

Ph.D.

TARUN PAL

JUIT, WAKNAGHAT

2017

DEVELOPMENT OF COMPUTATIONAL RESOURCES FOR PREDICTING DISEASE RESISTANCE GENES AND MAPPING NGS- TRANSCRIPTS TO SECONDARY METABOLISM IN PLANTS

Thesis submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

IN

BIOINFORMATICS

BY

TARUN PAL



Department of Biotechnology & Bioinformatics

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, DISTRICT SOLAN, H.P., INDIA

Month November Year 2017

**DEVELOPMENT OF COMPUTATIONAL
RESOURCES FOR PREDICTING DISEASE
RESISTANCE GENES AND MAPPING NGS-
TRANSCRIPTS TO SECONDARY
METABOLISM IN PLANTS**

Thesis submitted in fulfillment of the requirements for the Degree of

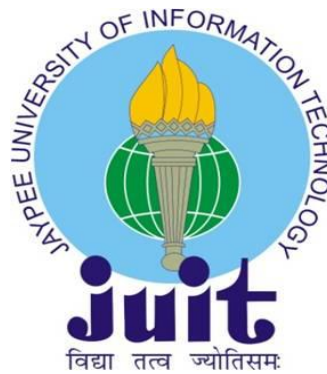
DOCTOR OF PHILOSOPHY

IN

BIOINFORMATICS

BY

TARUN PAL



Department of Biotechnology & Bioinformatics

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, DISTRICT SOLAN, H.P., INDIA

Month November Year 2017

Copyright

@

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

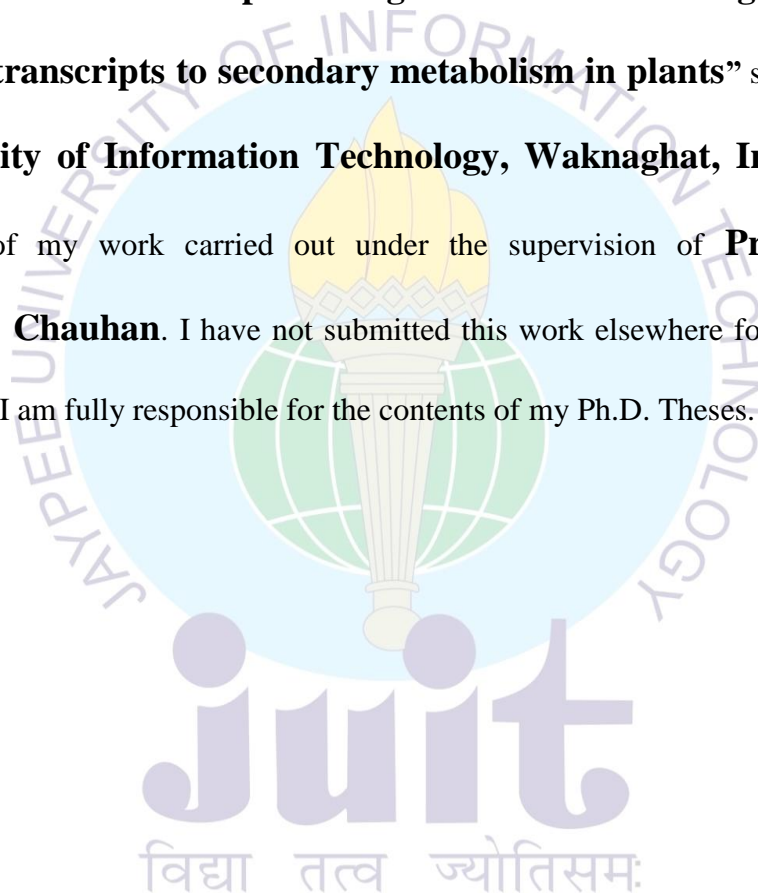
WAKNAGHAT

DECEMBER 2017

ALL RIGHTS RESERVED

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled “**Development of computational resources for predicting disease resistance genes and mapping NGS-transcripts to secondary metabolism in plants**” submitted at **Jaypee University of Information Technology, Wagnaghat, India**, is an authentic record of my work carried out under the supervision of **Prof. (Dr.) Rajinder Singh Chauhan**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D. Theses.



Tarun Pal

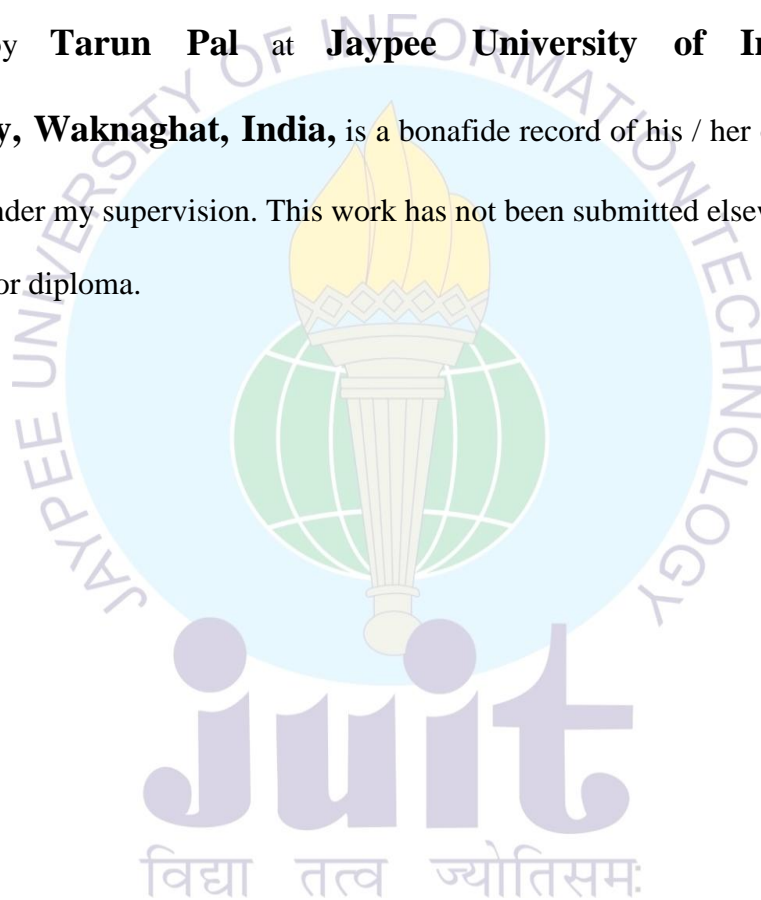
Department of Biotechnology & Bioinformatics

Jaypee University of Information Technology, Wagnaghat, India

Date:

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled “**Development of computational resources for predicting disease resistance genes and mapping NGS-transcripts to secondary metabolism in plants**”, submitted by **Tarun Pal** at **Jaypee University of Information Technology, Wagnaghat, India**, is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.



Prof. (Dr.) Rajinder Singh Chauhan (Supervisor)

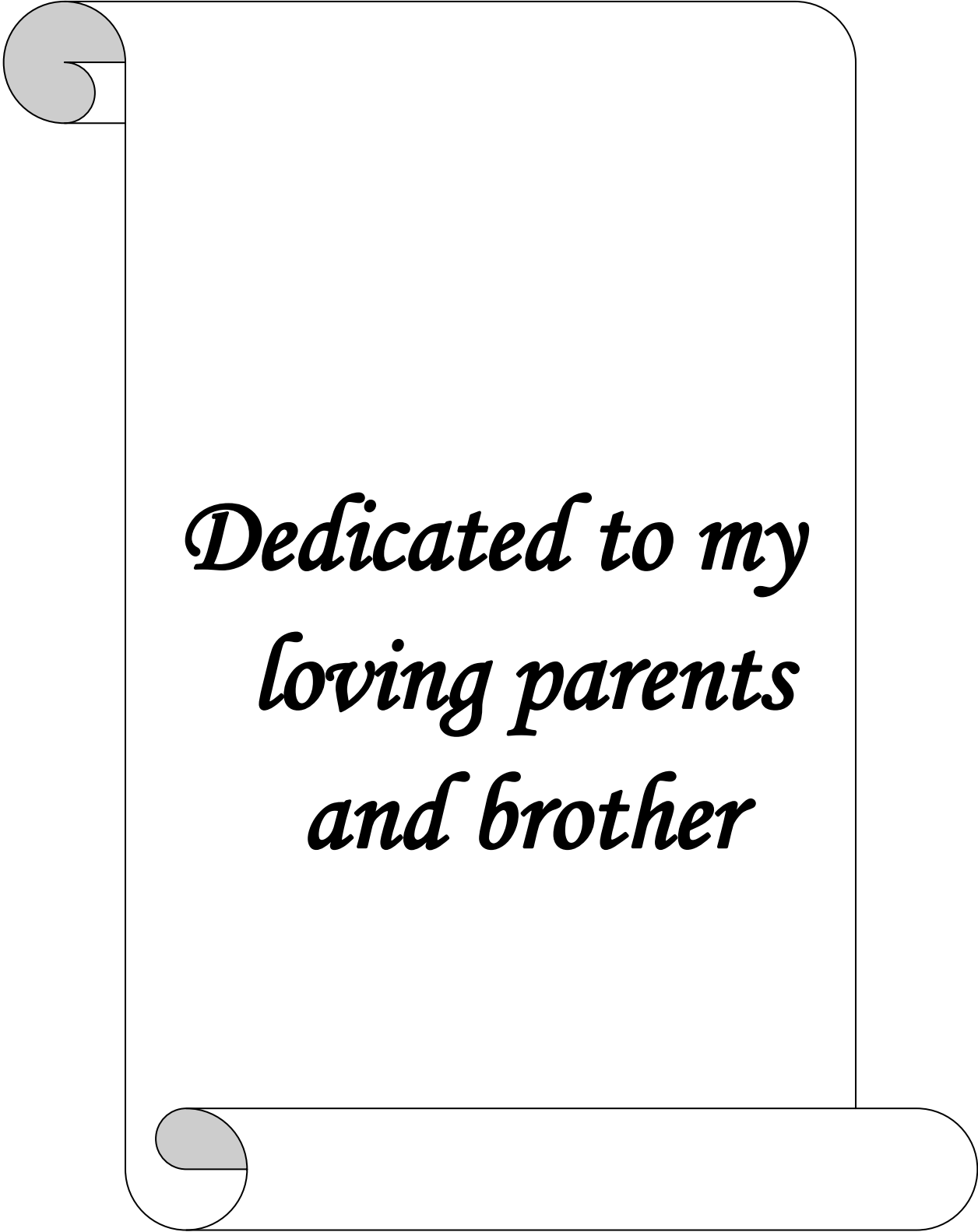
Former Dean (Biotechnology)

Department of Biotechnology & Bioinformatics

Jaypee University of Information Technology

Wagnaghat, India

Date:



*Dedicated to my
loving parents
and brother*

ACKNOWLEDGEMENT

“To speak gratitude is courteous and pleasant, to enact gratitude is generous and noble, but to live gratitude is to touch Heaven”

I feel proud and privileged to honour and place my gratitude to all those who have been supportive and encouraging throughout the completion of my Ph.D. thesis.

*With reverence and pleasure, I would like to express my unfathomable indebtedness to my esteemed supervisor/mentor, **Prof. (Dr.) Rajinder Singh Chauhan** for his unconditional support, immaculate guidance and affectionate encouragement to carry on my research. I deem it a privilege to thank him for all untiring efforts, which he has endeared to his students and scholars.*

*I feel indebted and express my loyal and venerable thanks to JUIT administration, **Prof. (Dr.) Vinod Kumar** (Vice Chancellor, JUIT); **Prof. (Dr.) Samir Dev Gupta** (Director & Academic Head, JUIT); **Maj Gen Rakesh Bassi (Retd.)** (Registrar and Dean of Student Welfare, JUIT). I gratefully acknowledge **Prof. (Dr.) Y. Medury** (former COO, Jaypee Education System, JUIT) and **Dr. Sudhir Kumar Syal** Acting Head-Department of Biotechnology and Bioinformatics for showing me this path and encouraging me to take up this journey.*

*I am also grateful to the valuable advices provided by the expert doctoral program monitoring committee (DPMC) members, **Prof. (Dr.) Satya Prakash Ghrera**, **Dr. Chittaranjan Rout** and **Dr. Tiratha Raj Singh**, which has always helped me to recognize my weakness and provided me the path to get rid of them.*

*I am tremendously thankful to the **Department of Biotechnology, Ministry of Science and Technology, Government of India** for providing financial support. I am also thankful to **Dr. Hemant Sood**, **Dr. Jayashree Ramana**, **Dr. Harvinder Singh** and **Prof. (Dr.) Pradeep Kumar Naik** for their help and valuable suggestions throughout my research work.*

I humbly acknowledge all the faculty members of Department of Biotechnology and Bioinformatics, for providing me conducive environment for carrying out this research work.

Compiling the Ph.D. degree was probably most challenging activity of my life. I

have shared deep sense of emotions while completing this degree. I feel indebted to thank two of my beloved seniors **Dr. Varun Jaiswal** and **Dr. Sree Krishna Chanumolu**.

I wish to convey my sincere thanks to all the members of non-technical staff of the department, especially **Mrs. Somlata Sharma** and **Mr. Baleshwar** for their assistance and kind cooperation extended to me.

I am indebted to all those friends who have encouraged me to do a valuable work. I extend my heartfelt thanks to **Dr. Aseem Chawla, Dr. Charu Suri, Dr. Kirti Shitiz Rohil, Dr. Archit Sood, Dr. Nikhil Malhota, Dr. Swapnil Jain, Jibesh Kumar Padhan, Pawan Kumar, Ira Vashisht, Neha Sharma, Ashwani Kumar and Vikrant Sharma**. I pay my sincere thanks to all the research scholars of the Department Biotechnology & Bioinformatics as they always kept me pushing forward with my goals.

This work would have never taken shape sans the moral imbibements inculcated in me by my grandfather **Late Shri Hira Singh Pal**. I wish to thank my parents, **Mr. K. C. Pal** and **Mrs. Rita Pal**. Their love, support and motivation provided me inspiration and was my driving force to successfully complete my work. I am also thankful to my brother **Er. Varun Pal**, my uncles **Prof. (Dr.) L. S. Pal**, **Prof. (Dr) A. C. Pal**; my aunts **Mrs. Satya Pal, Prof. (Dr.) Anita Pal** and **Prof. (Dr.) Krishna Pal** for providing me the support in one form or other.

I owe them everything and wish I could show them just how much I love and appreciate them.

Finally, I would like to express my heartfelt appreciation to all those who have contributed directly or indirectly towards obtaining my doctorate degree and extend apologies to any one whom I have failed to recognize. Last, but not the least, I thank the one above all of us, omnipresent God, for answering my prayers, for giving me the strength to plod on during each and every phase of my life.

All may not be mentioned, but no one is forgotten.

Tarun Pal

TABLE OF CONTENTS

CONTENTS	Page No.
LIST OF TABLES	I-II
LIST OF FIGURES	III-V
LIST OF ABBREVIATIONS	VI-VII
ABSTRACT	VIII-IX
CHAPTER 1: Prediction of disease resistance proteins in plants using support vector machine based learning tool	1-40
Abstract	1
1.1 Introduction	2-4
1.2 Review of Literature	5-10
1.2.1 Plant disease resistance genes and pathogen avirulence (Avr) genes	6-7
1.2.2 Classification of <i>R</i> genes	7
1.2.3 Machine learning	8-10
1.3 Materials and Methods	11-30
1.3.1 Data selection	11-12
1.3.2 Distribution of datasets into training and test sets	12
1.3.3 Feature extraction	13-15
1.3.4 Data preprocessing	15-16
1.3.5 Feature selection	16
1.3.6 Model generation	16-17
1.3.7 Model evaluations	17-18
1.3.7.1 Evaluation through test datasets	17
1.3.7.2 10-fold cross-validation	17
1.3.7.3 Performance measures	17-18
1.3.8 ROC curve	18
1.3.9 Implementation of model as tool and website development	18-30
1.4 Results and Discussion	31-34
1.4.1 Model parameter	31-33

1.4.1.1 Models parameter and accuracy	31
1.4.1.2 Performance of the support vector machine model using statistics	31-32
1.4.1.3 10-fold cross-validation in the training dataset	32
1.4.1.4 Receiver Operating Characteristic (ROC) plot	32-33
1.4.2 Evaluation on test dataset	33
1.4.3 Web Implementation	33-34
1.5 Conclusion	35
References	36-40
CHAPTER 2: Computational analysis of NGS transcriptomes of medicinal herbs (<i>Aconitum heterophyllum</i> and <i>Swertia chirayita</i>)	41-106
Abstract	41-42
2.1 Introduction	43-46
2.2 Review of Literature	47-65
2.2.1 DNA sequencing (First-generation technologies)	47
2.2.2 Need for next generation sequencing technology	47-48
2.2.3 Next generation sequencing	48
2.2.4 Transcriptome sequencing	48
2.2.5 Launching of NGS platforms	48-50
2.2.6 Illumina sequencing	50-51
2.2.7 Computational tools/pipelines for data analysis	51
2.2.8 Quality filtering	52
2.2.9 <i>De novo</i> assembly	52-53
2.2.10 Functional annotation	54
2.2.11 Gene ontology	54-55
2.2.12 COG classification	55-57
2.2.13 Domain search	57
2.2.14 <i>In silico</i> transcript abundance	57-58
2.2.15 Biological pathways and network connectivity diagrams	58-62
2.2.15.1 KEGG pathway database	59-60
2.2.15.2 BioCyc database	50-61
2.2.15.3 Reactome database	61

2.2.15.4 WikiPathways	61
2.2.15.5 NCBI Biosystems database	62
2.2.15.6 Network connectivity diagrams	62
2.2.16 <i>Aconitum heterophyllum</i>	62-63
2.2.17 <i>Swertia chirayita</i>	63-64
2.2.18 Secondary metabolism and associated pathways (MVA/MEP)	64-65
2.3 Materials and Methods	66-75
2.3.1 Medicinal plant species	66
2.3.2 Transcriptomes generation	66-67
2.3.3 <i>De novo</i> assembly using Illumina HiSeq 2000 platform	67
2.3.4 Functional annotation, GO mapping and COG analysis	68
2.3.5 Transcript abundance prediction using fragment mapping approach	69
2.3.6 Computational mining of transcriptomes for MVA/MEP pathway genes	69
2.3.7 Domain prediction using Pfam database	70
2.3.8 Pathway mapping using KEGG in <i>A. heterophyllum</i> and <i>S. chirayita</i> transcriptomes	70
2.3.9 Biosystems classification and network connectivity diagrams	70-75
2.3.10 Data availability	75
2.4 Results and Discussion	76-96
2.4.1 <i>De novo</i> sequencing using Illumina platform	76-77
2.4.2 Functional annotation and classification of transcripts	78-82
2.4.3 GO and COG functional classification	82-88
2.4.4 Identification of domains	88-89
2.4.5 Mining transcriptomes for genes involved in MVA/MEP biosynthesis pathways	89-93
2.4.6 Mapping of transcriptomes on KEGG pathways	93-94
2.4.7 NCBI Biosystems for functional classification	94-96
2.5 Conclusion	97
References	98-106

CONCLUSION AND FUTURE PROSPECTS	107-108
APPENDICES	109-181
LIST OF PUBLICATIONS	182-183

LIST OF TABLES

Table no.	Title	Page No.
Table 1.1	Reference <i>R</i> proteins and their seven domain classes	11
Table 1.2	Distribution of data set into training, test, positive and negative classes	12
Table 1.3	Protr package generated descriptors	15
Table 1.4	Performance of support vector machine model on dataset	32
Table 2.1	Comparison of leading NGS Platforms	50
Table 2.2	Comparison of next-generation sequencing Assemblers (single-end reads/paired -end) [adapted by Zhang et al. [32]]; “*” indicates any operating systems with Perl interpreter	53
Table 2.3	COG functional classification (Tatusov et al. 2003 [40])	56-57
Table 2.4	The list of important existing pathway databases	59
Table 2.5	The contents and database of KEGG	60
Table 2.6	Important pharmacological values of <i>A. heterophyllum</i>	63
Table 2.7	Important medicinal properties of <i>S. chirayita</i>	64
Table 2.8	Assembly statistics for transcriptomes from roots (AHSR) and shoots (AHSS) samples of <i>A. heterophyllum</i>	77
Table 2.9	Assembly statistics for transcriptomes from greenhouse (SCFG) and tissue cultured (SCTC) samples of <i>S. chirayita</i>	77
Table 2.10	Prediction statistics of CDS, exons and peptides from roots (AHSR) and shoots (AHSS) samples of <i>A. heterophyllum</i>	78
Table 2.11	Prediction statistics of CDS, exons and peptides from greenhouse (SCFG) and tissue cultured (SCTC) transcriptomes of <i>S. Chirayita</i>	80

