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinforma.* 17(suppl_1): S296-S305.

- Zhou, H., Jin, J., Zhang, H., Yi, B., Wozniak, M. and Wong, L. 2012, December. IntPath--an integrated pathway gene relationship database for model organisms and important pathogens. In *BMC sys. biol.* 6 (2): S2p. BioMed Central.
- Zhou, H., Gao, S., Nguyen, N.N., Fan, M., Jin, J., Liu, B., Zhao, L., Xiong, G., Tan, M., Li, S. and Wong, L., 2014. Stringent homology-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. *Biol. direct*, 9(1): 5.
- Zhou, X., Park, B., Choi, D. and Han, K. 2018. A generalized approach to predicting protein-protein interactions between virus and host. *BMC genomics*, 19(6): 165p.
- Zoraghi, R. and Reiner, N.E. 2013. Protein interaction networks as starting points to identify novel antimicrobial drug targets. *Current opinion in microbiol.* 16(5): 566-572.

APPENDICES

8. APPENDIX I

Appendix I. Functional annotation result of the predicted PPIs in cassava
(Tags represents: Interpro [I], Blast [B] Mapping [M] and Annotation [A]).

Sl no.	Seq name	Tags	Description	Length	Sim mean
1	tr A0A251K7S8 A0A251K7S8_M ANES	I,B,M,A	Eukaryotic translation initiation factor 3A	1013	85.79
2	tr A0A2C9VFE3 A0A2C9VFE3_ MANES	I,B,M,A	Eukaryotic translation initiation factor 3A	1003	92.12
3	tr A0A2C9WHR8 A0A2C9WHR8_ MANES	I,B,M,A	Nascent polypeptide-associated complex (NAC), alpha subunit family protein	221	89.63
4	tr A0A2C9W4Z3 A0A2C9W4Z3_ MANES	I,B,M,A	Nascent polypeptide-associated complex (NAC), alpha subunit family protein	206	87.29
5	tr A0A2C9W041 A0A2C9W041_ MANES	I,B,M,A	Nascent polypeptide-associated complex (NAC), alpha subunit family protein	268	89.93
6	tr A0A199UAY8 A0A199UAY8_ MANES	I,B,M,A	Plasma-membrane associated cation-binding protein 1	205	86.49
7	tr A0A2C9W3E8 A0A2C9W3E8_ MANES	I,B,M,A	Translation initiation factor 3B1	720	68.1
8	tr A0A2C9UCF2 A0A2C9UCF2_ MANES	I,B,M,A	Translation initiation factor 3B1	720	68.41
9	tr A0A2C9VKP1 A0A2C9VKP1_ MANES	I,B,M,A	RNA-binding (RRM/RBD/RNP motifs) family protein	294	60.12
10	tr A0A2C9V4Q7 A0A2C9V4Q7_	I,B,M,A	RNA-binding (RRM/RBD/RNP motifs) family protein	295	59.63

	MANES				
11	tr A0A2C9WG65 A0A2C9WG65_ MANES	I,B,M,A	Argonaute family protein	1070	64.07
12	tr A0A2C9W3T7 A0A2C9W3T7_ MANES	I,B,M,A	Overexpressor of cationic peroxidase 3	353	80.46
13	tr A0A2C9W3M2 A0A2C9W3M2_ MANES	I, No- Blast	---NA---	951	
14	tr A0A2C9WC92 A0A2C9WC92_ MANES	I,B,M,A	Homeobox-leucine zipper protein family	288	68.04
15	tr A0A2C9V8V6 A0A2C9V8V6_ MANES	I,B,M,A	Homeobox-leucine zipper protein family	472	63.03
16	tr A0A2C9VUS4 A0A2C9VUS4_ MANES	I,B,M,A	Homeobox protein 5	303	80.98
17	tr A0A2C9VPI5 A0A2C9VPI5_M ANES	I,B,M,A	Homeobox protein 5	319	71.42
18	tr A0A2C9WNB8 A0A2C9WNB8_ MANES	I,B,M,A	Homeobox-leucine zipper protein family	291	67.66
19	tr A0A2C9VUV8 A0A2C9VUV8_ MANES	I,B,M,A	Homeobox protein 5	296	79.59
20	tr A0A2C9WLH4 A0A2C9WLH4_ MANES	I,B,M,A	Proliferating cell nuclear antigen 2	266	97.35
21	tr A0A2C9WD70 A0A2C9WD70_ MANES	I,B,M,A	Proliferating cell nuclear antigen 2	264	97.35
22	tr A0A2C9W164 A0A2C9W164_ MANES	I,B,M,A	BEL1-like homeodomain 1	806	58.28