Table 2.12	<i>In silico</i> transcript quantification for MVA and MEP pathways genes predicted in <i>A. heterophyllum</i> root and shoot transcriptomes	89-90
Table 2.13	<i>In silico</i> transcript quantification for MVA and MEP pathways genes predicted in greenhouse and tissue cultured samples of <i>S. chirayita</i>	91
Table A1	Reference <i>R</i> genes (manually curated) of PRGDB	109-115
Table A2	KOs associated with secondary metabolism in AHSR	115-128
Table A3	KOs associated with secondary metabolism in AHSS	128-141
Table A4	KOs associated pathways in SCFG	141-153
Table A5	KOs associated pathways in SCTC	153-165
Table A6	Functional classification by NCBI Biosystems (AHSR)	165-169
Table A7	Functional classification by NCBI Biosystems (AHSS)	169-172
Table A8	Functional classification of SCFG transcriptome by NCBI Biosystems	172-177
Table A9	Functional classification of SCTC transcriptome by NCBI Biosystems	177-181

LIST OF FIGURES

Fig. no.	Title	Page No.
Fig. 1.1	Illustrative representation of various pathogens and their effect on plants (medicinal) [31]	6
Fig. 1.2	Structure of NBS-LRR gene (NBS- Nucleotide binding site; LRR- Leucine rich repeat; TIR- Toll and interleukin-1 receptor-like domain; CC- Coiled-coil domain; N- Amino terminus; C- Carboxyl terminus) [38]	7
Fig. 1.3	The optimal hyperplane and maximum margin linearly separating 2D-points	9
Fig. 1.4	An Artificial neural network	10
Fig. 1.5	Workflow for the methodology used in DRPPP	12
Fig. 1.6	ROC Curve for Binary SVM classifier: ROC plot depicts relative trade-offs between true positive and false positives	33
Fig. 1.7	Snapshot of DRPPP's interface	34
Fig. 2.1	Mature tuberous roots of <i>A. heterophyllum</i> plant	44
Fig. 2.2	A representative diagram showing different modules for alkaloids biosynthesis in <i>A. heterophyllum</i> [adapted from Rodriguez-Concepción et al. [8]]	44
Fig. 2.3	Mature green house grown <i>S. chirayita</i> plant	45
Fig. 2.4	Flow diagram of whole transcriptome sequencing, assembly, annotation and analysis for root (AHSR) and shoot (AHSS) transcriptomes of <i>A. heterophyllum</i> and tissue cultured (SCTC) and greenhouse grown (SCFG) plant transcriptomes of <i>S. chirayita</i> .	67
Fig. 2.5	Species distribution of the top Blastx hits against <i>A.</i>	79

	<i>heterophyllum</i> root transcriptome	
Fig. 2.6	Species distribution of the top Blastx hits against <i>A. heterophyllum</i> shoot transcriptome	79
Fig. 2.7	Proportion of SCFG transcripts matching to different plant species	81
Fig. 2.8	Proportion of SCTC transcripts matching to different plant species	82
Fig. 2.9	Distribution of GO annotated transcripts for <i>A. heterophyllum</i> root transcriptomes	83
Fig. 2.10	Distribution of GO annotated transcripts for <i>A. heterophyllum</i> shoot transcriptomes	83
Fig. 2.11	Distribution of GO annotated transcripts in SCFG transcriptome of <i>S. chirayita</i>	84
Fig. 2.12	Distribution of GO annotated transcripts in SCTC transcriptome of <i>S. chirayita</i>	85
Fig. 2.13	Distribution of COG classified transcripts of root and shoot transcriptomes of <i>A. heterophyllum</i>	87
Fig. 2.14	Distribution of COG classified transcripts of greenhouse and tissueculture transcriptomes of <i>S. chirayita</i>	88
Fig. 2.15	Graphical representation of <i>in silico</i> transcript quantification for secondary metabolism (MVA and MEP) pathway genes in root versus shoot transcriptomes	90
Fig. 2.16	Graphical representation of <i>in silico</i> transcript quantification for secondary metabolism (MVA and MEP) pathway genes in greenhouse versus tissue cultured transcriptomes	92
Fig. 2.17	Flow diagram used for constructing network connectivity diagram	95

Fig. 2.18	Construction of isoquinoline alkaloids biosynthesis network connectivity diagram in root transcriptome of <i>A. heterophyllum</i>	96
-----------	---	----

LIST OF ABBREVIATIONS

ACTH	Acetoacetyl-CoA thiolase
AHSR	Root transcriptome
AHSS	Shoot transcriptome
AUC	Area under the curve
COG	Cluster of orthologous group
DRPPP	Disease resistance plant protein predictor
DXPR/DXR	1-deoxy-D-xylulose 5-phosphate reductase
DXPS/DXS	1-deoxy-D-xylulose 5-phosphate synthase
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GDPS	Geranyl diphosphate synthase
GO	Gene ontology
HDS/ISPG	1-hydroxy-2-methyl-2-(<i>E</i>)-butenyl 4-diphosphate synthase
HMGR	3-Hydroxy-3-methyl glutaryl CoA reductase
HMGS	3-Hydroxy-3-methylglutaryl-CoA synthase
IPPI	Isopentenyl pyrophosphate isomerase
ISPD	2- <i>C</i> -methylerythritol 4-phosphate cytidyl transferase
ISPE	4-(cytidine-5'-diphospho)-2- <i>C</i> -methylerythritol kinase
ISPH	1-hydroxy-2-methyl-2-(<i>E</i>)-butenyl 4-diphosphate reductase
KO	KEGG orthology
MECPS/ISPF	2- <i>C</i> -methylerythritol-2,4-cyclophosphate synthase
MEP	2- <i>C</i> -Methyl-D-erythritol 4-phosphate
MVA	Mevalonate

MVDD	Mevalonate diphosphate decarboxylase
MVK	Mevalonate kinase
NBS-LRR	Nucleotide binding site and leucine-rich repeat
NGS	Next-generation sequencing
PMK	Phosphomevalonate kinase
RLK	Receptor-like kinases
SCFG	Greenhouse transcriptome
SCTC	Tissue cultured transcriptome
SVM	Support vector machine
RSEM	RNA-Seq by Expectation Maximization
NR	Non-Redundant database compiled by the NCBI

ABSTRACT

Plant disease outbreak is increasing rapidly around the globe and is a major cause for crop loss worldwide. Plants, in turn, have developed diverse defense mechanisms to identify and evade different pathogenic microorganisms. Early identification of plant disease resistance genes (*R* genes) can be exploited for crop improvement programs. The existing prediction methods are either based on sequence similarity/domain-based methods or electronically annotated sequences, which might miss existing unrecognized proteins or low similarity proteins. Therefore, there was an urgent need to devise a novel machine learning tool to address this problem. Considering these gaps and importance of disease resistance genes, DRPPP (Disease resistance plant protein predictor), a support vector machine (SVM) learning based tool was developed. 16 different methods were generated through feature extraction method and were employed to generate 10,270 features. Radial basis function was used and ten-fold cross validation was performed to optimize SVM parameters. The model for DRPPP was derived using LibSVM and achieved an overall accuracy of 91.11% on the test dataset. The tool was found to be robust and can be used for high-throughput datasets.

Furthermore, in plants the importance of medicinal herbs can be derived from the fact that the demand for herbal medicines is estimated to increase upto US\$3 trillion by 2020. However, their genetic improvement has been hampered due to lack of genome resources. Next-generation sequencing (NGS) has provided unprecedented opportunities for high throughput research on medicinal plants, especially for those whose genome/transcriptome datasets were still not available. Therefore, NGS transcriptomes for two critically endangered species i.e. *Aconitum heterophyllum* and *Swertia chirayita* having high therapeutic values were generated and computationally analyzed for varying conditions of secondary metabolites. In total, four transcriptomes were generated for differential tissues (root versus shoot) in *Aconitum heterophyllum* and differential conditions (greenhouse versus tissue cultured) in *Swertia chirayita* differing for biosynthesis and accumulation of secondary metabolites. The *in silico* transcript quantification for aconites biosynthesis through mevalonate (MVA) and (MEP) pathway genes revealed that 4 genes *HDS* (1-hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate synthase), *HMGR* (3-hydroxy-3-methylglutaryl-CoA reductase), *MVK* (mevalonate kinase and *MVDD* (mevalonate diphosphate decarboxylase) showed higher expression in root (AHSR) as compared to shoot (AHSS) transcriptome. Similarly, in case of *Swertia*

chirayita transcript abundance analysis for MVA/MEP revealed that most of the genes (9 genes) demonstrated higher transcript abundance in SCFG as compared to SCTC transcriptomes. Construction of isoquinoline alkaloid biosynthesis network connectivity diagrams associated with secondary metabolism were drawn in the root transcriptome of *A. heterophyllum*. Moreover, the transcriptomes of these two important plant species were assembled, annotated and characterized for the first time to reveal the molecular components contributing to biosynthesis and accumulation of secondary metabolites.

The study highlights the importance of an efficient machine learning based computational tool for predicting disease resistance proteins in plants together with the development of transcriptomic resources for two important medicinal herbs. These transcriptomic resources can be exploited to decipher not only discovering candidate genes involved in secondary metabolites production, but also for understanding the molecular basis of various biological processes. The generated resources can be used for planning a suitable genetic intervention strategy for the improvement of these plant species.

CHAPTER 1

Prediction of disease resistance proteins in plants using support vector machine based learning tool

Abstract

Plant disease outbreak is increasing rapidly around the globe and is a major cause for crop loss worldwide. Plants, in turn, have developed diverse defense mechanisms to identify and evade different pathogenic microorganisms. Early identification of plant disease resistance genes (*R* genes) can be exploited for crop improvement programs. The existing prediction methods are either based on sequence similarity/domain-based methods or electronically annotated sequences, which might miss existing unrecognized proteins or low similarity proteins. Therefore, there is an urgent need to devise a novel machine learning technique to address this problem.

In the current study, a SVM-based tool was developed for prediction of disease resistance proteins in plants. All known disease resistance (*R*) proteins (112) were taken as a positive set, whereas manually curated negative dataset consisted of 119 *R* proteins. Feature extraction generated 10,270 features using 16 different methods. The ten-fold cross validation was performed to optimize SVM parameters using radial basis function. The model was derived using LibSVM and achieved an overall accuracy of 91.11% on the test dataset. The tool was found to be robust and can be used for high-throughput datasets. The current study provides instant identification of *R* proteins using machine learning approach, in addition to the similarity or domain prediction methods.

1.1 Introduction

Plants are in continuous attack from various invading pathogens, including bacteria, virus, fungi, nematodes and oomycetes [1, 2]. To combat such defence-related attacks plants-in turn have disease resistance (R) genes, which can recognize and respond to infections caused by pathogens. Pathogenic microorganisms have evolved several mechanisms, including structural and enzymatic components in order to enter their hosts. Diverse life strategies are being followed by different pathogens to pierce into plant system.

Plants do not have defined immune system as in case of animals, though they have developed two defense systems for recognition and evading pathogenic microorganisms and pests. The first mechanism consists of basal defense, where extracellular transmembrane receptors identify pathogen-associated molecular patterns (PAMPs) also known as microbe-associated molecular patterns (MAMPs) and bring forth basal defence. Whereas the second mechanism consists of the adaptive immune system, which mainly involves defense layer consisting of NBS-LRR proteins and proteins associated with effector-triggered immunity (ETI). Although if either of the two mechanisms activates it leads to overlapping set of responses such as transcriptional reprogramming of the plant cells to invade pathogen. Plants are equipped with specialised pattern recognition receptors (PRRs), present at plasma membrane in order to recognise pathogens. These PRRs identify pathogen-associated molecular patterns (PAMPs), and damage-associated molecular patterns (DAMPs). PAMPs are conserved molecular motifs found within a class of microbes such as bacterial flagellin, Ef-Tu and fungal chitin [3, 4]. The detection of PAMPs by PRRs leads to PAMP-triggered immunity (PTI) which further induces mitogen-activated protein kinases (MAPKs) signalling. PTI results in global transcriptional reprogramming carried out by plant WRKY transcription factors.

The second mechanism effector-triggered immunity (ETI) reinstates PTI basal transcriptional programs and antimicrobial defenses and is also coupled with plant cell death [5]. Pathogens produce avirulence (Avr) genes, which are recognized by plant disease resistance (R) proteins. These *R* proteins recognize elicitor, leading to initiation of downstream signaling responses. Majority of *R* genes in plants are composed of nucleotide binding site (NBS) and a leucine-rich repeat (LRR) domain(s) known as NBS-LRR genes. On the basis of motif and domain classification NBS-LRR proteins are further classified into two sub-classes. The first class contains N-terminal

Toll/Interleukin1 (TIR)-like domain known as TIR-NBS-LRR (TNL), whereas the second class non-TIR-NBS-LRRs consists of domains lacking TIR-domain. The non-TIR-NBS-LRRs proteins contain coiled-coil (CC) domain in place of TIR domain. The NBS-LRR class is the most abundant class of genes found in plant families [6-8]. In NBS-LRR genes the NBS domain plays a role for binding to ATP [9], whereas the C-terminal leucine-rich repeat (LRR) is responsible for binding to pathogen-derived molecules and regulation of signal transduction [10, 11]. The NBS domain attached either to TIR or to non-TIR proteins consists of motifs kinase 1a (P-loop), kinase 2 and kinase 3a collectively referred as NB subdomain [12, 13]. These *R* genes are functionally classified on the basis of domains into seven different classes namely CNL (CC-NB-LRR), TNL (TIR-NB-LRR), NL (NBS-LRR), RLP (ser/thr-LRR), RLK (Kin-LRR), TN (TIR-NBS) and others [14].

Since these *R* genes are important, but only few computational resources had been developed in the past. One of the most popular, comprehensive resource is PRGdb (Plant Resistance Gene database), it is a dedicated repository storing 112 reference resistance proteins and 1,04,335 putative *R* proteins from around 233 plant species [14]. These 112 reference resistance proteins have been recognized from these species: *Arabidopsis thaliana* (25), *Oryza sativa* (20), *Solanum lycopersicum* (13), *Hordeum vulgare* (8), *Solanum pimpinellifolium* (5), *Solanum tuberosum* (5), *Triticum aestivum* (5), *Cucumis melo* (4), *Glycine max* (3), *Linum usitatissimum* (3), *Zea mays* (3), *Capsicum annuum* (2), *Nicotiana benthamiana* (2), *Solanum bulbocastanum* (2), *Solanum habrochaites* (2), *Aegilops tauschii* (1), *Beta vulgaris* (1), *Capsicum chacoense* (1), *Helianthus annuus* (1), *Lactuca sativa* (1), *Nicotiana glutinosa* (1), *Nicotiana tabacum* (1), *Phaseolus vulgaris* (1), *Solanum acaule* (1) and *Solanum demissum* (1) species. It's predictive pipeline DRAGO classifies putative data into CN, CNL, Mlo-like, N, NL, RLK, RLK-GNK2, RLP, RPW8-NL T, TN, TNL and Other classes using domain dependent approach.

The breeding programs and process of genetic improvement can get an immediate boost if there is an instant identification *R* proteins and thereby leading to disease-resistant varieties [15]. Conversely, this prediction is not so easy and is complex due to the repetitive nature of NBS-LRR genes. In the past, several computational methods have been developed for predicting *R* proteins, including sequence alignment, BLAST search, domain/motif analysis and phylogenetic analysis. These methods employ different tools to predict *R* proteins, such as Pfam (<http://pfam.xfam.org/>), Hidden Markov Model

(HMM) [16], Prosite (<http://prosite.expasy.org/>), SMART (<http://smart.embl-heidelberg.de/>) [17] and InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan5/>). These existing methods are more laborious and sometimes even require the capability of extensive data processing. In the recent past, two pipelines namely NLR-parser and NBSPred were developed for identification of *R* proteins. The current limitations with NLR-parser is that it uses motif alignment and search tool (MAST) output to recognize TNL or CNL class, while NBSPred is a machine learning pipeline, but it considers electronically curated datasets for model building [18, 19]. Moreover, while dealing with genomic assemblies NLR-parser is incompatible and using electronically annotated sequences by NBSPred is not a precise approach for model building.

Current prediction methods are based on sequence similarity or domain-based methods, which may miss some already existing unrecognized proteins. These methods have limited prediction accuracy while dealing with less sequence similarity *R* proteins, but are more effective while predicting similar proteins to existing *R* proteins. Therefore, to get rid of all these problems there is an immediate need for alignment and domain-independent methods for prediction of *R* proteins. In the past, the prediction problems have been efficiently tackled by machine learning methods, especially support vector machine (SVM). The support vector machine (SVM) has been very helpful in many biomedical fields particularly while dealing with prediction problems. The performance of SVM classifier is reported to be high while classifying various biomedical datasets [20, 21]. In such a scenario, development of an efficient machine learning tool for identification of *R* proteins would prove to be ideal for tackling rapidly increasing genomic and transcriptome datasets. Keeping in view lack of an efficient machine learning tool for prediction of disease resistance proteins in plants the present research was carried with the following objectives:

Objective 1: Development of a machine learning based computational tool for prediction of disease resistance proteins in plants

1.2 Review of Literature

Plants are a valuable source of human livelihood besides providing basic necessities such as food, clothing, shelter and health care; they also feed livestock and act as a raw material for the pharmaceutical industry. The potential of plants to gratify this increase in demand is an all time concern. More urbanization has resulted in increase in demand. Currently, there are about 3,74,000 thousand known species of plants [22]. Out of these some species are declared as critically extinct due various biotic and abiotic challenges faced in routine by them. To counter biotic challenges plants have a special set of genes known as disease resistance genes or *R* genes. These *R* genes encode proteins, which can identify avirulence (*Avr*) genes expressed by various pathogens and trigger downstream signalling process. The first *R* gene was isolated from maize known as *HMI*, it was found to be responsible for developing resistance against the leaf spot fungus *Cochliobolus carbonum* [23]. The most popular *R* genes class is nucleotide binding site (NBS) and a leucine-rich repeat (LRR) domain(s) known as NBS-LRR genes [6-8].

The comparative analysis of these genes indicated that they belong to a large gene family with varied copy number and have unsimilar distributions among subclasses [24, 25]. The apple genome, which contained around 1,000 NBS–LRR genes, while its genome size is not large (about 740 Mb) [26], whereas in large genomes such as of maize (over 2 Gb), only around 120 of them had been reported [27]. Since these *R* genes are important and their early identification can boost the process of genetic improvement, but there are only few computational resources available for their detection. Methods such as the sequence alignment, BLAST search, domain/motif analysis and phylogenetic analysis, were used for a period of time to predict these genes [28-30]. In 2010, a BLAST search based pipeline namely disease resistance analysis and gene orthology (DRAGO) pipeline of PRG database came into existence [14], whereas, in 2015 another motif and search based pipeline the NLR-parser was introduced. In 2016, first machine learning based pipeline came into existence [18].