23	tr A0A2C9U0M9 A0A2C9U0M9_ MANES	I,B,M,A	BEL1-like homeodomain 1	665	62.89
24	tr A0A2C9WCA3 A0A2C9WCA3_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	288	61.35
25	tr A0A2C9UNR1 A0A2C9UNR1_ MANES	I,B,M,A	NAC transcription factor-like 9	623	67.59
26	tr A0A2C9U746 A0A2C9U746_M ANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	450	50.63
27	tr A0A2C9V1V9 A0A2C9V1V9_ MANES	I,B,M,A	NAC domain containing protein 35	244	62.9
28	tr A0A2C9VPE6 A0A2C9VPE6_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	400	72.55
29	tr A0A2C9U5Z3 A0A2C9U5Z3_M ANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	319	71.09
30	tr A0A2C9V824 A0A2C9V824_M ANES	I,B,M,A	NAC domain containing protein 35	286	61.06
31	tr A0A2C9VIE7 A0A2C9VIE7_M ANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	440	58.99
32	tr A0A2C9VU82 A0A2C9VU82_ MANES	I,B,M,A	NAC domain containing protein 82	494	57.02
33	tr A0A2C9VS27 A0A2C9VS27_M ANES	I,B,M,A	NAC domain containing protein 82	486	56.7
34	tr A0A2C9UAD5 A0A2C9UAD5_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	455	54.38
35	tr A0A2C9VS73 A0A2C9VS73_M	I,B,M,A	NAC domain containing protein	484	57.06

	ANES		82		
36	tr A0A2C9UQA8 A0A2C9UQA8_ MANES	I,B,M,A	NAC domain containing protein 50	349	55.88
37	tr A0A2C9UP13 A0A2C9UP13_M ANES	I,B,M,A	NAC domain containing protein 50	341	55.95
38	tr A0A2C9VNM8 A0A2C9VNM8_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	456	52.62
39	tr A0A251L6B6 A0A251L6B6_M ANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	464	54.18
40	tr A0A251L698 A 0A251L698_MA NES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	463	54.19
41	tr A0A2C9VTU5 A0A2C9VTU5_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	421	63.42
42	tr A0A2C9WM25 A0A2C9WM25_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	354	48.56
43	tr A0A2C9WN37 A0A2C9WN37_ MANES	I,B,M,A	NAC domain containing protein 52	348	44.32
44	tr A0A2C9WKT2 A0A2C9WKT2_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	354	48.09
45	tr A0A2C9W4N0 A0A2C9W4N0_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	337	72.21
46	tr A0A2C9VRR0 A0A2C9VRR0_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	247	66.1
47	tr A0A2C9W5R6 A0A2C9W5R6_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	336	72.44

48	tr A0A2C9WC46 A0A2C9WC46_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	401	72.86
49	tr A0A251KRR0 A0A251KRR0_ MANES	I,B,M,A	NAC domain containing protein 82	373	60.01
50	tr A0A2C9VRP8 A0A2C9VRP8_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	298	64.92
51	tr A0A2C9VUQ0 A0A2C9VUQ0_ MANES	I,B,M,A	NAC domain containing protein 50	319	49.03
52	tr A0A2C9VE39 A0A2C9VE39_M ANES	I,B,M,A	NAC with transmembrane motif1	231	52.63
53	tr A0A2C9VIE3 A0A2C9VIE3_M ANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	173	65.87
54	tr A0A2C9VSE5 A0A2C9VSE5_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	393	69.26
55	tr A0A199UA28 A0A199UA28_M ANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	383	57.81
56	tr A0A2C9UKB8 A0A2C9UKB8_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	429	52.36
57	tr A0A2C9U024 A0A2C9U024_M ANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	329	80.71
58	tr A0A2C9UFZ4 A0A2C9UFZ4_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	411	51.26
59	tr A0A2C9URD4 A0A2C9URD4_ MANES	I,B,M,A	NAC domain containing protein 82	315	74.26
60	tr A0A2C9W4Q2 A0A2C9W4Q2_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator	323	72.25

	MANES		superfamily protein		
61	tr A0A2C9WR64 A0A2C9WR64_ MANES	I,B,M,A	NAC domain containing protein 41	68	68.57
62	tr A0A2C9U2C3 A0A2C9U2C3_ MANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	288	71.66
63	tr A0A251LEQ3 A0A251LEQ3_M ANES	I,B,M,A	NAC (No apical meristem) domain transcriptional regulator superfamily protein	357	70.04
64	tr A0A251L5A8 A0A251L5A8_M ANES	I,B,M,A	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	322	71.95
65	tr A0A251K2X1 A0A251K2X1_M ANES	I,B,M,A	Eukaryotic translation initiation factor 3C	929	83.45
66	tr A0A2C9VFH4 A0A2C9VFH4_ MANES	I,B,M,A	Eukaryotic translation initiation factor 3C	930	79.53
67	tr A0A2C9VV50 A0A2C9VV50_ MANES	I,B,M,A	Eukaryotic translation initiation factor 5A-1 (eIF-5A 1) protein	159	92.58
68	tr A0A2C9VP44 A0A2C9VP44_M ANES	I,B,M,A	Eukaryotic translation initiation factor 5A-1 (eIF-5A 1) protein	159	92.26
69	tr A0A2C9VMR2 A0A2C9VMR2_ MANES	I,B,M,A	Eukaryotic translation initiation factor 5A-1 (eIF-5A 1) protein	160	92.31
70	tr A0A2C9UXZ5 A0A2C9UXZ5_ MANES	I,B,M,A	Eukaryotic translation initiation factor 3 subunit 7 (eIF-3)	572	86.09
71	tr A0A2C9W286 A0A2C9W286_ MANES	I,B,M,A	Eukaryotic translation initiation factor 3 subunit 7 (eIF-3)	557	81.91
72	tr A0A2C9VM01 A0A2C9VM01_ MANES	I,B,M,A	Translation initiation factor eIF3 subunit	223	73.17

73	tr A0A2C9WCR1 A0A2C9WCR1_MANES	I,B,M,A	Proteasome component (PCI) domain protein	412	79.1
74	tr A0A2C9UAB5 A0A2C9UAB5_MANES	I,B,M,A	Eukaryotic translation initiation factor 3K	239	90.41
75	tr A0A2C9WK52 A0A2C9WK52_MANES	I,B,M,A	Translation initiation factor 3 subunit H1	340	57.45
76	tr A0A2C9UGV6 A0A2C9UGV6_MANES	I,B,M,A	Transducin/WD40 repeat-like superfamily protein	326	50.89
77	tr A0A2C9U7A7 A0A2C9U7A7_MANES	I,B,M,A	Transducin/WD40 repeat-like superfamily protein	326	51.23
78	tr A0A2C9VMX9 A0A2C9VMX9_MANES	I,B,M,A	Eukaryotic translation initiation factor 2	287	55.23
79	tr A0A2C9VKV5 A0A2C9VKV5_MANES	I,B,M,A	Eukaryotic translation initiation factor 2	289	54.91
80	tr A0A2C9VMB7 A0A2C9VMB7_MANES	I,B,M,A	Eukaryotic translation initiation factor 2	315	53.8
81	tr A0A2C9VX86 A0A2C9VX86_MANES	I,B,M,A	catalase 3	492	90.84
82	tr A0A2C9VVT7 A0A2C9VVT7_MANES	I,B,M,A	catalase 3	461	91.34
83	tr A0A2C9VVU3 A0A2C9VVU3_MANES	I,B,M,A	catalase 3	492	88.4
84	tr Q9SW99 Q9SW99_MANES	I,B,M,A	catalase 3	492	89.22
85	tr A0A2C9WMD1 A0A2C9WMD1_MANES	I,B,M,A	catalase 3	344	91.35

	1_MANES				
86	tr A0A2C9TZH3 A0A2C9TZH3_ MANES	I,B,M,A	catalase 3	358	89.27
87	tr A9YME8 A9Y ME8_MANES	I,B,M,A	catalase 3	261	87.02
88	tr A0A2C9U8S7 A0A2C9U8S7_M ANES	I,B,M,A	CBL-interacting protein kinase 9	499	74.2
89	tr A0A2C9U1Q2 A0A2C9U1Q2_ MANES	I,B,M,A	CBL-interacting protein kinase 3	415	73.78
90	tr A0A2C9U4V6 A0A2C9U4V6_ MANES	I,B,M,A	CBL-interacting protein kinase 9	459	73.82
91	tr A0A2C9UYQ8 A0A2C9UYQ8_ MANES	I,B,M,A	CBL-interacting protein kinase 9	457	70.73
92	tr A0A2C9VFA0 A0A2C9VFA0_ MANES	I,B,M,A	ARM repeat superfamily protein	533	63.14
93	tr A0A2C9VFC8 A0A2C9VFC8_ MANES	I,B,M,A	ARM repeat superfamily protein	530	64.04
94	tr A0A2C9VBQ4 A0A2C9VBQ4_ MANES	I,B,M,A	Importin alpha isoform 4	534	69.72
95	tr A0A2C9VBR2 A0A2C9VBR2_ MANES	I,B,M,A	ARM repeat superfamily protein	530	64.23
96	tr A0A2C9VTE1 A0A2C9VTE1_ MANES	I,B,M,A	ARM repeat superfamily protein	533	63.67
97	tr A0A251LCR3 A0A251LCR3_M ANES	I,B,M,A	ARM repeat superfamily protein	529	64.54
98	tr A0A2C9U3K5	I,B,M,A	Disease resistance protein (TIR-	1029	51.22