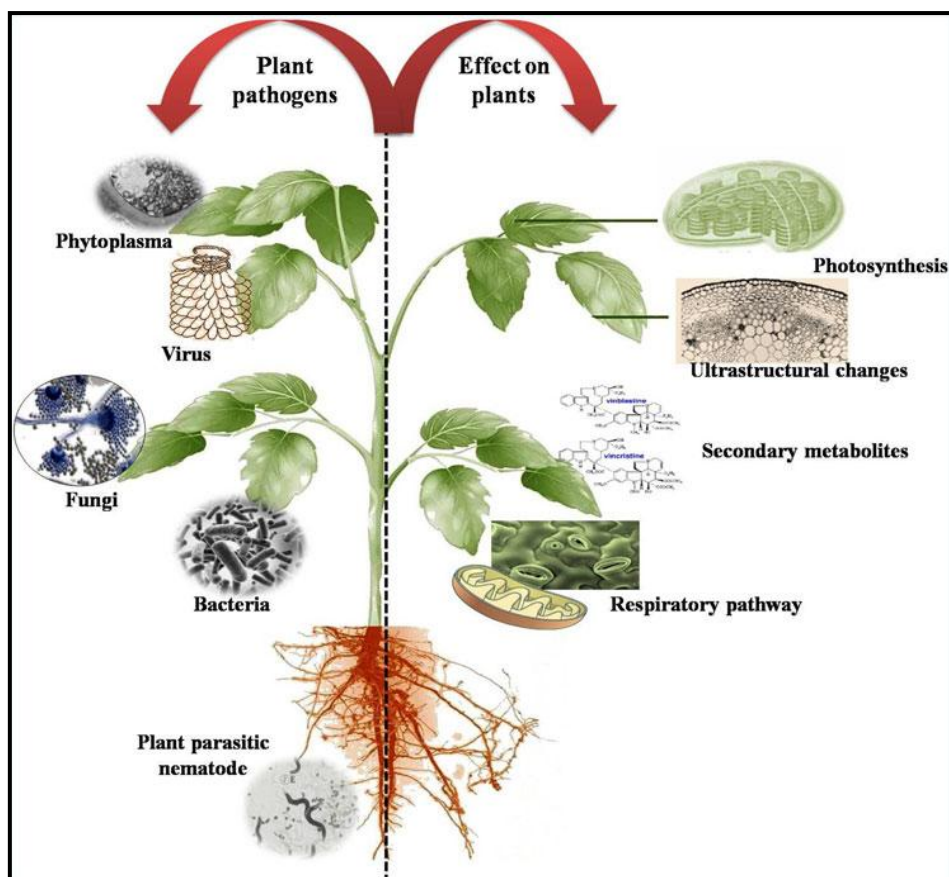


Fig. 1.1 Illustrative representation of various pathogens and their effect on plants (medicinal) [31]

The overall loss caused by these pathogens (bacteria, fungi, viruses, phytoplasmas, phytonematodes) to plants (Fig. 1.1), is also a major threat to the yield, biomass, bioactive potential and prospects of plants. The current status of literature pertaining to disease resistance genes (*R* genes) has been reviewed as under:

1.2.1 Plant disease resistance genes and pathogen avirulence (*Avr*) genes

Plants are in constant attack from viruses, microbes, invertebrates, and even other plants. They do not possess a circulatory system as in the case of animals, and therefore individual cells should possess inducible defense capability against pathogens. Plant pathogens can be classified on the basis of their modes of nutrition: necrotrophy, biotrophy or hemibiotrophy [32-34]. In 1940s, HH Flor proposed the “gene-for-gene” model concept, stating that in order to take place resistance there should be complementary dominant genes in both host as well as pathogen. The first isolated *R* gene Hm1 encodes for reductase enzyme, which detoxifies *C. carbonum* toxins. The ultimate

aim of the plant is to inhibit growth of invading pathogens by specialized genes known as disease resistance genes or *R* genes.

1.2.2 Classification of *R* genes

The major class of *R* genes is NBS-LRR genes, which encodes NBS-LRR proteins. These proteins are further subcategorized into two classes on the basis of motifs and domains present. The domain containing N-terminal Toll/Interleukin1 (TIR) are referred as NBS-LRR (TNL) proteins, whereas those lacking are known as non-TIR-NBS-LRRs. Further, these can also be referred as CC-NBS-LRR proteins if coiled-coil (CC) domain replaces TIR domain (Fig. 1.2). In NBS-LRR, the NBS domain binds to ATP, whereas the leucine-rich repeat LRR domain is responsible for binding to pathogen-derived molecules and regulation of signal transduction [9, 10]. The NBS domain consists of small motifs, kinase 1a (P-loop), kinase 2 and kinase 3a, which are jointly referred as NB subdomain [12]. The domains of NBS-LRR proteins functions together to detect pathogen effectors and thereby activate downstream signalling responses. These genes are found in all angiosperms inspected till date, but there exists a difference w.r.t. monocot and dicot species, i.e. the majority of the genes in *A. thaliana* encode for TIR domain [35], whereas, in case of cereal species the subclasses of these genes are found to be absent [36, 37]. This observation indicated that dicots preserved TIR domain associated with NBS-encoding genes, but not by monocots. These *R* genes are also classified into seven different types on domain basis, composing of CNL (CC-NB-LRR), TNL (TIR-NB-LRR), NL (NBS-LRR), RLP (ser/thr-LRR), RLK (Kin-LRR), TN (TIR-NBS) and others [51]. Total 112 reference *R* genes (manually curated) known till date in the literature [14].

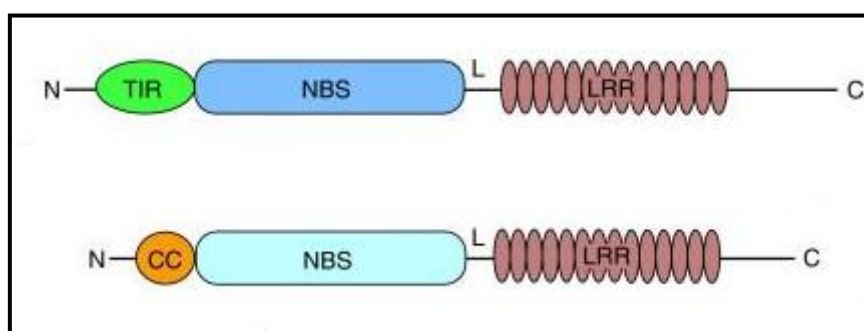


Fig. 1.2 Structure of NBS-LRR gene (NBS- Nucleotide binding site; LRR- Leucine rich repeat; TIR- Toll and interleukin-1 receptor-like domain; CC- Coiled-coil domain; N- Amino terminus; C- Carboxyl terminus) [38]

1.2.3 Machine learning

Machine learning is the subset of the computer science domain that provides a computer with the ability to learn without being explicitly programmed [39]. Later in 1998 Tom M. Mitchell defined it as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." Machine learning problems are broadly categorized into three categories:

- 1) **Supervised learning:** The labels are provided to the learning algorithm, i.e. inputs together with desired output and the algorithm utilizes this information to extract important rules and these can be used to build model and predict results for various inputs.
- 2) **Unsupervised learning:** The labels are not provided to the learning algorithm, therefore it has to recognize the rules on its own, which can be further utilized for various inputs.
- 3) **Reinforcement learning:** In this field of machine learning software program interact with a dynamic environment and carry out certain specific goals.

There also exists semi-supervised learning layer in between supervised learning and unsupervised learning layer. This learning considers unlabelled data for prediction purposes, which may consist of relatively small amount of labelled data. Machine learning utilizes several algorithms such as:

- 1) **Support vector machines (SVM):** The concept of support vector machine algorithm erupted in 1963 by Vladimir N. Vapnik and Alexey Ya. Chervonenkis. The current module of this invention was published by Corinna Cortes and Vapnik in the year 1995. It is categorized as a supervised machine learning algorithm used to classify binary data or regression. This algorithm inputs labelled training datasets and builds a model, which can be used to classify any non-labelled dataset. This algorithm uses kernel function to transform input space into a multidimensional space. The kernel has to optimize two key parameters, i.e. c and gamma. Where c parameter controls overfitting of the model and parameter gamma specifies the degree of nonlinearity of the model. For classification and regression, it constructs hyperplane in high dimensional space (Fig. 1.3). In the past, the application of this algorithm has proven good performance not only in the field of computer science, but also for biomedical datasets [21, 40].

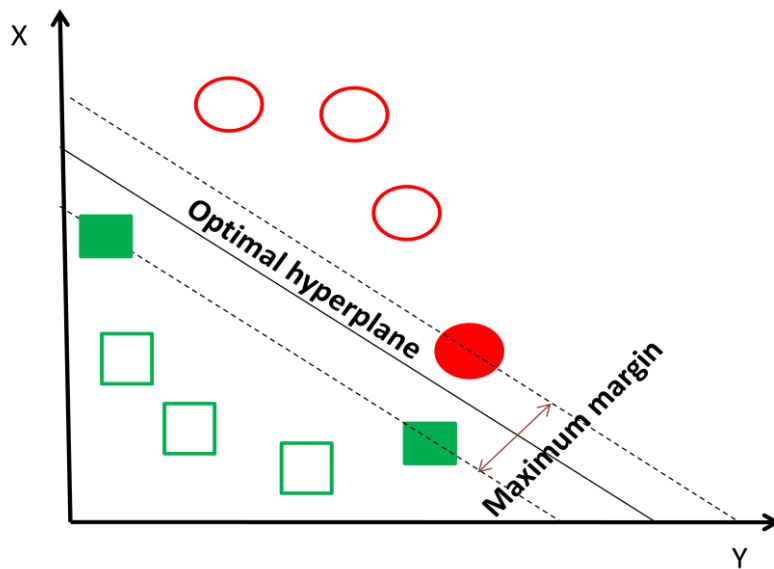


Fig. 1.3 The optimal hyperplane and maximum margin linearly separating 2D-points

Some of the popular support vector machine software's include LIBSVM, SVM^{light} and Weka, etc. LIBSVM (Library for Support Vector Machines) is a freely available and extensively used SVM classification and regression package. It has an interface of Python and important kernel includes linear, polynomial and radial basis function. SVM^{light} it is also a popular SVM software with fast optimization algorithm and distributed in C++ sources. Whereas, Weka is basically designed for data mining tasks and consists of collection of machine learning algorithms.

- 2) **Artificial neural networks (ANN):** Nerural networks are a computational model motivated by biological nervous system. The brain is composed of neurons connected by axons and work together to solve various problems. The brain learns from the past experiences rather than being programmed. The ANN is inspired by biological nervous system and contains multiple layers such as input, hidden and output layer (Fig. 1.4).

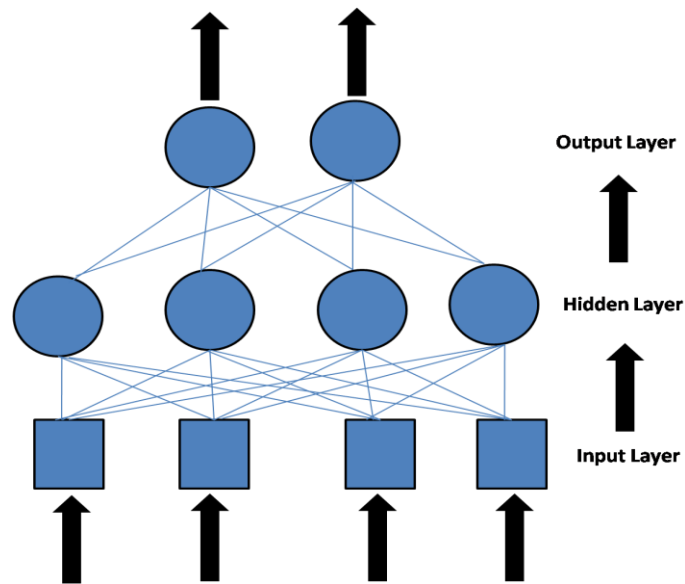


Fig. 1.4 An Artificial neural network

- 3) **Clustering:** It is the process to group similar objects in a such a way that the objects in particular cluster are more similar to each other than to those present in other clusters. This technique of unsupervised machine learning is used for statistical data analysis and in data mining. There are different algorithms for clustering such as hierarchical clustering, k-means clustering and centroid-based clustering, etc.

- 4) **Decision tree learning:** It uses decision tree for classification of data, where data is classified by submitting it to a series of tests that determine the class label. These series of test are represented in a hierarchical structure fashion. It uses predictive modelling strategy, which forms a part of statistics and machine learning. For a finite set of values in a tree model are known as classification trees. These trees are of different types such as classification and regression trees.

1.3 Materials and Methods

1.3.1 Data selection

In total, 112 known reference *R* proteins from 25 plant species were downloaded from the existing PRGdb database. These 112 *R* proteins belonging to seven different domain classes served as a positive set for building the tool (Table 1.1). Whereas, for building the negative set all known protein sequences from the same 25 plant species were taken from the NCBI protein database. From these sequences, 158 protein sequences were screened randomly using in-house developed Perl script. Further, these 158 protein sequences were manually checked to confirm as non-*R* proteins. These sequences were scaled down to reduce the redundancy using CD-HIT (<http://weizhong-lab.ucsd.edu/cd-hit/>) [41] program, which resulted in 119 non-*R* protein sequences labelled as negative dataset. The workflow for the methodology adopted in the current tool is given in Fig. 1.5.

Table 1.1 Reference *R* proteins and their seven domain classes

Sr. no.	Reference Domain classes	Number of reference Proteins
1	TNL (TIR-NBS-LRR)	16
2	CNL (CC-NBS-LRR)	50
3	NL (NBS-LRR)	4
4	RLP (Receptor-like proteins)	12
5	RLX (Receptor-Like Kinases)	10
6	TN (TIR-NBS)	1
7	Others (conferring resistance)	19
	Total	112

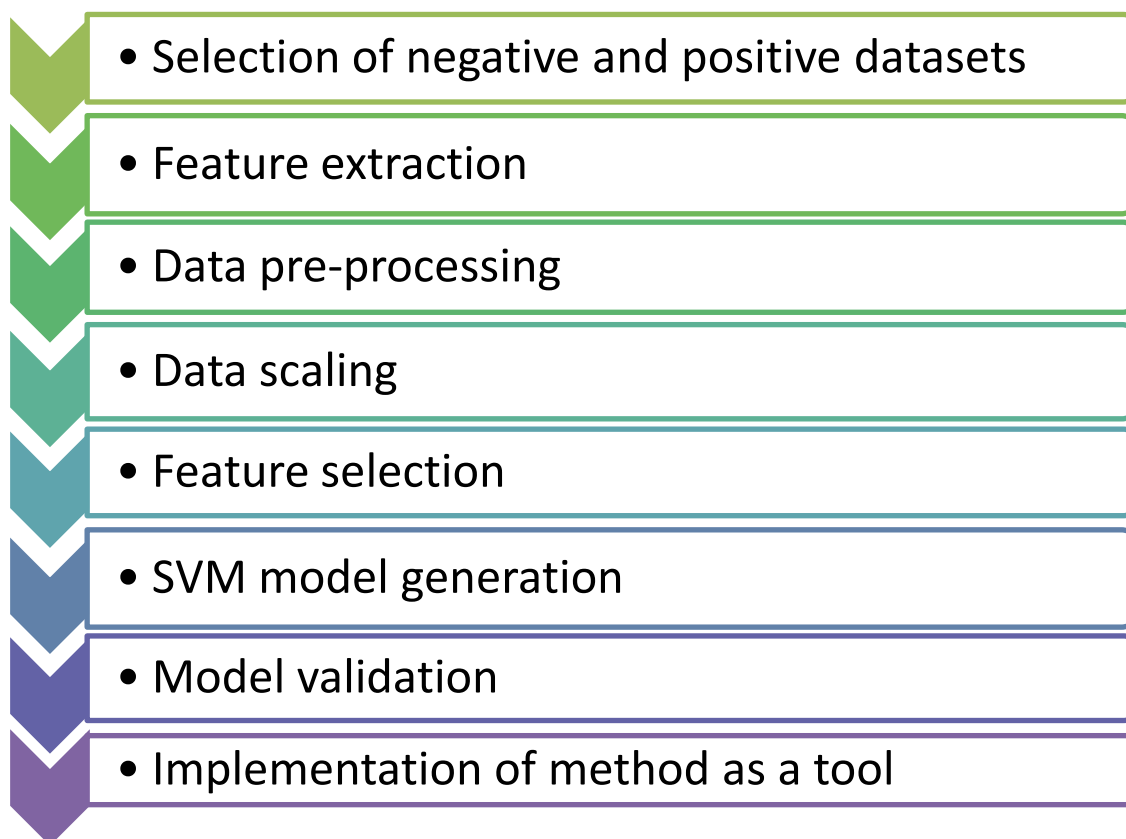


Fig. 1.5 Workflow for the methodology used in DRPPP

1.3.2 Distribution of datasets into training and test sets

Identified positive and negative datasets were further distributed into training and test datasets. From the above-mentioned sequences, 80% of the random sequences were selected in the training set and the remaining were included in the test sets. The 112 positive protein sequence set contributed to 89 in training set and 23 test sets, while 119 negative protein sequences were distributed into 95 and 24 training and test sets, respectively (Table 1.2).

Table 1.2 Distribution of data set into training, test, positive and negative classes

Sr. no.	Data set	Positive	Negative	Total
1	Training	90	96	186
2	Test	22	23	45
	Total	112	119	

1.3.3 Feature extraction

In machine learning, feature extraction is the key step to extract all the features, which forms a part of training and testing data sets rather than the whole sequences. The extracted features after screening are directly utilized in the future for model building. For feature extraction, all important methods available till date were considered. A comprehensive package Protr from R language and environment was used to generate descriptors (features). This package is capable of producing important features from protein sequences and has been used exclusively in the field of Bioinformatics and Chemogenomics research [42]. This package includes important descriptors consisting of amino acid composition, autocorrelation, CTD, conjoint traid, quasi-sequence order, pseudo amino acid composition, and profile-based descriptors, etc. The prediction model of our tool used features as described in Table 1.3.

1. **Amino acid composition:** It is described as portion of individual amino acid type within all amino acids in protein. This descriptor is defined by these extractAAC(), extractDC() and extractTC() functions as:

$$f(r) = \frac{N_r}{N} \quad r = 1, 2, \dots, 20$$

Where N is the length of the sequence and N_r is the number of the amino acid type.

2. **Autocorrelation:** This descriptor describes the properties of amino acids along the sequence. The implication of this descriptor is given by these functions extractMoreauBroto(), extractMoran() and extractGeary() and is defined as:

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \quad d = 1, 2, \dots, \text{nlag}$$

Where d is referred as the lag of the autocorrelation.

3. **Composition/Transition/Distribution:** Composition is described as the percent of each encoded class in the sequence. The transition from one class to another in an encoded class is referred as the percent with which one class is followed by the other or second class is followed by first. The distribution is stated as the distribution of each attribute in the sequence. This descriptor calculates extractCTDC(), extractCTDT(), extractCTDD() functions. Where a composition is represented by below equation:

$$C_r = \frac{n_r}{n} \quad r = 1, 2, 3$$

Where n_r is the number of amino acid type and n is the length of the sequence.

Transition is defined as:

$$T_{rs} = \frac{n_{rs} + n_{sr}}{N - 1}, \quad rs = '12', '13', '23'$$

Where N is the length of the sequence and n_{rs} , n_{sr} is the numbers of dipeptide encoded as “rs” and “sr” in the sequence.

4. **Conjoint triad descriptors:** It considers the properties of one amino acid and its two adjacent amino acids combined as a unit. It generates more than twenty discrete numbers to represent a protein. The descriptor uses descriptor is `extractCTriad()` function and is described as:

$$d_i = \frac{f_i - \min\{f_1, f_2, \dots, f_{343}\}}{\max\{f_1, f_2, \dots, f_{343}\}}$$

The numerical value of d_i for each protein ranges from 0 to 1; f_i stands for frequencies of occurrence of three amino acids as a unit in sequence.

5. **Quasi-sequence-order descriptors:** Physicochemical distance between amino acids, a set of sequence-order coupling numbers were introduced to reflect the sequence order effect, or in a rigorous term, the quasi-sequence-order effect. They are derived from the distance matrix between the 20 amino acids. It uses functions `extractQSO()` and `extractSOCN()` and is calculated by:

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{\maxlag} \tau_d} \quad r = 1, 2, 3, \dots, 20$$

Where w is a weighting factor ($w = 0:1$) and f_r is the normalized occurrence for amino acid type i .

6. **Pseudo-amino acid composition (PAAC):** They generate $20 + \lambda$ discrete numbers to represent a protein. PAAC descriptors are calculated using these functions `extractAPAAC()` and `extractPAAC()` and defined as:

$$H_1(i) = \frac{H_1^0(i) - \frac{1}{20} \sum_{i=1}^{20} H_1^0(i)}{\sqrt{\frac{\sum_{i=1}^{20} [H_1^0(i) - \frac{1}{20} \sum_{i=1}^{20} H_1^0(i)]^2}{20}}}$$

7. **Generalized BLOSUM and PAM matrix-derived:** This descriptor calculates the generalized BLOSUM and PAM matrix-derived descriptors using Blosum and PAM matrices for the 20 amino acids. The function used was extractBLOSUM() by this descriptor.
8. **Scales-based descriptors:** It has 20+ classes of Molecular Descriptors, uses these function extractDescScales() and extractPropScales().

Table 1.3 Protr package generated descriptors

Descriptors groups	Descriptors	Number
Amino acid compositions	Amino acid composition	20
	Dipeptide composition	400
	Tripeptide composition	8000
Autocorrelation	Moran	240
	Geary	240
CTD	Composition	21
	Transition	21
	Distribution	105
Conjoint triad	Conjoint triad	343
Quasi-sequence-order	Sequence-order-coupling number	60
	Quasi-sequence-order descriptors	100
Pseudo-amino acid composition	Type 1	50
	Type 2	80

1.3.4 Data preprocessing

A combination of linearly and nonlinearly features are received through feature extraction technique. After extraction of all the features from the protein sequences, there is a need

to pre-process them before feeding them to LibSVM. In order to produce good accuracy, scaling was performed for all the features between 0 and 1.

1.3.5 Feature selection

Feature selection is one of the important steps to retain important features only and exclude the rest of features. The included features form original feature space, which reduces the dimensions of feature vector and increases prediction accuracy. The selected subsets of features are used for model building. It provides an edge by reducing the complexity of the model and checks overfitting of data. For feature selection, Fselect script from LibSVM package was used. It calculates F-score, which is one of the best-optimized method with respect to SVM package. This script can be downloaded from <https://www.csie.ntu.edu.tw/~cjlin/LibSVMtools/fselect/fselect.py> [43].

$$F_{(i)} \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_{(i)})^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

Where the training vector is x_k , $k=1, \dots, m$, whereas n_+ and n_- stands for positive and negative instances; $\bar{x}_i, \bar{x}_i^{(+)}, \bar{x}_i^{(-)}$ are defined as average of the i^{th} feature of the whole, positive, and negative data sets.

The higher the F-score is, it is more probable that the feature is more discriminative.

1.3.6 Model generation

Support vector machine (SVM) is a supervised machine learning technique widely used to resolve regression and classification problems. This machine learning technique can analyze data and infer patterns for solving problems. It uses linear and non-linear methods (using kernel function) for pattern analysis. The kernel helps to map the input space to high dimensional space. The two key parameters, namely c ('tolerance of misprediction') and γ ('shape of separating hyperplane') have to be optimized manually for better performance. In the current research work, freely accessible LibSVM package was used to generate SVM model. The Radial basis function (RBF) kernel which has gained immense popularity in this domain was used. The LibSVM package classifies the data into two sets, namely training dataset (dataset used to build the model) and this model will be tested on the test dataset to evaluate its performance. Once model was built, it was implemented through a tool named as DRPPP (Disease resistance plant protein predictor).

1.3.7 Model evaluations

The evaluation of the proposed model is a mandatory step to confirm the performance of the binary classifier. Listed below are some of the popular evaluation techniques followed for the model building:

1.3.7.1 Evaluation through test datasets

The Support Vector Machine (SVMs) based model was evaluated using test datasets (sequences not involved in building models). Evaluation is an important step to keep a check on overfitting of data. Therefore, to evaluate the model test dataset consisted of 45 sequences, including 22 positive sequences and 23 negative sequences. The negative dataset was normalized between 0 and 1 to match with the model requirement. To predict better accuracy than the existing NBSPred server the test dataset was also checked on it.

1.3.7.2 10-fold cross-validation

Cross-validation is the technique to evaluate the model by deciding the training set into k subsets, where $k-1$ subsets are used in training of data and a single subset is used to test the data. It is a popular model validation technique to test the model on independent data sets taken from training dataset. In 10 fold cross-validation k is 10 and training data was divided into ten equal subsets, followed by training of nine subsets and test on remaining one subset and at last mean accuracy was calculated [44].

1.3.7.3 Performance measures

The predicted support vector machine model was validated using statistical measures. The statistical measures used to test the performance of the model were:

1. **Sensitivity:** It is defined as the percentage of R proteins that are correctly predicted as R proteins.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100$$

2. **Specificity:** It is the percentage of non- R protein that are correctly predicted.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} * 100$$

3. **Accuracy:** It is the percentage of correct predictions carried out from the total number of predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} * 100$$

4. **Matthew's correlation coefficient (MCC):** It is a combined measure of both sensitivity and specificity measures.

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FN} * \text{FP}}{\sqrt{(\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TP} + \text{FP}) * (\text{TN} + \text{FN})}}$$

In above case TP, TN, FP and FN represents the number of true positives, true negative, false positive and false negative, respectively.

1.3.8 ROC curve

ROC stands for Receiver operating characteristic curve it is a popular graphical plot to access the performance of binary classifier. Where AUC stands for the area under the curve it used to sum up the performance in a single number. It predicts the trade off between sensitivity and 1-specificity.

1.3.9 Implementation of model as tool and website development

Once the model was built, its implementation was carried out using a graphical user interface tool DRPPP. Perl/Tk was used to create a Graphical user interface (GUI). The codes/scripts used for development of the tool were:

1. Code 1: R_Predict.pl

```
#!/usr/bin/perl
use strict;
use warnings;
use Tk;
my $current_file;
my $rb_value1;
my $mw = MainWindow->new;
    $mw->minsize( qw(800 300));
    $mw->configure(-title=>'DRPPP',-background=>'grey');
    $mw->geometry('+100+300');
my $menu=$mw->Frame(-relief=>'groove',-borderwidth=> 3,-background=>'#90CA77')->pack(-side=>'top',-fill=>'x');
my $file_mb=$menu->Menubutton(-text =>'File',-underline => 0,-background=>'#90CA77',-font => [-weight => 'bold',-size => 10])->pack(-side=>'left');
```

```
$file_mb->command(-label => "Close",-accelerator => 'Ctrl-w',-underline => 0,-  
command => \&exit,);
```

```
my $edit=$menu->Menubutton(-text =>'Edit',-underline => 0,-background=>#90CA77',-  
font => [-weight => 'bold',-size=>10])->pack(-side=>'left');  
$edit->command(-label=>'Input file',-activebackground=>#9E3B33',-  
command=>\&organismtypee1);  
$edit->separator;  
$edit->command(-label => 'Preferences',-font => [-weight => 'bold',-size=>10],-  
activebackground=>#9E3B33');
```

```
my $window=$menu->Menubutton(-text =>'Window',-underline => 0,-  
background=>#90CA77',-font => [-weight => 'bold',-size=>10])->pack(-side=>'left');  
$window->command(-label=>'Website',-activebackground=>#9E3B33',-  
command=>\&web);
```

```
sub web{  
my $url = 'http://14.139.240.55/NGS/download.php';  
open_browser($url);  
}
```

```
my $help=$menu->Menubutton(-text =>'Help',-underline => 0,-  
background=>#90CA77',-font => [-weight => 'bold',-size=>10])->pack(-side=>'left');  
$help->command(-label => 'Version',-font => [-weight => 'bold',-size=>10], -command  
=>\&button22,-activebackground=>#9E3B33');  
$help->separator;  
#$help->command(-label => 'About',-activebackground=>#9E3B33');
```

```
$help->command(-label => 'About',-font => [-weight => 'bold',-size=>10], -command  
=>\&button21,-activebackground=>#9E3B33');
```

```
sub button22{  
my $yesno_button22 = $mw->messageBox(-message => "DRPPP Verion 1.0",-type =>  
"OK", -icon => "info");  
}
```

```
sub button21{  
my $yesno_button21 = $mw->messageBox(-message => "The tool aims to provide the  
scientific community with a novel machine learning method for prediction of disease  
resistance proteins in plants.",-type => "OK", -icon => "info");  
}
```



```

my $lb1=$mw->Label(-text => "DRPPP",-font => [-size => 50,-weight => 'bold'])-
>pack(-anchor => "nw");

my $lb2=$mw->Label(-text => "DRPPP stands for Disease resistance Protein prediction
in plants.The model predicted in this study is implemented as a free standalone tool
DRPPP to predict R-genes.",-font => [-size => 7,-weight => 'bold']->pack(-anchor =>
"nw");

my $fr2=$mw->Frame(-relief=>'groove',-borderwidth=> 3,-background=>'lightblue')-
>pack(-side=>'top',-fill=>'x');
my $lb4=$fr2->Label(-text =>'Upload Input file in fasta format',-
background=>'lightblue',-foreground=>'black',-font => [-weight => 'bold',-size=>20])-
>pack(-side=>'left');
my $button2= $fr2->Button(-text=>'Browse...',-command =>\&organismtypee1,-
background=>'grey')->pack(-side=>'right',-anchor => 'se');

my $lb5=$mw->Label(-text => "Note:Submit Transcriptome file (protein sequences) in
fasta format.",-font => [-size => 7,-weight => 'bold',])->pack(-side=>'bottom');

my $fr3=$mw->Frame(-relief=>'groove',-borderwidth=> 3,-background=>'lightcyan')-
>pack(-side=>'top',-fill=>'x');

my $button3= $fr3->Button( -text=> ' Help ',-command => \&button23, -relief =>
'raised',-background=>'grey')->pack(-side=>'left', -expand => 1);
my $button4= $fr3->Button( -text=> ' Run ',-command => \&button24, -relief =>
'raised',-background=>'grey')->pack(-side=>'left', -expand => 1);
my $button5= $fr3->Button( -text=> ' Quit ',-command => sub { exit })-> pack();

sub organismtypee1 {
    my @types =
    (["All files", '*'],
    ["Text files", [qw/.txt/]],
    ["Fasta files", [qw/.fasta/]],
    ["Fasta files", [qw/.fa/]],
    );
    $current_file= $mw->getOpenFile(-filetypes => \@types);
    print "$current_file\n";
    my @w=split(/\//,$current_file);
    open (F2 , ">./address.txt")or die "file not found";
    print F2 "$w[-1]\n";
    close(F2);
    #my $filename = 'scriptforblast.pl';

```

```

        #open(my $fh, '<:encoding(UTF-8)', $filename) or die "Could not open file
'$filename' $!";
        #my $row = <$fh>;
        #print "heelo$row";
        # $row =~ s/home/varun/mlpal/f.txt/$current_file/g;
        #write_file($filename, $row);
    }

sub button23 {
    my $yesno_button33 = $mw->messageBox(-message => "This
application uses in-house developed scripts and predicts disease resistant vrs non-disease
resistance proteins in plants transcriptomes/proteomes.It accepts input file in fasta
format containing protein/peptide sequences.",-type => "OK", -icon => "info");
}

sub button24 {
    print "$current_file\n";
    print "The tool is currently running please wait....";
    system("perl scriptforblast.pl $current_file >out1_seq_feature.LibSVM");
    system("perl extrct.pl out1_seq_feature.LibSVM
>out2_seq_feature_preprocessed.LibSVM");
    system("./LibSVM-3.20/svm-scale -l -l -u 1 -s range3
out2_seq_feature_preprocessed.LibSVM >
out3_seq_feature_preprocessed_scaled.LibSVM");
    system("perl grepse.pl out3_seq_feature_preprocessed_scaled.LibSVM
>out4_seq_feature_preprocessed_scaled_select.LibSVM");
    system("./LibSVM-3.20/svm-predict
out4_seq_feature_preprocessed_scaled_select.LibSVM
training_set_feature_preprocessed_scaled_selected.LibSVM.model
outu_svm_predict.LibSVM >/dev/null");
    system("grep \">\" $current_file > outu2");
    my @l = `cat outu2`;
    my @l2 = `cat outu_svm_predict.LibSVM`;
    open(F1, ">Result.xls");
    for(my$i=0;$i<@l;$i++)
    {
        if($l2[$i]==1)
        {
            print F1 "Predicted as DR Protein\t${l[$i]";
        }else
        {
            print F1 "Not Predicted as DR Protein\t${l[$i]";
        }
    }
}

```

```

print "\n The tool has predicted the results please refer to Result.xls file";
close(F1);
}
sub button47 {
    my $address;
    open(F1, "./address.txt") or die "file not find";
    while(<F1>)
    {
        chomp($_);
        $address=$_;
    }
    system("./LibSVM/svm-predict $address ./drprotein1.LibSVM
./Output/" . $address . ".txt");
}

MainLoop;

```

2. Code 2: scriptforblast.pl

```

my $file = $ARGV[0];

open FH,$file;
my $file1;
my @query = <FH>;
my $i = 0;
my $fastal;
my $fastal1;
my $liness;
my $count =0;
close (FH);

foreach my $xyz (@query)
{

    if ($xyz =~ />/ && $count == 0)
    { $count = 1;
        $liness = $xyz;

        $i++;
        next;
    }
    if ($xyz !~ />/)
    {
        $liness = "$liness$xyz";
        next;
    }
    if ($xyz =~ />/ && $count == 1)
    {
        # $file1 = "file_ $i.fasta";
        $file1 = "f.txt";
    }
}

```

```

        open (FH1,">$file1");
        print FH1 "$liness";
        close (FH1);
        Rrun();
        $liness = $xyz;
        $i++;
        next;
    }
}
#$file1 = "file_$.fasta";
$file1 = "f.txt";

        open (FH1,">$file1");
        print FH1 "$liness";
        close (FH1);

Rrun();

sub Rrun
{

    my $R = Statistics::R->new();
    $R->startR ;
    $R -> send('library(protr);');
    for(my$i=0;$i<1;$i++)
    {

        $R -> send('setwd("/home/JUIT/test/");');
        $R -> send('LL =
readFASTA(system.file("../..../..../home/JUIT/test/f.txt", package = "protr"))[[1]]);');
        $R -> send('y = extractAAC(LL);');
        $R -> send('write.table(y, "/home/JUIT/test/g0.txt");');
        $R -> send('y = extractDC(LL);');
        $R -> send('write.table(y, "/home/JUIT/test/g1.txt");');
        $R -> send('y = extractTC(LL);');
        $R -> send('write.table(y, "/home/JUIT/test/g2.txt");');
        $R -> send('y = extractMoreauBroto(LL);');
        $R -> send('write.table(y, "/home/JUIT/test/g3.txt");');
        $R -> send('y = extractMoran(LL);');
        $R -> send('write.table(y, "/home/JUIT/test/g4.txt");');
        $R -> send('y = extractGeary(LL);');
        $R -> send('write.table(y, "/home/JUIT/test/g5.txt");');
        $R -> send('y = extractCTDC(LL);');
        $R -> send('write.table(y, "/home/JUIT/test/g6.txt");');
        $R -> send('y = extractCTDT(LL);');
        $R -> send('write.table(y, "/home/JUIT/test/g7.txt");');
        $R -> send('y = extractCTDD(LL);');
        $R -> send('write.table(y, "/home/JUIT/test/g8.txt");');
        $R -> send('y = extractCTriad(LL);');
    }
}

```

```

$R -> send('write.table(y, "/home/JUIT/test/g9.txt");');
$R -> send('y = extractAPAAC(LL);');
$R -> send('write.table(y, "/home/JUIT/test/g10.txt");');
$R -> send('y = extractPAAC(LL);');
$R -> send('write.table(y, "/home/JUIT/test/g11.txt");');
$R -> send('y = extractQSO(LL);');
$R -> send('write.table(y, "/home/JUIT/test/g12.txt");');
$R -> send('y = extractSOCN(LL);');
$R -> send('write.table(y, "/home/JUIT/test/g13.txt");');
$R -> send('y = extractBLOSUM(LL, submat = "AABLOSUM62", k = 5,
lag = 7, scale = TRUE, silent = TRUE);');
$R -> send('write.table(y, "/home/JUIT/test/g14.txt");');
$R -> send('y = extractDescScales(LL, propmat = "AATopo", index =
c(37:41, 43:47), pc = 5, lag = 7, silent = FALSE);');
$R -> send('write.table(y, "/home/JUIT/test/g15.txt");');
#$R -> send('y = extractPropScales(LL, index = c(160:165, 258:296), pc =
5, lag = 7, silent = FALSE);');
#$R -> send('write.table(y, "/home/JUIT/test/g16.txt");');

#$R -> send('LL = x;');
#$R -> send('y <- rbind(extractAAC(LL),extractAAC(LL))');
#$R -> send('tprops = AATopo[,c(37:41,43:47)];');
#$R -> send('AAidxmat = t(na.omit(as.matrix(AAindex[, 7:26])));');
#$R -> send('y = extractMDSScales(LL, propmat = "tprops", k = 5, lag =
7, silent = FALSE);');
#$R -> send('y = extractScales(LL, propmat = "AAidxmat", pc = 5, lag =
7, silent = FALSE);');
#$R -> send('y = extractFAScales(LL, propmat = "tprops", factors = 5, lag
= 7, silent = FALSE);');
#$R -> send('write.table(y, "/home/JUIT/test/g1.txt");');

my $file1 = "/home/JUIT/test/g1.txt";
my @l = `awk {'print \$2'} g0.txt g1.txt g2.txt g3.txt g4.txt g5.txt g6.txt
g7.txt g8.txt g9.txt g10.txt g11.txt g12.txt g13.txt g14.txt g15.txt`;
my @l1 = `awk {'print \$1'} g0.txt g1.txt g2.txt g3.txt g4.txt g5.txt g6.txt
g7.txt g8.txt g9.txt g10.txt g11.txt g12.txt g13.txt g14.txt g15.txt`;
#my @l = `awk {'print \$2'} g1.txt g0.txt`;
system ("rm -rf g*.txt");
my $cunter = 1;
for(my$i=0;$i<@l;$i++)
{
    if($l[$i] ne "\n")
    {
        chomp $l[$i];
        print "$cunter:$l[$i] ";
        $cunter++;
    }
}

}print "\n";

```

```

        open FH,$file1;

        close FH;
        #system ("rm g1.txt");
    }
    $R->stopR() ;
}

```

3. Code 3: extrct.pl

```

#!/usr/bin/perl -w
use strict;
open FH,$ARGV[0];
my @file = <FH>;
close FH;

#

for(my$j=0;$j<@file;$j++)
{
    my @temp = split (/s+/, $file[$j]);
    if($j<90){print "1 ";}else{print "0 ";}
#    print"$temp[0] ";
    my $counter=1;
    for(my$k=0;$k<@temp;$k++)
    {
        if($temp[$k])
        { chomp $temp[$k];
          my @temp = split(/:/,$temp[$k]);
          $temp[1] =~ s/NA/0/;
          my $rounded = sprintf("%.3f",$temp[1]);
          print"$counter:$rounded ";
          $counter++;
        }else{
          print"$counter:0 ";
          $counter++;}
    }
    print "\n";
}

```

4. Code 4: grepse.pl

```

#!/usr/bin/perl -w
use strict;
open FH,$ARGV[0];
my @file = <FH>;
close FH;
my @selfet = (9641, 9721, 9891, 9902, 9905, 9912, 9909, 9916, 9897, 9896, 9917, 9904,
9913, 9900, 9892, 9908, 9920, 9894, 9919, 9898, 9640, 9907, 9901, 9899, 9918, 9906,