	A0A2C9U3K5_MANES		NBS-LRR class) family		
99	tr A0A2C9U398 A0A2C9U398_MANES	I,B,M,A	Disease resistance protein (TIR-NBS-LRR class) family	1187	56.13
100	tr A0A2C9V0P3 A0A2C9V0P3_MANES	I,B,M,A	Disease resistance protein (TIR-NBS-LRR class) family	1135	54.78
101	tr A0A2C9V4R5 A0A2C9V4R5_MANES	I,B,M,A	Disease resistance protein (TIR-NBS-LRR class) family	1239	56.35
102	tr A0A251LK88 A0A251LK88_MANES	I,B,M,A	Disease resistance protein (TIR-NBS-LRR class) family	1158	55.79
103	tr A0A251LK92 A0A251LK92_MANES	I,B,M,A	Disease resistance protein (TIR-NBS-LRR class) family	1284	53.94
104	tr A0A2C9WFM4 A0A2C9WFM4_MANES	I,B,M,A	Disease resistance protein (TIR-NBS-LRR class) family	1133	55.35
105	tr A0A2C9UNM2 A0A2C9UNM2_MANES	I,B,M,A	Disease resistance protein (TIR-NBS-LRR class) family	1100	53.58
106	tr A0A2C9VW90 A0A2C9VW90_MANES	I,B,M,A	Calcium-dependent lipid-binding (CaLB domain) family protein	538	60.86
107	tr A0A2C9V4J0 A0A2C9V4J0_MANES	I,B,M,A	Calcium-dependent lipid-binding (CaLB domain) family protein	539	58.48
108	tr A0A2C9W0I7 A0A2C9W0I7_MANES	I,B,M,A	Calcium-dependent lipid-binding (CaLB domain) family protein	540	60.42
109	tr A0A2C9W634 A0A2C9W634_MANES	I,B,M,A	Calcium-dependent lipid-binding (CaLB domain) family protein	534	53.66
110	tr A0A2C9W0F0 A0A2C9W0F0_MANES	I,B,M,A	Calcium-dependent lipid-binding (CaLB domain) family protein	429	62.24



	MANES				
111	tr A0A2C9V0Q1 A0A2C9V0Q1_ MANES	I,B,M,A	Disease resistance protein (TIR- NBS-LRR class) family	967	53.27
112	tr A0A251LKM5 A0A251LKM5_ MANES	I,B,M,A	Disease resistance protein (TIR- NBS-LRR class) family	771	54.93
113	tr A0A2C9VKE1 A0A2C9VKE1_ MANES	I,B,M,A	Target of rapamycin	991	89.67
114	tr A0A2C9VT85 A0A2C9VT85_M ANES	I,B,M,A	Calcium-dependent lipid-binding (CaLB domain) family protein	511	54.03

**MODELING OF CASSAVA-CASSAVA MOSAIC VIRUS
INTERACTIONS WITH COMPUTATIONAL BIOLOGY AND
BIOINFORMATICS APPROACH**

By

RAJANI K. R.

(2014-09-105)

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



**DEPARTMENT OF PLANT BIOTECHNOLOGY
COLLEGE OF AGRICULTURE
VELLAYANI, THIRUVANANTHAPURAM-695 522
KERALA, INDIA**

2019

9. ABSTRACT

Every year pathogenic organisms cause billions of dollars' worth damage to crops and livestock. In agriculture, study of plant-microbe interactions is demanding a special attention to develop management strategies for the destructive pathogen induced diseases that cause huge crop losses every year worldwide. *Cassava Mosaic Virus* (CMV) is a major viral leaf pathogen that causes disease in cassava. Protein-Protein Interactions (PPIs) play a critical role in initiating pathogenesis and maintaining infection. Understanding the PPI network between a host and pathogen is a critical step for studying the molecular basis of pathogenesis. The experimental study of PPIs at a large scale is very scarce and also the high throughput experimental results show high false positive rate. Hence, there is a need for developing efficient computational models to predict the interaction between host and pathogen in a genome scale, and find novel candidate effectors and/or their targets.

In this study, interacting proteins in cassava-CMV interaction is predicted using interolog-based method. The interolog method relies on protein sequence similarity to conduct the PPI prediction. Using this method, 114 PPIs have been predicted between 114 proteins of cassava and 10 proteins of CMV. Functional annotation of the predicted proteins showed the presence of 10 disease resistance protein in cassava that interacts with CMV. The subcellular location of the predicted proteins was found and it showed that major interactions occur in nucleus and chloroplast region. This can be a useful resource to the plant community to characterize the host-pathogen interaction in cassava and CMV. Further, these prediction models can be applied to the agriculturally relevant crops.