```

9911, 9910, 9915, 9903, 9895, 9893, 9633, 9882, 9861, 9875, 9720, 9887, 9872, 9886, 9879, 9883, 9867, 9914, 9870, 9864, 9878, 9880, 9874, 9862, 9881, 9889, 9871, 9866, 11, 9890, 9876, 9885, 9877, 9868, 9888, 9865, 9873, 9869, 9863, 9884, 9646, 9713, 9635, 9644, 9632, 9715, 9638, 9634, 9712, 9636, 9714, 9724, 9726, 9642, 9716, 9718, 9637, 9722, 9647, 9163, 9649, 5023, 9169, 9648, 9650, 9643, 9145, 9639, 9729, 9717, 9727, 9921, 223, 9719, 9728, 9730, 9723, 9361, 7152, 9541, 9771, 9645, 9151, 1840, 9396, 9415, 8421, 236, 9373, 9355, 9652, 9651, 9166, 331, 9172, 131, 9833, 251, 9346, 9450, 9959, 8422, 9926, 9359, 232, 9725, 9494, 9142, 9803, 7023, 8901, 226, 9180, 9148, 9542, 9394, 8663, 8661, 4876, 9731, 8452, 9273, 9274, 9275, 9276, 9277, 91, 9496, 9362, 9248, 9249, 9250, 9251, 9252, 9300, 9444, 9352, 4475, 9263, 9264, 9265, 9266, 9267, 8432, 8902, 8903, 9445, 9654, 8932, 1157, 8631, 9569, 9144, 9398, 4555, 9347, 6639, 9368, 229, 3023, 2632, 9422, 8662, 7652, 9495, 9349, 9622, 9761, 4471, 239, 6623, 9395, 8512, 8692, 1, 224, 4802, 7836, 9150, 9353, 9348, 8462, 9375, 10098, 4476, 7931, 9835, 9360, 9544, 8442, 9931, 8585, 9631, 10103, 9258, 9259, 9260, 9261, 9262, 4583, 4782, 9808, 4469, 9443, 10138, 9228, 9229, 9230, 9231, 9232, 4491, 9860, 8451, 9996, 9233, 9234, 9235, 9236, 9237, 151, 9857, 6603, 9198, 9199, 9200, 9201, 9202, 4465, 10042, 1836, 7036, 9377, 9111, 9367, 161, 2076, 9513, 9203, 9204, 9205, 9206, 9207, 9317, 9856, 9543, 9223, 9224, 9225, 9226, 9227, 9657, 9193, 9194, 9195, 9196, 9197, 9183, 9184, 9185, 9186, 9187, 9278, 9279, 9280, 9281, 9282, 8931, 8942, 9143, 4757, 1351, 9936, 10182, 51, 9356, 9446, 9218, 9219, 9220, 9221, 9222, 8151, 9213, 9214, 9215, 9216, 9217, 4792, 9448, 9830, 6650, 9366, 6683, 9826, 8528, 6751, 8871, 8427, 9457, 6226, 6531, 8465, 6624, 9653, 9711, 9188, 9189, 9190, 9191, 9192, 4970, 5029, 5131, 2711, 8912, 4791, 9555, 9401, 8843, 1256, 8519, 4643, 9208, 9209, 9210, 9211, 9212, 9238, 9239, 9240, 9241, 9242, 8495, 8693, 6519, 227, 9170, 1531, 9310, 9838, 9600, 2636, 9997, 10032, 9674, 9268, 9269, 9270, 9271, 9272, 103, 3611, 10211, 1036, 9082, 9859, 4472, 9372, 9827, 1691, 8523, 8702, 9382, 9345, 4489, 2622, 9364, 4443, 9768, 9951, 8691, 9492, 9969, 5114, 5086, 1264, 6711, 2332, 9162, 8589, 9104, 7151, 7404, 4632, 4582, 1374, 420, 5380, 4736, 9822, 9858, 10238, 4466, 1825, 8852, 8842, 2651, 9083, 3096, 4992, 351, 9493, 4655, 10228, 1032, 8424, 4515, 15, 8522, 415, 9464, 3823, 6289, 9147, 6629, 9819, 9253, 9254, 9255, 9256, 9257, 8223, 2609, 65, 8672, 8521, 4663, 4063, 818, 2352, 9167, 2616, 1131, 4543, 9780, 1514, 9680, 4863, 9370, 9548, 4463, 28, 5024, 9540, 2696, 9805, 10084, 8933, 4086, 4331, 9477, 9658, 6450, 9243, 9244, 9245, 9246, 9247, 9376, 7604, 9354, 8894, 2631, 9485, 4594, 143, 408, 8439, 9499, 4587, 9324, 1240, 8457, 8518, 304, 9092, 8582, 9849, 678, 2874, 4596, 321, 4735, 8525, 4651, 2126, 1023, 8396, 76, 1439, 119, 4452, 9825, 175, 3570, 8612, 8593, 1830, 8499, 9829, 10168, 10099, 9363, 9497, 4903, 4722, 9408, 9579, 9733, 10005, 8602, 10143, 10009, 4056, 41, 3531, 2062, 5491, 2623, 1839, 4462, 7326, 9475, 10062, 9165, 63, 10128, 222, 8945, 6229, 965, 2312, 6631, 3271, 4742, 8864, 9350, 8624, 6103, 8634, 9491, 2633, 6412, 3, 6710, 10246, 10263, 8423, 2360, 8555, 1747, 4108, 8609, 8992, 3331, 9562, 9528, 9775, 9303, 8571, 9828, 10007, 8503, 9806, 10002, 6206, 10179, 9582, 4567, 10176, 1094, 3032, 9381, 9462, 3451, 9529, 1451, 4726, 1027, 8, 10116, 8524, 9664, 4597, 1530, 9852, 9777, 8469, 8572, 1534, 9843, 6236, 1367, 1932, 4991, 8435, 9369, 1205, 1302, 3847, 375, 21, 4202, 9732, 9358, 10100, 9161, 8438, 1976, 8459, 10046, 5111, 1025, 9384, 9296, 8999, 10047, 9365, 3499, 9134, 1262, 5026, 9483, 10242, 8752, 8638, 4068, 8907, 10040, 8849, 9520, 9342, 9294, 10129, 8211, 9673, 2629, 3291, 8579, 3871, 4835, 5376, 8615, 9397, 10201, 2051, 6281, 307, 4951, 1829, 2180, 8855, 3365, 2462, 36, 9655, 4556, 2771, 4933, 9292, 9089, 1590, 7211, 8539, 2530, 2086, 6447, 4727, 4486, 9436, 8527, 9999, 3026, 4723, 9571, 5764, 8851, 4051, 9351, 9341, 5701, 9607, 1730, 851, 4603, 8446, 2200, 3545, 3996, 4368, 4878, 7232, 7423, 7871, 8205, 9288, 3951, 3140, 9534, 2872, 1596, 348, 9840, 9283,

9284, 9285, 9286, 9287, 4035, 4126, 10044, 10141, 10006, 7029, 8603, 8577, 4332, 8873, 9988, 5522, 8429, 5784, 4467, 4685, 216, 5032, 72, 317, 242, 10185, 5036, 71, 9447, 4786, 3880, 5194, 1810, 8922, 312, 20, 6851, 4522, 4528, 9458, 3351, 8937, 9074, 4088, 359, 10136, 6951, 9957, 2211, 4636, 9338, 10196, 4967, 3131, 9839, 10082, 22, 7851, 8251, 9809, 8481, 13, 1793, 10004, 8774, 4538, 9380, 4728, 4146, 10139, 4457, 8515, 9597, 1376, 9970, 9813, 6108, 9003, 8649, 5262, 5923, 17, 1436, 1819, 1872, 2955, 3245, 3405, 3849, 4117, 7362, 8245, 8270, 4656, 8705, 9855, 469, 7832, 9173, 931, 711, 2570, 10262, 9980, 5286, 4682, 9591, 1510, 3430, 9481, 720, 10223, 206, 1770, 5256, 5272, 4687, 8684, 1910, 27, 9834, 407, 4626, 10186, 9611, 4537, 8586, 8581, 9754, 4336, 9823, 10213, 9171, 8559, 6276, 6212, 2551, 5049, 1286, 3851, 1668, 5084, 9176, 8491, 9618, 5827, 8226, 6637, 8904, 1331, 6296, 9095, 6931, 5743, 2448, 9466, 872, 1984, 1276, 9478, 9779, 904, 5379, 1316, 9002, 1372, 1548, 3872, 3011, 1733, 3915, 9322, 8202, 9113, 9836, 1396, 395, 301, 4532, 8529, 3429, 8563, 2863, 9663, 7432, 2004, 8472, 5218, 9734, 2891, 2241, 7207, 1268, 2800, 9471, 10183, 8695, 9091, 231, 9853, 1831, 1043, 6300, 8445, 5926, 6290, 4647, 9614, 1253, 1564, 1572, 1873, 1985, 2054, 2077, 2383, 3080, 3094, 3125, 3825, 5419, 6181, 6205, 6293, 7265, 7322, 7884, 230, 8281, 5105, 4864, 6511, 9146, 9152, 8654, 302, 3816, 8264, 162, 5326, 8939, 4308, 10158, 7051, 9545, 4731, 8775, 2714, 311, 10245, 9593, 10151, 5703, 8834, 9992, 6571, 4257, 1576, 3350, 5021, 10231, 1223, 10252, 6325, 8454, 5727, 315, 5816, 2046, 3013, 6870, 7612, 9426, 9820, 4612, 3500, 4232, 6460, 8492, 9506, 1051, 8703, 5950, 4548, 9993, 4739, 1033, 1408, 8482, 8924, 4744, 9344, 9298, 9504, 8551, 9476, 4371, 2639, 2230, 4010, 5270, 5791, 3900, 10068, 5766, 5004, 1769, 2237, 3991, 6906, 1371, 5927, 4511, 9811, 3515, 4564, 8660, 8131, 3221, 3408, 1663, 8667, 2591, 8682, 4494, 191, 7422, 3766, 6237, 8588, 1022, 8641, 2351, 7027, 6690, 8517, 9578, 9420, 10212, 4980, 9603, 9626, 8875, 4239, 97, 211, 381, 9114, 4450, 2705, 1334, 8288, 4531, 3321, 4557, 4161, 10113, 3456, 6731, 1287, 1303, 1320, 1449, 1678, 2204, 2259, 2417, 2478, 2485, 2929, 3176, 3735, 4154, 4233, 5292, 5343, 5365, 6287, 6877, 8005, 8119, 10145, 6456, 8556, 8172, 9977, 277, 5022, 9289, 8541, 10266, 417, 9841, 8526, 10024, 8514, 4227, 5524, 6134, 296, 1823, 2730, 8592, 34, 8408, 7751, 9297, 2236, 9773, 2555, 6570, 78, 9584, 10153, 9175, 6007, 8587, 8759, 8680, 455, 2072, 2862, 3626, 3711, 261, 8825, 2683, 3931, 1509, 2512, 2790, 1790, 3123, 8664, 4112, 9044, 6266, 1485, 2151, 2498, 3025, 4245, 5327, 5752, 5992, 6765, 7018, 7424, 4606, 4600, 9552, 4046, 10175, 8574, 9065, 7636, 2624, 1305, 10095, 9688, 5663, 8511, 4642, 4652, 1546, 1916, 4275, 4593, 5514, 8098, 10108, 1111, 2731, 856, 1394, 2140, 2296, 5242, 7556, 10173, 8360, 4971, 9984, 5432, 6607, 9001, 8496, 6569, 9075, 9602, 30, 8725, 6348, 8560, 8402, 1513, 4009, 9428, 6463, 4737, 6486, 3836, 401, 10222, 4340, 3424, 2531, 2096, 9762, 4224, 3029, 548, 173, 8384, 643, 10110, 7412, 4151, 9118, 4251, 1257, 1310, 2396, 4790, 8138, 8484, 8763, 4604, 5836, 9699, 10052, 9615, 5740, 4322, 9687, 86, 9105, 5843, 6297, 4492, 8497, 212, 8351, 9848, 2908, 1265, 9051, 7687, 6468, 9008, 10059, 4527, 9014, 391, 2634, 9693, 1403, 9374, 286, 4225, 6954, 478, 1538, 1703, 1989, 2117, 2130, 2141, 2382, 2759, 2867, 2984, 3279, 3345, 4361, 4519, 5354, 5462, 5765, 6218, 6549, 6693, 6903, 7235, 7586, 8107, 8345, 4131, 6686, 4215, 4328, 6268, 8578, 10000, 3201, 8324, 4128, 8498, 6696, 4987, 9391, 8537, 3832, 8611, 1732, 8538, 9629, 202, 2446, 4354, 16, 9804, 1272, 6574, 1875, 4530, 8532, 262, 1423, 8853, 2206, 3333, 6921, 8256, 7839, 8616, 3367, 950, 3334, 33, 6335, 9435, 9554, 7160, 3426, 7735, 1517, 2313, 3956, 4762, 152, 6676, 4464, 8520, 441, 10137, 5536, 10035, 4488, 10208, 8633, 4444, 9922, 721, 1504, 3237, 4797, 2854, 5260, 1124, 203, 10, 1484, 1845, 2877, 3483, 5058, 5737, 5885, 6053, 6253, 6547, 7455, 8012, 6406, 6216, 8697, 8920, 9679, 8076, 8375, 1445, 3338, 5268, 7141, 2766, 2567, 4007, 10181, 4733, 7032, 9656, 37, 6976, 2465, 4374, 5725, 7276, 1030, 6806, 6221, 147, 5935, 4568, 2029, 4287, 1085, 2765, 4641,

8874, 5622, 10045, 9807, 6122, 3391, 10120, 9763, 3416, 6271, 954, 1105, 2899, 7420, 7809, 9404, 6215, 8287, 9538, 5823, 9440, 2224, 8420, 422, 9178, 8597, 9412, 303, 1640, 9451, 912, 2303, 6915, 6995, 5944, 9340, 134, 8984, 5001, 8735, 8146, 10093, 8516, 8542, 4648, 10165, 8303, 9290, 1440, 2508, 2741, 2770, 4954, 6764, 8259, 8294, 9556, 10261, 4862, 9700, 306, 8773, 8591, 10178, 8943, 8919, 9743, 10013, 8530, 314, 1096, 2090, 3433, 9694, 345, 4734, 9461, 1211, 561, 5804, 8811, 4961, 8557, 6563, 4004, 1317, 6748, 2084, 6467, 494, 9924, 4875, 8544, 9129, 1551, 10268, 3112, 4910, 8079, 1970, 355, 681, 4711, 2561, 9966, 9599, 6961, 283, 4329, 5991, 2686, 8020, 2482, 4392, 10226, 8762, 10065, 4535, 160, 5931, 6691, 9577, 4692, 9416, 6522, 3296, 48, 8558, 5829, 9994, 6652, 5251, 3684, 1245, 2386, 1063, 4732, 3228, 9156, 6209, 10169, 8464, 316, 2246, 4830, 10147, 4542, 1426, 9093, 9670, 8975, 9081, 3357, 117, 5932, 1505, 2755, 1911, 539, 204, 1455, 5876, 6106, 7664, 7740, 6872, 2990, 2921, 370, 8949, 2697, 8935, 978, 8813, 458, 907, 994, 1385, 1475, 1483, 1561, 1945, 2106, 2172, 2258, 2271, 2343, 2795, 2869, 2905, 3188, 3199, 3583, 3633, 3699, 3712, 3795, 4293, 4669, 5093, 5335, 5466, 5547, 5564, 5566, 5567, 5669, 5749, 5862, 6074, 6192, 6418, 6901, 7133, 7325, 7447, 7462, 7571, 7599, 7610, 7635, 7678, 7733, 7838, 7970, 8153, 3234, 8540, 2519, 4461, 10144, 19, 280, 168, 276, 3362, 9818, 9566, 8601, 1912, 608, 8430, 3406, 2082, 8835, 9387, 6524, 460, 6254, 1620, 275, 3000, 10015, 1864, 2321, 3695, 7616, 6698, 6934, 9302, 802, 4490, 8139, 2620, 8216, 6719, 9697, 4589, 8744, 9758, 5643, 3651, 10078, 9623, 8643, 7406, 5731, 6436, 9182, 6712, 8312, 8804, 8051, 1759, 1944, 10236, 5044, 8688, 8567, 4938, 4675, 7246, 7256, 1012, 3621, 2626, 8120, 9386, 1311, 8699, 7108, 8426, 8086, 508, 3662, 5297, 6507, 6944, 1343, 553, 1258, 1308, 1413, 1629, 1706, 2074, 2381, 2699, 3082, 3874, 4187, 4879, 5274, 5614, 7337, 7769, 4143, 5820, 8865, 1234, 716, 10166, 6604, 8998, 8768, 10022, 7124, 9052, 8468, 8249, 1737, 4089, 2322, 218, 9140, 9947, 8870, 821, 9821, 6200, 4136, 4327, 4121, 9601, 8714, 9962, 10170, 6670, 8674, 9115, 4041, 2876, 4631, 10041, 1133, 1430, 32, 5818, 835, 9088, 3551, 10010, 8458, 8471, 8501, 5687, 5792, 6330, 3145, 917, 1862, 2131, 4178, 4738, 8856, 5657, 3751, 73, 1532, 2094, 170, 4192, 5672, 7083, 23, 8479, 8208, 4226, 3111, 2438, 3910, 5403, 5459, 8199, 9844, 3613, 2610, 9316, 4884, 98, 1414, 9045, 5121, 3522, 384, 9328, 10171, 6651, 7017, 483, 7136, 8622, 9605, 1755, 3473, 8453, 3817, 9473, 9062, 171, 7200, 7916, 9985, 10126, 9050, 8311, 1092, 38, 2116, 2663, 5328, 1766, 2095, 2605, 10114, 5948, 7090, 80, 6612, 8156, 4746, 9560, 8934, 433, 6551, 431, 8857, 8840, 6801, 6508, 4401, 3834, 3323, 6840, 2431, 8841, 1666, 310, 10127, 9735, 1671, 2751, 7210, 9669, 3246, 1735, 5934, 9531, 328, 2656, 8331, 1296, 4366, 8009, 5619, 6401, 1454, 10221, 440, 3631, 4504, 1527, 1635, 6342, 7690, 6224, 9330, 8500, 4430, 571, 9532, 3839, 4493, 5956, 8547, 9164, 532, 1136, 108, 8489, 814, 1249, 8418, 2786, 3036, 10087, 4031, 6141, 9097, 534, 6, 10216, 9334, 6241, 3610, 9110, 1422, 8761, 9810, 1723, 8290, 178, 5691, 4431, 6634, 7812, 6886, 5351, 5266, 8822, 8552, 4460, 9053, 8965, 3697, 9080, 7528, 901, 1283, 1419, 1458, 1729, 2070, 2128, 2227, 5571, 8045, 10248, 8868, 9414, 624, 4216, 9101, 740, 6655, 6104, 1591, 5470, 461, 9430, 9972, 9108, 8665, 2310, 6622, 9594, 4148, 4235, 4477, 2816, 6262, 6485, 9514, 9332, 1014, 6703, 5416, 7833, 8810, 6160, 6115, 10092, 9490, 104, 651, 8513, 7675, 6017, 7053, 9596, 6491, 4521, 3332, 8022, 3502, 6023, 246, 3517, 7731, 8460, 1535, 962, 10161, 7094, 2739, 7995, 1601, 10057, 3028, 333, 5621, 9433, 9306, 2851, 9935, 3440, 3442, 5651, 4919, 5889, 1284, 4021, 9685, 9157, 3534, 1236, 8549, 9845, 4026, 8961, 10131, 8486, 3439, 9946, 10177, 1960, 172, 3361, 5763, 2174, 3142, 8224, 6808, 930, 4242, 1128, 8679, 8824, 138, 8608, 726, 10218, 8507, 6317, 620, 4539, 1319, 1966, 3132, 3343, 4115, 5739, 6898, 7087, 8812, 5754, 8355, 8175, 4595, 9149, 9572, 5013, 10011, 4180, 1959, 503, 594, 758, 898, 905, 999, 1073, 1118, 1129, 1301, 1353, 1489, 1878, 2099, 2182, 2209, 2253, 2267, 2284, 2317, 2361, 2376, 2400, 2669,

2907, 2919, 2945, 2993, 3058, 3074, 3075, 3187, 3385, 3677, 3693, 3744, 3853, 3864, 3876, 3883, 3932, 3947, 3966, 3994, 4005, 4138, 4509, 4779, 4998, 5198, 5199, 5293, 5329, 5389, 5448, 5464, 5476, 5485, 5565, 5579, 5597, 5607, 5653, 5698, 5797, 6105, 6120, 6129, 6379, 6673, 6769, 6858, 7278, 7289, 7340, 7392, 7407, 7418, 7471, 7485, 7543, 7582, 7665, 7704, 7743, 7793, 7818, 7945, 7961, 7971, 7994, 8373, 8227, 4841, 6975, 10064, 2068, 2404, 8436, 2710, 6449, 3005, 4982, 9570, 9851, 5548, 2091, 9455, 9707, 7651, 31, 1120, 7591, 5252, 1863, 1693, 7642, 2541, 1417, 372, 2736, 6625, 4304, 6795, 8640, 1834, 857, 6302, 8543, 4710, 8878, 5842, 5428, 8607, 6564, 5623, 6037, 7042, 225, 8673, 3310, 115, 3308, 5692, 8985, 6956, 5896, 324, 180, 5208, 4686, 9832, 279, 4708, 3050, 9778, 4536, 3505, 1071, 8487, 4122, 9539, 4310, 237, 710, 4218, 4644, 1390, 732, 2521, 201, 2471, 2726, 4387, 2435, 807, 8428, 8890, 4305, 10091, 5620, 8077, 2176, 4264, 6169, 6529, 3368, 5175, 6606, 7513, 6056, 9460, 8580, 8494, 5240, 6223, 1981, 8900, 1263, 2841, 3608, 5291, 3476, 4517, 42, 8232, 4246, 1405, 1438, 6546, 714, 1683, 3826, 1920, 6715, 2252, 1433, 7130, 8709, 10152, 837, 1005, 1062, 1176, 1282, 1563, 1815, 1961, 2135, 2250, 2325, 2458, 2589, 3105, 3320, 3646, 3921, 4189, 4549, 5312, 5406, 5427, 5783, 5794, 5947, 6376, 6669, 6919, 7295, 7524, 7850, 8145, 8546, 421, 1953, 8795, 4844, 5451, 6451, 309, 6965, 10025, 1639, 8307, 5060, 2184, 7527, 7545, 6820, 6746, 3121, 9087, 62, 8415, 8441, 425, 8861, 9953, 6474, 3501, 4524, 6871, 9469, 619, 7154, 2612, 1299, 2897, 8776, 815, 4729, 8074, 472, 5, 903, 2502, 2925, 82, 9005, 9508, 5601, 366, 4747, 7355, 5195, 6610, 6065, 1281, 2987, 2503, 8889, 8241, 736, 8952, 8177, 6981, 1292, 560, 9141, 9160, 9096, 8015, 4547, 6633, 3846, 4076, 6828, 9321, 1000, 7081, 4248, 1336, 3537, 8055, 1076, 6013, 8166, 7275, 1271, 4891, 9929, 193, 423, 10043, 2148, 2798, 4258, 6545, 7304, 10149, 7021, 1942, 9949, 9568, 637, 1621, 6887, 7888, 2104, 1495, 2688, 5666, 824, 9048, 8449, 4859, 968, 6986, 7198, 6427, 10191, 9974, 8174, 10250, 2840, 8476, 7915, 9078, 8030, 5047, 5683, 3293, 9986, 4928, 141, 2526, 7834, 5146, 707, 3732, 578, 784, 9518, 6752);

#

```
for(my$j=0;$j<@file;$j++)
{
    my @temp = split (/s+/, $file[$j]);
    print "$temp[0] ";
    #my $counter=1;
    for(my$k=1;$k<@temp;$k++)
    {
        my @tempp = split(/:/, $temp[$k]);
        my $counter = 0;
        for (my$w=0;$w<@selfet;$w++)
        {
            $counter = 0;
            if($tempp[0]==$selfet[$w])
            {
                $counter = 1;
                last;
            }
        }
        if($counter){
```

```
        print"$temp[$k] ";  
    #     $counter++;  
        }  
    }  
    print "\n";  
}
```

1.4 Results and Discussion

Development of disease resistance has always been one of the major objectives for any crop improvement program and early identification of *R* genes can be a major step towards achieving this goal. In total, 112 positives and 119 negative sequences were collected from PRGdb and protein database maintained by the NCBI (National Center for Biotechnology Information), respectively. The detailed description of the distribution into training and test datasets is represented in Table 1.2. Total data were divided into 80% for training and about 20% for testing the trained SVM model. A total of 10,270 features were generated using 16 different methods. These features were pre-processed to adjust NA values. Finally, all these features were scaled between 0 and 1 value. In-house developed Perl script was used to extract features from all the training and test datasets. Databases and tools such as PRGdb and DRAGO are available in the literature, but they rely on similarity or domain prediction methods only, neglecting the relevance of low similarity proteins and also hindering the identification of new or novel *R* genes, which cannot be predicted through these tools.

1.4.1 Models parameter

1.4.1.1 Models parameter and accuracy

The kernel function was optimized to obtain best *c* and gamma values. Essential parameters such as *c* and gamma were tuned to 2 and 70 values. 10-fold cross-validation was performed on the training dataset using the radial basis function (RBF) kernel.

The kernel function was optimized using *c* and gamma function and their values were tuned as large *C* can give you low bias and high variance, whereas, small *C* can provide you higher bias and lower variance. The gamma was also optimized as small gamma will give you low bias and high variance while a large gamma will give you higher bias and low variance.

1.4.1.2 Performance of the support vector machine model using statistics

The numbers of predicted sequences from TP, TN, FP and FN were 18, 23, 0 and 4, respectively from a total of 45 test datasets. The performance of the module is provided in Table 1.4.

Table 1.4 Performance of support vector machine model on dataset

Sr. no.	Model	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
1	SVM	81.82	100	91.11	0.83

The high sensitivity obtained from the model indicated that 81.82% of *R* proteins are correctly predicted as *R* protein. The specificity came out to be 100%, which indicates that the model has correctly identified all non-*R* proteins as non-*R* proteins. Accuracy of 91.11% indicated that the model is 91.11% efficient for predicting correct prediction from the total number of predictions. MCC (Matthews correlation coefficient) indicated the combined effect of both sensitivity and specificity, it is high (0.83) indicating that the model used for binary classification is 83% correct.

1.4.1.3 10-fold cross validation in the training dataset

Ten-fold cross validation was performed to optimize SVM parameter using the radial basis function (RBF) kernel. The kernel function was then optimized to obtain the best C and γ corresponding to the highest values of sensitivity, specificity and accuracy. Training dataset cross validation accuracy was achieved to be 85.4839%. The high percentage of ten-fold cross validation indicates that the results of statistical analysis will generalize to a good extent on an independent data set.

1.4.1.4 Receiver Operating Characteristic (ROC) plot

The Receiver Operating Characteristic (ROC) curve was used to show the trade-off between true positive rate (sensitivity) and false positive rate (specificity) for all possible values (Fig. 1.6). The area under the curve (AUC) was observed to be 0.8708 proving efficacy of the model.

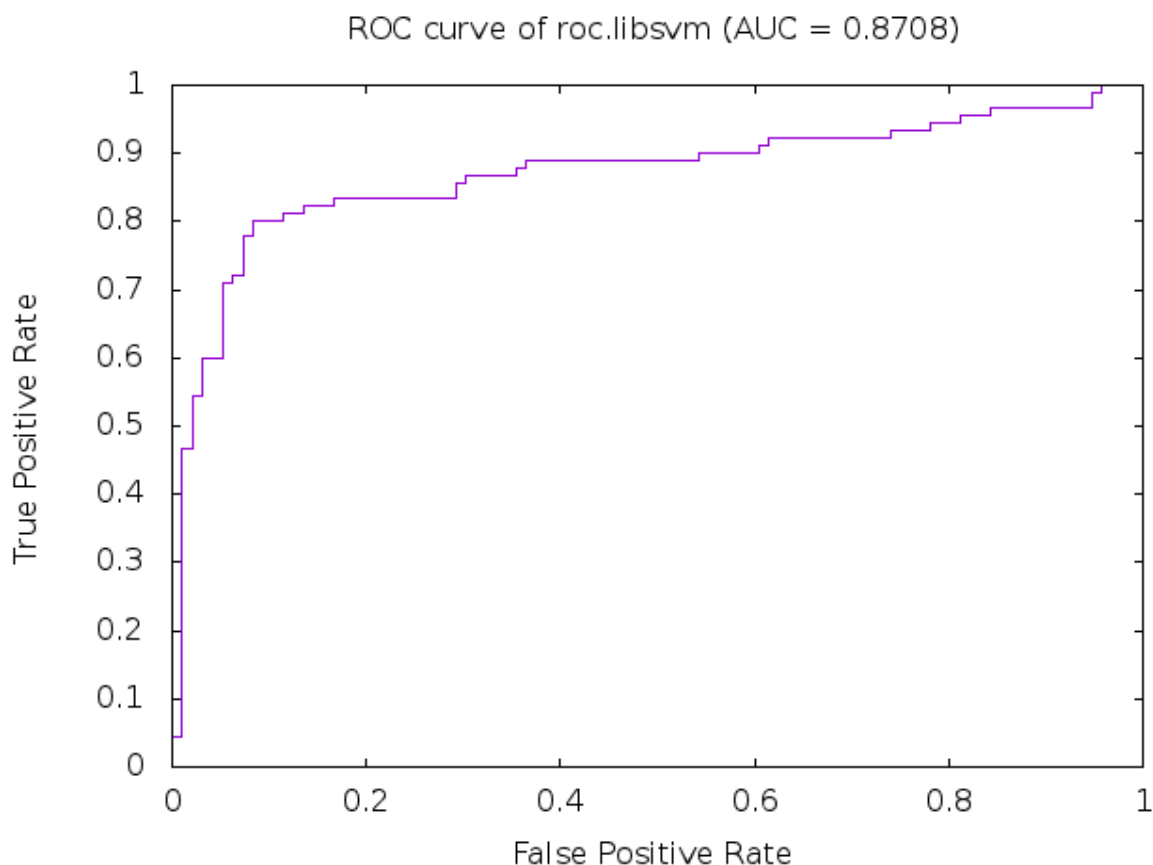


Fig. 1.6 ROC Curve for Binary SVM classifier: ROC plot depicts relative trade-offs between true positive and false positives.

1.4.2 Evaluation on test dataset

The test dataset of DRPPP was run on NBSPred, DRPPP detected 91.11 % of the sequences as *R* protein, whereas NBSPred detected 88.88% indicating more exploration potential of DRPPP for predicting *R* proteins. More exploration potential suggests that besides compositional frequencies descriptors (used by NBSPred) there are also other descriptors needed to efficiently discriminate between *R* and non-*R* proteins.

1.4.3 Web implementation

The SVM based model presented in the study is implemented as a freely available standalone tool ‘DRPPP’ to predict *R* proteins. The interface of the tool is simple and adaptable, which works by browsing an input file and thereby running it directly (Fig. 1.7). The tool is can be downloaded from the <http://14.139.240.55/NGS/download.php> website, the website is designed using PHP and HTML in combination with custom Java and Perl programs. Remaining all the scripts used in this study are coded in Perl programming language. Specific instructions are provided in the read me file. DRPPP is

hosted on a DELL PowerEdge™ T410 server with 16 cores 2.67 GHz Intel® Xenon processors running on CentOS 6.5 based 64-bit operating system.

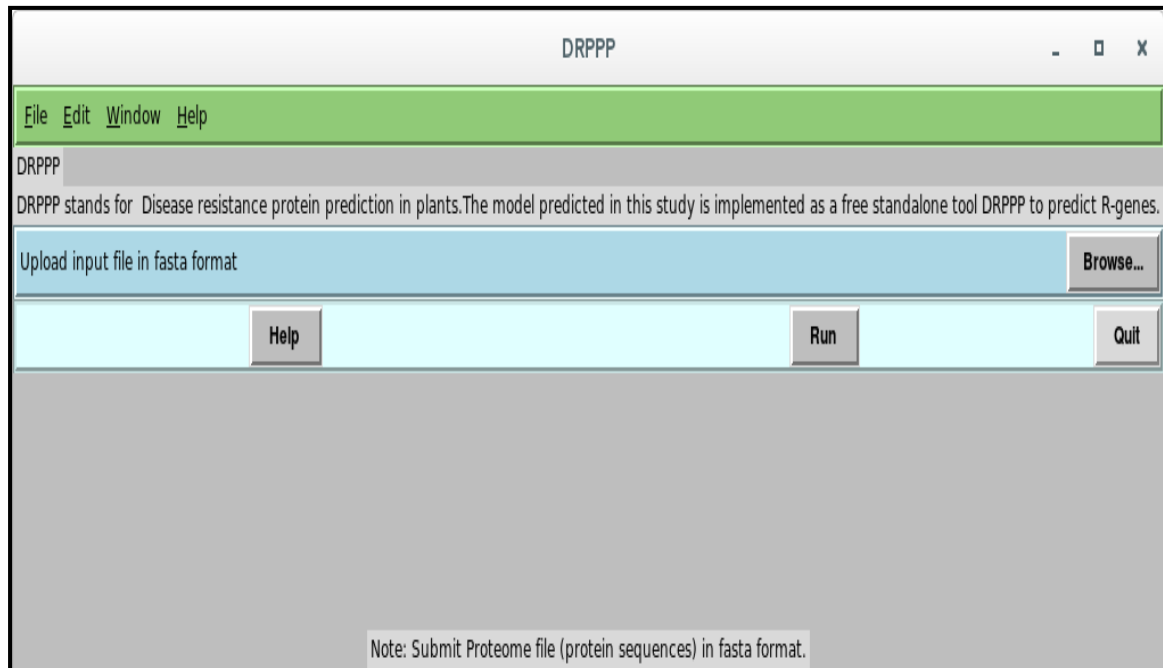


Fig. 1.7 Snapshot of DRPPP's interface

1.5 Conclusion

The influential features for predicting *R* genes are not known completely, therefore, a set of all important features was generated through feature extraction techniques. The developed method and implemented tool i.e. DRPPP can efficiently predict *R* protein with highest accuracy (91.11%) as compared to other existing tools. In future, DRPPP will be updated with the inclusion of more *R* genes discovered, thereby enhancing the efficiency of DRPPP.

REFERENCES

- [1] Z. Chen, A. P. Kloek, J. Boch, F. Katagiri, and B. N. Kunkel, "The *Pseudomonas syringae* avrRpt2 gene product promotes pathogen virulence from inside plant cells," *Molecular Plant-Microbe Interactions*, vol. 13, pp. 1312-1321, 2000.
- [2] J. L. Dangl and J. D. G. Jones, "Plant pathogens and integrated defence responses to infection," *Nature*, vol. 411, pp. 826-833, 2001.
- [3] C. Zipfel, G. Kunze, D. Chinchilla, A. Caniard, J. D. G. Jones, T. Boller, et al., "Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts *Agrobacterium*-mediated transformation," *Cell*, vol. 125, pp. 749-760, 2006.
- [4] L. Gomez-Gomez and T. Boller, "FLS2: An LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in *Arabidopsis*," *Molecular Cell*, vol. 5, pp. 1003-1011, 2000.
- [5] J. D. G. Jones and J. L. Dangl, "The plant immune system," *Nature*, vol. 444, pp. 323-329, 2006.
- [6] J. Ellis, P. Dodds, and T. Pryor, "Structure, function and evolution of plant disease resistance genes," *Current Opinion in Plant Biology*, vol. 3, pp. 278-284, 2000.
- [7] N. D. Young, "The genetic architecture of resistance," *Current Opinion in Plant Biology*, vol. 3, pp. 285-290, 2000.
- [8] D. A. Jones and J. D. G. Jones, "The role of leucine-rich repeat proteins in plant defences," *Advances in Botanical Research*, vol. 24, pp. 89-167, 1997.
- [9] W. I. L. Tameling, J. H. Vossen, M. Albrecht, T. Lengauer, J. A. Berden, M. A. Haring, et al., "Mutations in the NB-ARC domain of I-2 that impair ATP hydrolysis cause autoactivation," *Plant Physiology*, vol. 140, pp. 1233-1245, 2006.
- [10] G. B. Martin, A. J. Bogdanove, and G. Sessa, "Understanding the functions of plant disease resistance proteins," *Annual Review of Plant Biology*, vol. 54, pp. 23-61, 2003.
- [11] A. Sood, V. Jaiswal, S. K. Chanumolu, N. Malhotra, T. Pal, and R. S. Chauhan, "Mining whole genomes and transcriptomes of *Jatropha* (*Jatropha curcas*) and Castor bean (*Ricinus communis*) for NBS-LRR genes and defense response associated transcription factors," *Molecular Biology Reports*, vol. 41, pp. 7683-7695, 2014.

- [12] A. Bendahmane, G. Farnham, P. Moffett, and D. C. Baulcombe, "Constitutive gain-of-function mutants in a nucleotide binding site-leucine rich repeat protein encoded at the Rx locus of potato," *The Plant Journal*, vol. 32, pp. 195-204, 2002.
- [13] T. W. Traut, "The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide-binding sites," *European Journal of Biochemistry*, vol. 222, pp. 9-19, 1994.
- [14] W. Sanseverino, A. Hermoso, R. D'Alessandro, A. Vlasova, G. Andolfo, L. Frusciante, et al., "PRGdb 2.0: towards a community-based database model for the analysis of R-genes in plants," *Nucleic Acids Research*, vol. 41(Database issue), pp. D1167-D1171, 2010.
- [15] D. Marone, M. A. Russo, G. Laido, A. M. De Leonardis, and A. M. Mastrangelo, "Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses," *International Journal of Molecular Sciences*, vol. 14, pp. 7302-7326, 2013.
- [16] B. J. Yoon, "Hidden Markov models and their applications in biological sequence analysis," *Current Genomics*, vol. 10, pp. 402-415, 2009.
- [17] J. r. Schultz, R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork, "SMART: a web-based tool for the study of genetically mobile domains," *Nucleic Acids Research*, vol. 28, pp. 231-234, 2000.
- [18] S. K. Kushwaha, P. Chauhan, K. Hedlund, and D. Ahren, "NBSPred: a support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR prediction," *Bioinformatics*, vol. 32, pp. 1223- 1225, 2016.
- [19] B. Steuernagel, F. Jupe, K. Witek, J. D. G. Jones, and B. B. H. Wulff, "NLR-parser: rapid annotation of plant NLR complements," *Bioinformatics*, vol. 31, pp. 1665-1667, 2015.
- [20] K. L. S. Ng and S. K. Mishra, "De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures," *Bioinformatics*, vol. 23, pp. 1321-1330, 2007.
- [21] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol. 10, pp. 16, 2010.

- [22] M. J. M. Christenhusz and J. W. Byng, "The number of known plants species in the world and its annual increase," *Phytotaxa*, vol. 261, pp. 201-217, 2016.
- [23] G. S. Johal and S. P. Briggs, "Reductase activity encoded by the HM1 disease resistance gene in maize," *Science-New York then Washington*, vol. 258, pp. 985-985, 1992.
- [24] L. Gu, W. Si, L. Zhao, S. Yang, and X. Zhang, "Dynamic evolution of NBS-LRR genes in bread wheat and its progenitors," *Molecular Genetics and Genomics*, vol. 290, pp. 727-738, 2014.
- [25] S. Yang, Z. Feng, X. Zhang, K. Jiang, X. Jin, Y. Hang, et al., "Genome-wide investigation on the genetic variations of rice disease resistance genes," *Plant Molecular Biology*, vol. 62, pp. 181-193, 2006.
- [26] R. Velasco, A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, et al., "The genome of the domesticated apple (*Malus × domestica* Borkh.)," *Nature Genetics*, vol. 42, pp. 833-839, 2010.
- [27] S. Luo, Y. Zhang, Q. Hu, J. Chen, K. Li, C. Lu, et al., "Dynamic nucleotide-binding site and leucine-rich repeat-encoding genes in the grass family," *Plant Physiology*, vol. 159, pp. 197-210, 2012.
- [28] S. Tan and S. Wu, "Genome wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon*," *Comparative and Functional Genomics*, vol. 2012, pp. 1-12, 2012.
- [29] J. Shang, Y. Tao, X. Chen, Y. Zou, C. Lei, J. Wang, et al., "Identification of a new rice blast resistance gene, Pid3, by genomewide comparison of paired nucleotide-binding site-leucine-rich repeat genes and their pseudogene alleles between the two sequenced rice genomes," *Genetics*, vol. 182, pp. 1303-1311, 2009.
- [30] W. Sanseverino, G. Roma, M. De Simone, L. Faino, S. Melito, E. Stupka, et al., "PRGdb: a bioinformatics platform for plant resistance gene analysis," *Nucleic Acids Research*, vol. 38, pp. D814-D821, 2012.
- [31] A. Singh, R. Gupta, S. K. Saikia, A. Pant, and R. Pandey, "Diseases of medicinal and aromatic plants, their biological impact and management," *Plant Genetic Resources*, vol. 14, pp. 1-14, 2016.
- [32] S. E. Perfect, H. B. Hughes, R. J. O'Connell, and J. R. Green, "Colletotrichum: a model genus for studies on pathology and fungal-plant interactions," *Fungal Genetics and Biology*, vol. 27, pp. 186-198, 1999.

- [33] M. Koeck, A. R. Hardham, and P. N. Dodds, "The role of effectors of biotrophic and hemibiotrophic fungi in infection," *Cellular Microbiology*, vol. 13, pp. 1849-1857, 2011
- [34] O. C. Maloy and T. D. Murray, *Encyclopedia of Plant Pathology*: Wiley, 2001.
- [35] J. Yu, S. Tehrim, F. Zhang, C. Tong, J. Huang, X. Cheng, et al., "Genome-wide comparative analysis of NBS-encoding genes between *Brassica* species and *Arabidopsis thaliana*," *BMC Genomics*, vol. 15, pp. 3, 2014.
- [36] T. Zhou, Y. Wang, J. Q. Chen, H. Araki, Z. Jing, K. Jiang, et al., "Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes," *Molecular Genetics and Genomics*, vol. 271, pp. 402-415, 2004.
- [37] Q. Pan, J. Wendel, and R. Fluhr, "Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes," *Journal of Molecular Evolution*, vol. 50, pp. 203-213, 2000.
- [38] L. McHale, X. Tan, P. Koehl, and R. W. Michelmore, "Plant NBS-LRR proteins: adaptable guards," *Genome Biology*, vol. 7, pp. 212, 2006.
- [39] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, pp. 210-229, 1959.
- [40] S. B. Rice, G. Nenadic, and B. J. Stapley, "Mining protein function from text using term-based support vector machines," *BMC Bioinformatics*, vol. 6, pp. S22, 2005.
- [41] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, pp. 1658-1659, 2006.
- [42] N. Xiao, D.S. Cao, M.F. Zhu, and Q.S. Xu, "protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences," *Bioinformatics*, vol. 31, pp. 1857-1859, 2015.
- [43] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, pp. 27, 2013.
- [44] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, et al., "Application of high-dimensional feature selection: evaluation for genomic prediction in man," *Scientific Reports*, vol. 5, pp.1-12, 2015.

Computational analysis of NGS transcriptomes of medicinal herbs (*Aconitum heterophyllum* and *Swertia chirayita*)

Abstract

The *Aconitum heterophyllum* and *Swertia chirayita* are high altitude medicinal herbs native to North-Western Himalayas and possessing diverse pharmacological properties, thereby widely used in the treatment of diarrhea, malaria, cough, vomiting, cold, etc. Till date, there exists no information on genetic factors contributing to the biosynthesis and accumulation of secondary metabolites in these plant species. Therefore, comparative transcriptomes were generated for the two tissues root versus shoot of *Aconitum heterophyllum* and for two differential conditions i.e. photoautotrophic versus photoheterotrophic modes of nutrition of *Swertia chirayita*. These transcriptomes were generated to decipher important molecular components associated with the secondary metabolites biosynthesis in these plant species.

The paired-end (PE) Illumina sequencing technology generated 46,612,687 (22.1 GB paired-end data) and 28,777,415 (13.6 GB paired-end data) high-quality reads for root and shoot tissues of *Aconitum heterophyllum*, whereas 41,031,326 and 21,859,688 high quality (HQ) reads were observed for greenhouse(SCFG) and tissue cultured (SCTC) *Swertia chirayita* samples, respectively after quality filtering.

The Velvet pipeline was optimized to assemble these reads into 75,548 and 39,100 transcripts for root transcriptome and shoot transcriptome of *Aconitum heterophyllum*, whereas in case of *Swertia chirayita* 57,460 and 43,702 transcripts for greenhouse grown (SCFG) and tissue cultured (SCTC) plants were observed, respectively. Gene ontology analysis of root versus shoot transcriptomes of *Aconitum heterophyllum* assigned biological functions to 27,596 and 12,340 transcripts, respectively; while for *Swertia chirayita* 18,090 and 2,102 transcripts were annotated for greenhouse grown (SCFG) and tissue cultured plants. The clusters of orthologous group analysis, classified 16,604 and 9,398 assembled root and shoot transcripts of *Aconitum heterophyllum*, where in *Swertia* this number was found to be 12,826 and 9,565 for greenhouse grown (SCFG) and tissue cultured (SCTC) plants, respectively. FPKM (Fragments per kilobase of transcript

per million) approach was used to carry out large scale expression profiling of the mevalonate (MVA)/2-C-methyl-D-erythritol 4-phosphate (MEP, non-mevalonate) pathway genes revealing 4 genes *HDS* (1-hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate synthase), *HMGR* (3-hydroxy-3-methylglutaryl-CoA reductase), *MVK* (mevalonate kinase) and *MVDD* (mevalonate diphosphate decarboxylase) with higher expression in roots as compared to shoots of *Aconitum heterophyllum*, whereas in *Swertia* 9 genes (encoding HMGS, MVK, PMK, ISPD, ISPE, ISPF, IPPI, GDPS and MVDD) showed higher transcript abundance in SCFG compared to SCTC transcriptomes. Network connectivity diagrams were constructed for isoquinoline alkaloid biosynthesis pathway associated with secondary metabolism in root transcriptome of *A. heterophyllum*.

2.1 Introduction

Furthermore, in plants the importance of medicinal herbs can be derived from the fact that the demand for herbal medicines is estimated to increase up to US\$3 trillion by 2020 [1]. However, their commercial cultivation has been hampered due to lack of genome resources so as to take up genetic improvement programmes. Next-generation sequencing (NGS) has provided unprecedented opportunities for high throughput research on medicinal plants, whose genome/transcriptome datasets were still not available. This cost-effective technique has provided us with opportunities to explore genome, transcriptome, exome, small RNA and targeted DNA/RNA at a much rapid pace. The evolution of this field has enabled understanding of important biological processes through differential transcriptomics studies, phylogenomic analysis, etc. The medicinal herbs, *Aconitum heterophyllum* and *Swertia chirayita* have been extensively used in various herbal drug formulations resulting in their excessive utilization; thereby, placing them in high value endangered category. Keeping in view, the burgeoning amount of information produced by the transcriptomes, the current work was designed to perform computational analysis of the transcriptomes generated from different tissues, of *Aconitum heterophyllum* and *Swertia chirayita* and mapping transcripts to biosynthetic pathways contributing to biosynthesis and accumulation of secondary metabolites.

Aconitum heterophyllum (Atis) is a biennial herb from the Ranunculaceae family found in the North-Western Himalayas region at an altitude of 2400–3600 m. The nontoxic alkaloids like aconitine, atisine, heterophyllinine and hetidine are key components of its tuberous roots [2, 3] and is extensively used for the treatment of cough, cold, diarrhea, vomiting, etc. [4, 5] (Fig. 2.1). In fact, this is the only non-toxic species of the genus *Aconitum*. The over harvesting of this plant species has led to considerable depletion of its population, placing this species in list of ‘critically endangered species’ by International Union for Conservation of Nature and Natural Resources [6, 7]. The major constituent aconites including atisine is a well known marker compound of *A. heterophyllum*. The biosynthesis and accumulation of aconites is pursued through the MVA/MEP pathways recognized for isoprenoid production (Fig. 2.2).



Fig. 2.1 Mature tuberos roots of *A. heterophyllum* plant

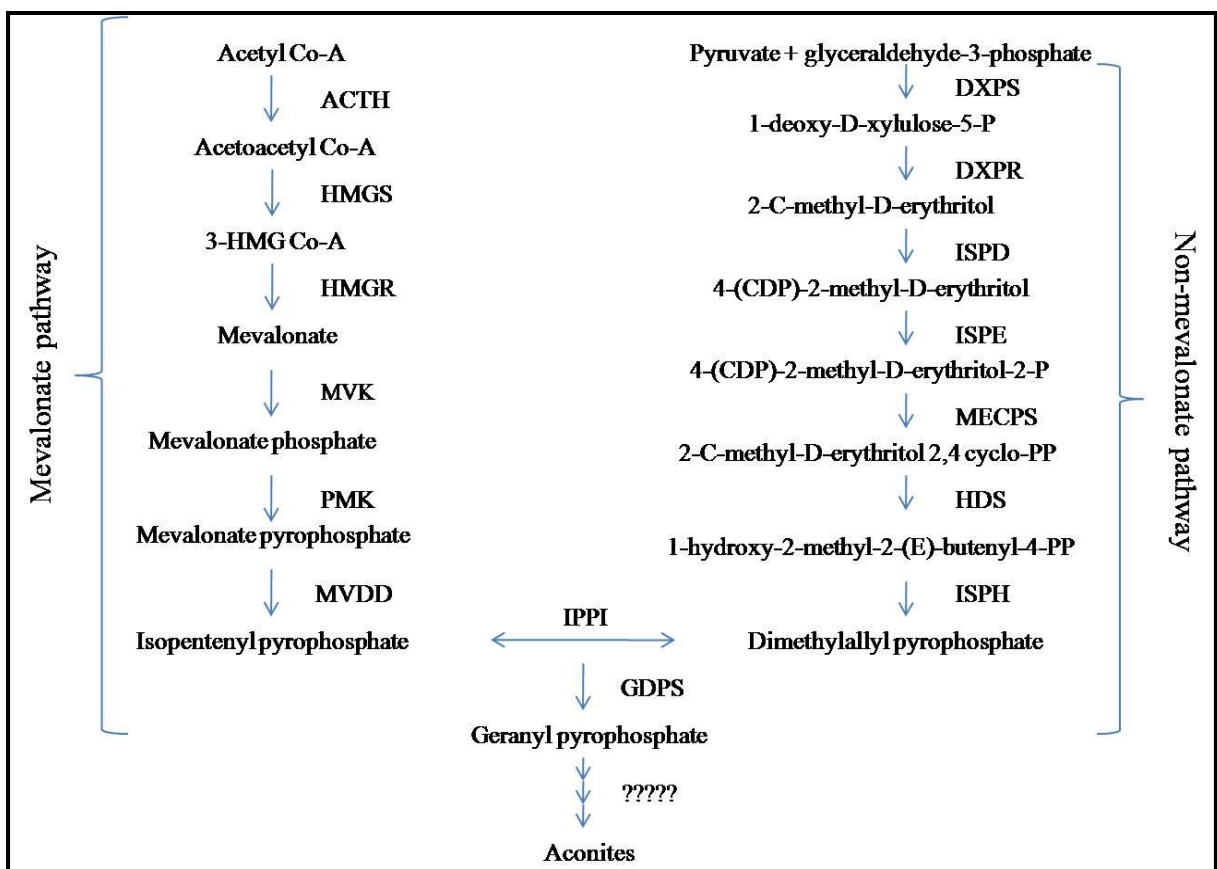


Fig. 2.2 A representative diagram showing different modules for alkaloids biosynthesis in *A. heterophyllum* [adapted from Rodríguez-Concepción et al. [8]]

Swertia chirayita (*S. chirayita*), commonly known as ‘Chirata’, is a pharmacological important medicinal herb endemic to the temperate Himalayas region (Kashmir to Bhutan, and Kashia hills) at an altitude of 1200-3000 m (Fig. 2.3). It possesses multifarious therapeutic values and belongs to Gentianaceae family. It is mainly valued for its pharmacological properties like antimalarial, antidiabetic, hepatoprotective, anticarcinogenic, anti-inflammatory, antioxidant, antiviral, antibacterial and antifungal activities. These properties are due to the presence of various secondary metabolites, including swertiamarin, mangiferin and amarogentin. Increased market demand have led to over exploitation of this medicinal herb and therefore, the plant is now placed in critically endangered category by New International Union for Conservation of Nature and Natural resources (IUCN) [9]. The endangered status of *S. chirayita* has reduced the availability of its raw material as well as secondary metabolites, thus compelling for the development of alternative routes for production of its biomass. However, the secondary metabolite content in *in vitro* grown plants is very low as compared to their natural habitat.



Fig. 2.3 Mature greenhouse grown *S. chirayita* plant

Due to lack of transcriptomic information for these species there exists a gap not only for discovering candidate genes involved in secondary metabolites production, but also for understanding the molecular basis of various biological processes. The genes involved in isoprenoid production through a combined biosynthetic route MVA/MEP route need to be identified and thereby employed to enhance the production of secondary metabolites (Fig. 2.2). Thus, the transcriptomes were generated for two tissues of *A. heterophyllum* and two tissues of *S. chirayita* plant species for the first time; and, thereby, computational transcriptome analyses were performed on all four tissues to decipher various molecular components responsible for secondary metabolism. The assembly, annotation and analyses was performed for root (AHSR) and shoot (AHSS) transcriptomes of *A. heterophyllum* and greenhouse grown (SCFG) and tissue cultured (SCTC) plant transcriptomes of *S. chirayita*.

Keeping in view the need for streamlining the annotation of transcriptome resources for secondary metabolism in medicinal herbs the present research was carried with the following objective:

Objective 2: Computational analysis of *Aconitum heterophyllum* and *Swertia chirayita* transcriptomes to unravel genes responsible for secondary metabolism

2.2 Review of Literature

2.2.1 DNA sequencing (First-generation technologies)

DNA sequencing is defined as the process of defining the correct order of the nucleotides present in a DNA molecule. In the early 1976–1977s, Allan Maxam and Walter Gilbert published widely accepted sequencing method based on nucleobase-specific partial chemical modification of DNA followed by chain breakage at specific nucleotides [10]. Later, the discovery of chain-terminating dideoxynucleotides sequencing method was a revolution by Frederick Sanger and colleagues, thereby defining new horizons in sequencing history. Further, this concept was improvised for the development of automated Sanger sequencing [11, 12].

Automated Sanger sequencing lead to the major landmark funded in 1990 for determining the complete base pairs, which make up human reference genome. It took decades to produce first draft sequence of human genome in 2001 [13] and finally in 2003 a more completed finished version was published [14]. The project costed more than three billion dollars and paved the way for development of high computing and advanced sophisticated algorithms/tool together with cheaper high-throughput sequencing techniques required to deal with such vast amount of data. This method was efficient, fast and simple; evolved from radioactive to dye labelling of nucleotides and using capillary electrophoresis. Whereas it also had some limitations, which included labour intensiveness, reagent cost and was time-consuming and thus involved major expenses.

Till this time a lot of other projects such as International HapMap Project and the prominent 1000 Genomes Project were launched and high demand for large scale sequencing and computational analysis was felt. All this triggered to produce a more efficient method, which could be capable of generating millions of sequences in a single run at an unprecedented pace, thereby leading to the production of next-generation sequencing (NGS) or second-generation sequencing methods.

2.2.2 Need for next generation sequencing technology

The “first-generation” sequencing technologies were further refined to more efficient technology named as next-generation sequencing (NGS). These systems are capable of generating millions of reads in a single run as compared to other traditional methods. The cost associated with the next generation sequencing is also reduced multiple times as compared to other sequencing techniques. The template consumed by them is in lower

amounts than the rest of the sequencing techniques. Today, this technology has unrivalled explosion not only in the field of biological sciences but has also aided medical scientists [15].

2.2.3 Next generation sequencing

Further, sequencing market saw a boom with vendors from different sequencing technologies competing and producing a dramatic rise of data output with falling cost. This unseen broad range of NGS applications made this technology as the method of the year 2007 [16]. These technologies generated millions of short sequence reads at a stretch, creating a bottleneck for large set of data storage, data backup and analysis, depending on hardware and labour intensiveness. In order to tackle such huge volume of data there is a need to develop more advanced and sophisticated algorithms, pipelines, tools with high computing capability. In order to process and analyze parallel computing with distributed clouds are needed to handle terabyte of data.

2.2.4 Transcriptome sequencing

Transcriptome is defined as the sum total of all RNA molecules that are being actively expressed at any given time by an organism. Thereby, the study of the transcriptome is stated as transcriptomics. Transcriptome provides insights of the genes that are being actively expressed, thereby identifying genes of interest, uncovering new molecular markers, gene expression analysis and performing comparative transcriptomics studies. The information of the genes that are being actively expressed at any given time and expression in different conditions can be used in pathway elucidation. Although becoming cheaper, transcriptome sequencing still remains an expensive endeavour.

2.2.5 Launching of NGS platforms

The beginning of NGS platforms era was seen with the beginning of 2000. During this year, MPSS Lynx Therapeutics (USA) Company started the beginning of NGS series, but soon this company was sold to American company Illumina.

The year 2004 saw a major boon in this field with the second company, namely 454 Life Sciences (Branford, CT, USA) launching pyrosequencing based method for high-throughput sequencing at its 454 SC. The reduction in the sequencing cost was observed to be as low as six times as compared to automated Sanger sequencing technology. During 2005-2006, the 454 signed the agreement with Roche and launched 454 GS 20 (Roche sequencing platform) together with its promotion, sales and distribution. In the

year 2007-2008, this platform was replaced with the GS FLX model to produce 400 Mbp. Later this model was upgraded to 454 GS-FLX+ Titanium sequencing platform producing capacity increased to 600 Mbp of sequence data [17, 18].

In 2005-2006, another company Solexa produced Genome Analyzer (GA) based on sequencing by synthesis (SBS) technique. In 2007, this company was purchased by the Illumina group. Their platform GAIIx produced 50 billion bases. After this, Illumina launched series of HiSeq platforms, including HiSeq 1000, HiSeq 1500, HiSeq 2000 and HiSeq 2500. The read length capacity using Illumina HiSeq 2500 also increased to 200 bp (capacity 120 billion bases (120 Bbp) of data produced in 27 hours). Later in 2011, Illumina also launched MiSeq platform with capacity of producing data raised to 1.5 Gbp per run in about 10 h [19].

In 2007, another leading company Applied Biosystems launched the first solid system with a capacity of 35bp and 3G data per run. Later in 2010, it also released SOLiD 5500 w and SOLiD 5500 xlw platforms.

During this era competition flourished and there were many other platforms launched such as Helicos sequencer became available in 2009 followed by Life Technologies Ion Torrent sequencer released in 2011 [20], Pacific Biosciences (Menlo Park, CA, USA) single molecule real-time (smrt) sequencer released in 2011 [21]. Currently, the Oxford Technologies Nanopore (Oxford, UK) is available from 2012–2013 [22]. The comparison leading NGS platform is given in Table 2.1.

Table 2.1 Comparison of leading NGS Platforms

Platform	Chemistry	Read Length	Run Time	Gb/Run	Website
454 GS FLX+ (Roche)	Pyro-sequencing	700	23 hrs.	0.7	http://www.454.com/
HiSeq (Illumina)	Reversible Terminator	2*100	2 days (rapid mode)	120 (rapid mode)	http://www.illumina.com/
SOLiD (Life)	Ligation	85	8 days	150	http://www.thermofisher.com/in/en/home/life-science/sequencing/next-generation-sequencing.html
PacBio RS	Real-time Sequencing	3000 (up to 15,000)	20 min	3	http://www.pacb.com/

2.2.6 Illumina sequencing

The initial contribution to the Illumina sequencing technology was made by Shankar Balasubramanian and David Klenerman in mid-1990s itself. Later in 1998, they formed Solexa company. Illumina sequencing technology utilizes sequencing by synthesis (SBS) chemistry capable of generating millions of reads with massively parallel sequencing fashion. It is competent to generate both single-read as well as paired-end libraries. The Genome Analyzer, first Solexa sequencer, was launched in 2006 and had a capacity of 1 gigabase (Gb) of data in a single run [19]. Later in 2007, Solexa was acquired for \$650 million by Illumina.

The generalized workflow of Illumina sequencing includes these basic steps:

1. **Library preparation** – Library preparation begins with fragmentation followed by repairing ends (add A overhangs) and adaptor ligation of the DNA or cDNA sample.

Otherwise, to increase efficiency fragmentation and ligation reactions steps are combined under tagmentation.

2. **Cluster generation** – Cluster generation is a process carried out by hybridization of template molecules onto the oligonucleotide-coated surface of the flow cell, thereby fragment amplification to generate clonal clusters. Cluster generation is followed by sequencing process.
3. **Sequencing** – The process to identify the correct order of nucleotides using SBS technology. Sequencing begins by hybridization of a sequencing primer followed by incorporation of single base at each cycle. All the four colours are detected by two lasers of total internal reflection fluorescence (TIFR).
4. **Data analysis** – After sequencing, single or paired end reads data analysis proceeds either with assembly or alignment of reads. Where assembly aims to assemble all the generated overlap fragments alignment deals with aligning the generated fragments to the reference genome.

Illumina Sequencers have eight flow cell lanes, where several samples can be loaded at the same time for analysis. Each lane has 100 tiles producing ~1.5 Gigabases per flow cell. After image, analysis and base calling the end output are files in Illumina's FASTQ format. These files can be further filtered to remove poor quality reads and generate high quality reads. These high quality reads can be used for detection of single nucleotide polymorphisms, insertion and deletion (indel), estimation of transcript abundance and more.

2.2.7 Computational tools/pipelines for data analysis

There are numerous tools developed for computational analysis of NGS datasets the majority of them categorized under quality filtering, assembly/alignment of sequence reads, *de novo* genome/transcriptome browsing and annotation. Various assembly tools such as Velvet, ABYSS, SOAPdenovo and SSAKE were popular during 2008-2010 [23-26].

Even today, more efficient assemblers are being produced commercially for aligning short-read alignments. Most NGS companies such as Roche, Illumina, SOLiD, etc., are providing their own software tools/pipelines compatible to the sequencing data provided by them. A more comprehensive review of software tools available in the market can be summarized as under.

2.2.8 Quality filtering

Alike to Sanger sequencing technique, the quality values or quality scores are also provided with NGS platforms, this value can reflect probability of incorrect bases. NGS platforms provide quality values or quality scores stating the likelihood that base call is incorrect. The Phred quality score denoted by Q is defined as $Q = -10\log_{10} P$, where P is base-calling error probabilities. For example, a Q20 value demonstrates the probability of incorrect base call is 1 in 100 and accuracy is 99% [27]. Different NGS platforms have different error profiles and, thereby, accuracy varies accordingly. Moreover, the range for the Phred quality scores varies for Sanger as compared to Illumina for Sanger Phred score ranges from 0 to 93, whereas for Illumina it ranges from -5 to 40 or from 0 to 40 (using ASCII 64–104 in fastq) based on the platform used [Understand Illumina link]. Whereas, in ABI Solid sequencing platforms, quality scores are assigned to each colour, but the range varies from range of Phred score i.e. from 0 to 45 [28]. For 454 pyrosequencing device, the quality values vary between 0 to 40. Discussion forums such as <http://seqanswers.com> are also popular in this much evolving field.

The Data filtering is an important step to remove low quality reads and improve the overall quality of the assembly. It is helpful in reducing the noise associated with the data. These days, many software packages are popular which includes prominent software's like FastQC [29] and Trimmomatic [30]. These softwares uses in-built codes to trim the adapters followed by removal of repeated reads so as to retain only high quality reads for further downstream analysis.

2.2.9 De novo assembly

Once the sequencing has been performed, there is a task to assemble all the fragmented reads as per the organism's chromosomes, which requires complex computations to be performed for combining million of reads. With the advent of NGS platforms the size of reads reduced, which outperformed various existing assemblers based on overlap graphs. Some of the important assemblers include ABySS, Velvet, SOAPdenovo, SSAKE and Trinity [31]. The comparison of popular next-generation sequencing assemblers is given in Table 2.2. The de Bruijn graphs (advance overlap graphs) are the famous directed graphs used by some of them due to the incapability of overlapping graphs to scale well with increasing reads. De Bruijn graphs break the reads into smaller number of subsequences denoted by k-mer. It follows the approach of converging non-intersecting paths into single nodes. For assembling k-mer is not a fixed parameter and it has to be

optimized for different values of k. Different values of k produce different assemblies according to k size. Different metrics are used to compare the quality of assembly's assembled using different k-mer. Two important metrics include:

1. **N50 value** - It is the smallest size contig with other larger contigs, which can cover 50% of the genome or transcriptome.
2. **Coverage** - It is the percentage of bases covered by assembled contigs in the reference genome. It can only be calculated if there exists reference genome.

Table 2.2 Comparison of next-generation sequencing Assemblers (single-end reads/paired -end) [adapted by Zhang et al. [32]]; “*” indicates any operating systems with Perl interpreter

Program	Algorithm	Program- ming Language	Running Platform	Required read length	Single/ Paired end	Input file format
SSAKE (V3.5)	Greedy- extension	Perl	*	25-36nt	Y,Y	Fasta/raw
VCAKE (V1.0)	Greedy- extension	Perl	*	<40nt	Y,N	Fasta/raw
Edena (V2.1.1)	OLC	C++	Win/linux	N/A	Y,N	Fasta/Fas tq
VELVET (V0.7.59)	De Bruijn	C	Linux/Ma cOS/Cyg win	N/A	Y,Y	Fasta/Fas tq
SOAPden ovo (V1.04)	De Bruijn	C	Linux/Ma cOS	N/A	Y,Y	Fasta/Fas tq

2.2.10 Functional annotation

Nr (non-redundant) protein sequence database is maintained by NCBI [33] and contains entries from various sources, including GenPept [33], Swissprot [34], PIR [35], PDB [36] and NCBI RefSeq [37]. In this comprehensive database, identical sequences are merged into a single sequence from both curated and non-curated databases. The condition to

merge two sequences is that they must have identical lengths and each individual residue in both the sequences must be same. Each different sequence is separated by a fasta signature (>) and common sequences are separated by control-A characters. This database is most popularly used for functional annotation by aligning the contigs or CDS to the non-redundant (nr) protein database of NCBI. Detailed annotation is essential for the determination of biological functions of newly sequenced transcripts, which is important for downstream biological analysis.

2.2.11 Gene ontology

The vast amount of genomic/transcriptomic data is being generated at a rapid pace and there is a need to allot well-defined vocabularies to these datasets. The most popular example of well-defined vocabularies is GO project. The project was originally launched in 1998 and included only three model organism database, namely FlyBase (*Drosophila*), the Mouse Genome Database (MGD) and the *Saccharomyces* Genome Database (SGD) [38]. The project has maintained common structure and style for maintaining information related to gene and gene product. The project includes three major functions: first, to develop and maintain ontology on the regular basis; secondly, annotating all gene and gene products; thirdly, development of various tools/software needed to maintain ontologies and provide easy access of these ontologies to customers.

The GO consortium basically classifies all the datasets into three primary classes, i.e. molecular functions, biological processes and cellular components among various species. The molecular functions refers to all the biochemical activities related to gene taking place at the molecular level such as binding activity and catalytic activity. It contains broad and narrower functional terms, such as transporter activity and toll receptor binding, respectively. A biological process is defined as biological events, which are contributed by gene or gene products. The examples of broad biological process consist of cell growth and maintenance, whereas specific process includes alpha-glucoside transport. The main difference between molecular functions and biological process is that biological process includes more than one step of molecular functions. The cellular components describe the location in the cell where the gene or gene product is active. Its example includes ribosome and proteasome indicating the place in the cell where gene product is found. Currently, there are around 48,410 GO id's stored in 'go-basic.obo' file which can be downloaded from the GO consortium website <http://www.geneontology.org/page/download-ontology>.

2.2.12 COG classification

The Clusters of Orthologous Groups (COG) database came into existence in 1997, it included comparison of proteins encoded from seven different genomes (five bacterial, one archaeal and one eukaryotic genomes) depicting 720 clusters of orthologous groups (COGs). COG gained due importance due to accelerating availability of molecular sequences, primarily the sequences of entire genomes. The prediction of orthologs is vital for predicting reliable functions of genes, especially in newly sequenced genomes. It is also important for establishing phylogenetic relationships as these can be deciphered only in orthologs. Every cluster of COG contains proteins or group of paralogs from a minimum of three lineages [39]. Later in 1998, sixth bacterial genome was included, enhancing the number from 720 to 860 clusters. In 2000, the group reported the increase to 2,091 COGs and included the proteins from 21 complete genomes. Initially, the COG database contained prokaryotic clusters (COGs) with single eukaryotic genome, but in 2003, 7 eukaryotic genomes were added to it, thereby increasing the overall tally to 1,38,458 proteins from 66 genomes [40]. The COG classifies all the sequences into 25 functional categories (Table 2.3).

Table 2.3 COG functional classification (Tatusov et al. 2003 [40])

Symbols	Functional Categories
“A”	“RNA processing and modification”
“B”	“Chromatin structure and dynamics”
“C”	“Energy production and conversion”
“D”	“Cell cycle control, cell division, chromosome partitioning”
“E”	“Amino acid transport and metabolism”
“F”	“Nucleotide transport and metabolism”
“G”	“Carbohydrate transport and metabolism”
“H”	“Coenzyme transport and metabolism”
“I”	“Lipid transport and metabolism”
“J”	“Translation, ribosomal structure and biogenesis”
“K”	“Transcription”
“L”	“Replication, recombination and repair”
“M”	“Cell wall/membrane/envelope biogenesis”
“N”	“Cell motility”
“O”	“Posttranslational modification, protein turnover, chaperones”
“P”	“Inorganic ion transport and metabolism”
“Q”	“Secondary metabolites biosynthesis, transport and catabolism”
“R”	“General function prediction only”
“S”	“Function unknown”
“T”	“Signal transduction mechanisms”
“U”	“Intracellular trafficking, secretion, and vesicular transport”

“V”	“Defense mechanisms”
“W”	“Extracellular structures”
“Y”	“Nuclear structure”
“Z”	“Cytoskeleton”

2.2.13 Domain search

Domains are the core conserved functional units of the proteins. These units form a three-dimensional compact structure and are responsible for a particular function. These can differ in length from 25 amino acids to 500 amino acids. A domain can consist of multiple motifs (super secondary structure).

2.2.14 *In silico* transcript abundance

The recent next generation sequencing technology has provided us with an advantage of deeply sequenced RNA-Seq data (mRNA sequencing). RNA-Seq is a prevailing technology, which has vanished microarrays [41]. This technology of massive parallel sequencing produces million of short reads from the cDNAs corresponding to the fragment of RNA. These short reads can be exploited for various transcriptomic analyses, such as the *de novo* transcript assembly [31, 42], transcript quantification, differential expression analysis [43, 44] and annotation using reference genes [45, 46].

The Rsem is one of the popular tools used for measuring *in silico* transcript abundance. It has an advantage of measuring abundance without the requirement of reference genome. It can directly take transcript sequences as an input for instance the transcripts produced by *de novo* transcriptome assembler [47]. The counts to *in silico* expression profiling can be measured in:

1. **RPKM** – It stands for ‘Reads Per Kilobase per Million mapped reads’. This level of measure is defined as:

$$RPKM = C/N * L$$

Where C stands for “Number of mappable reads on a feature (e.g. transcript, exon, etc.)”; N stands for “Total number of mappable reads (in millions)” and L stands for “Length of feature (in kb)”.

2. **FPKM**- It stands for ‘Fragments Per Kilobase of transcript per Million fragments mapped’. It is similar to RPKM but does not use read count rather uses transcripts fragment.

3. **TPM** – TPM stands for ‘Transcripts Per Million’. It is defined as :-

$$\text{TPM} = (10^6) * Z * (C/N * L)$$

Where additional Z parameter has been used to combat normalization factor [48].

2.2.15 Biological pathways and network connectivity diagrams

The biological pathways have been studied over several centuries. It may be defined as a series of biochemical reactions taking place in a cell [49]. Pathways are important as they lead to the ultimate end product of a series of biochemical reactions. There are multiple pathways simultaneously running in an organism. The basic pathways include metabolic pathway, genetic pathway and signal transduction pathway.

The metabolic studies were conducted from the thirteenth century by Ibn al-Nafis (1213-1288). He reported that “the body and its parts are in a continuous state of dissolution and nourishment, so they are inevitably undergoing permanent change”. Later in the 1940s, the complete glycolytic pathway was elucidated by Gustav Embden, Otto Meyerhof, Carl Neuberg, Jacob Parnas, Otto Warburg, Gerty Cori, and Carl Cori. Glycolysis is also known as the Embden-Meyerhof pathway. Currently there are many source databases such as, KEGG (Kyoto Encyclopedia of Genes and Genomes) [50], BioCyc (including its Tier 1 EcoCyc and MetaCyc databases, and its Tier 2 databases) [51], Reactome [52] and WikiPathways [53] (Table 2.4).

Table 2.4 The list of important existing pathway databases

Name of Database	Brief description	Availability
KEGG PATHWAYS	It consists of manually drawn pathway maps representing molecular interactions and reaction networks.	First made public in September 1997. url: http://www.genome.jp/KEGG/pathway.html
BioCyc	It consists of organism specific pathways/Genome databases. Each PGDB stores metabolic pathways information for a particular organism.	First made public in July 1999. url: http://biocyc.org/
Reactome	It is a manually curated and peer-reviewed database and majorly focuses on human pathways.	First made public in September 2008. url: http://www.reactome.org/
WikiPathways	It serves as a repository of biological pathways as pathway diagrams and as a platform for curating them.	First made public in 2008. url: http://www.wikipathways.org/
NCBI Biosystems database	The database is archival and each BioSystem record receives a unique identifier known as a bsid.	First made public in September 2008. url: http://www.ncbi.nlm.nih.gov/Biosystems/

2.2.15.1 KEGG pathway database

Whereas, Kyoto Encyclopedia of Genes and Genomes (KEGG) was originally developed in 1995, but later upgraded for KEGG drug and disease databases in 2005 and 2008 released versions. KEGG consists of sixteen databases (Table 2.5) [54]. At present, it consists of 509 pathway maps and 4,90,555 references (total) entries. At the backend, KEGG uses KAAS (KEGG automatic annotation server) for mapping transcripts to KEGG genes [55].

Table 2.5 The contents and database of KEGG

Database name	Content
“KEGG PATHWAY”	“KEGG pathway maps”
“KEGG BRITE”	“BRITE functional hierarchies”
“KEGG MODULE”	“KEGG modules of functional units”
“KEGG DISEASE”	“Information on human diseases”
“KEGG DRUG”	“Information related to drugs”
“KEGG ENVIRON”	“Crude drugs and health-related substances”
“KEGG ORTHOLOGY”	“KEGG Orthology groups”
“KEGG GENOME”	“Organisms with complete genomes”
“KEGG GENES”	“Gene catalogues in complete genomes”
“KEGG SSDB”	“Sequence similarity database for GENES”
“KEGG COMPOUND”	“Metabolites and other small molecules”
“KEGG GLYCAN”	“Glycans”
“KEGG REACTION”	“Biochemical reactions”
“KEGG RPAIR”	“Reactant pair chemical transformations”
“KEGG RCLASS”	“Reaction class defined by RPAIR”
“KEGG ENZYME”	“Enzyme nomenclature”

KEGG PATHWAY is a database consisting of flat files and has DBGET/LinkDB as its data retrieval system. It contains a collection of manually drawn pathway maps and the reaction network for these categories: metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems and human diseases.

2.2.15.2 BioCyc database

The BioCyc was developed by Stanford Research Institute and contains collection of 9,390 Pathway/Genome Databases (PGDBs) in combination with software/tools for accessing their data [56]. This database contains different organism specific databases having a unique name that starts with initials of an organism followed by ‘Cyc’. For example, EcoCyc database contains information about pathways for *Escherichia coli* K-12; AgroCyc database contains information about the pathways of *Agrobacterium*

tumefaciens C58. This database is divided into three tier based on quality of manually curated data and stored in an object-oriented database management system. Tier 1 databases, these are the most manually accurate curation and contains at least one person-year of literature. Tier 2 and Tier 3 databases contains computationally predicted pathways, if moderate curation (1-4 months) placed in tier 2 else if no manual curation tier 3. The EcoCyc is a freely accessible database for the bacterium *Escherichia coli* K-12 MG1655. This database conducts literature-based curation of the *E. coli* genome, transcriptional regulation, transporters, and metabolic pathways. The MetaCyc was originally released in 1997 and contains information with respect to metabolic pathways and enzymes from 2816 organisms. It contains 2,491 pathways involved in primary, secondary metabolism and also metabolites associated to reactions, genes and enzymes. The MetaCyc information for metabolites includes predicted Gibbs free energies of formation, chemical structures and links to external databases.

2.2.15.3 Reactome database

Reactome, is a manually curated open resource of human pathways and reactions [57]. It was introduced thirteen years ago and since then it has grown exceptionally to include entries for 8,701 human genes containing the annotation of 18,658 specific forms of protein. The pathway structure in this database is of hierarchical order and thereby pathways for translation, protein folding and post-translational modification have been grouped into larger domains of biological function like protein metabolism.

2.2.15.4 WikiPathways

The WikiPathways were built to contribute and maintain pathway information provided by the biology community. These pathways enhance additionally to ongoing primary pathways, such as KEGG, Reactome and Pathway Commons. Initially started with 500 pathways across six species maintained by four people [53]. At present, in 2017 they contain around 2300 pathways across over 25 different species. Its multiple pathways can be edited from its wiki page by initiating an embedded pathway editor. The pathways are available to be downloaded in multiple formats including GPML format.

2.2.15.5 NCBI Biosystems database

The NCBI Biosystems database is a type of secondary databases, which incorporates records from various source databases such as, KEGG, Reactome, BioCyc, Pathway Interaction Database, Gene Ontology and Wikipathways [58]. It categorizes and records

proteins, genes and small molecules, which are involved in biological systems together with their pathway information.

2.2.15.6 Network connectivity diagrams

A network is groups of two or more interconnected nodes. In biology, a network can be any connection of biological information in nodes, which applies to biological systems [59]. The networks provide a significant mean of mathematical representation together with the interconnections for e.g. neural networks. Network connectivity diagrams are referred as the diagrams connecting multiple networks. In case, we consider all the pathways as a biological system then the interconnections between all the pathways can be referred as network connectivity diagrams for pathways. The general rational behind the construction of the network based connectivity diagrams are:

1. To gain insight, as to how the assembled transcripts from our transcriptome map to the known molecules that interact in a biological system [60].
2. They can be used to decipher the interaction between expressed transcribed genes [61].
3. They can be used to depict high connectivity between what are typically considered distinct pathways [62].
4. All possible interactions between the pathways can be depicted using this approach

2.2.16 *Aconitum heterophyllum*

Aconitum heterophyllum is a rare, critically endangered Himalayan species. It is only non-toxic species in the genus *Aconitum* belonging to family Ranunculaceae. It is a biennial herb found in the North-Western Himalayan region and starts flowering from the second year. At present, a total of 250 known species of *Aconitum* exists. The popular species of *Aconitum* include *Aconitum balfourii*, *Aconitum dienorrhizum*, *Aconitum ferox*, *Aconitum heterophyllum*, *Aconitum napellus* and *Aconitum japonicum* etc.

The roots of *A. heterophyllum* contain tubers, contributing to the biosynthesis and accumulation of secondary metabolites in this plant species. This medicinal plant can be found in Himachal Pradesh, Jammu and Kashmir, Sikkim, Uttarakhand and Arunachal Pradesh states in India. In India, the Director General of Foreign Trade has restricted the export of this species, plant portions or any derivatives or extract [63]. The International

Union for Conservation of Nature and Natural Resources has listed this plant as a critically endangered species [6, 7].

Its popular herbal formulation includes balachaturbhadra churna, sudarshana churna, rasnerandadi kwatha and panchatiktaka puggulu ghrta [64]. The important pharmacological properties of *A. heterophyllum* are stated in the Table 2.6.

Table 2.6 Important pharmacological values of *A. heterophyllum*

Medicinal Property	Reference(s)
Anti-diarrheal	[65]
Antibacterial	[66]
Anti-diabetic	[65]
Antioxidant	[67]
Aphrodisiac	[68]
Arthritis	[69]
Hypolipidemic	[70]

2.2.17 *Swertia chirayita*

Swertia chirayita (Gentianaceae) commonly known as ‘Chirata’, found at an altitude of 1200-3000 m and indigenous to temperate Himalayas (Kashmir to Bhutan, and Kashia hills) [71]. Its medicinal properties are well documented in the Indian pharmaceutical codex. This medicinal herb has a bitter taste due to the presence of different bioactive compounds such as amarogentin, swertiamarin, mangiferin, swerchirin, sweroside, amaroswerin, gentianine, oleanolic acid, ursolic acid, swertanone, syringaresinol, bellidifolin, isobellidifolin, 1-hydroxy-3,5,8-trimethoxyxanthone, 1-hydroxy-3,7,8-trimethoxyxanthone, 1,5,8-trihydroxy-3-methoxyxanthone, β -amyrin and chiratol [71]. The increase in demand for *S. chirayita* has led to its overharvesting and had brought this species on a verge of extinction [72].

It is perhaps best known in India as the main ingredient in laghu sudarshana churna, mahasudarshana churna, chinnodbhavadi kvatha churna, ayush-64, himalaya diabecon,

mensturyl syrup and sudarshan churna. Some important pharmacological properties of *S. chirayita* are given in Table 2.7.

Table 2.7 Important medicinal properties of *S. chirayita*

Medicinal Property	Reference(s)
Antimalarial	[73]
Anticancer	[74]
Antidiabetic	[75]
Antifungal	[76]
Anti-inflammatory	[77]
Antioxidant activities	[78]
Antiviral	[79]

2.2.18 Secondary metabolism and associated pathways (MVA/MEP)

The idea of secondary metabolism is dedicated to Kossel [80]. His work was first to define and differentiate these metabolites from the primary ones. Secondary metabolism includes all metabolic pathways and other small molecules, which are not associated with growth and development of an organism or are not required for the survival of the organism. This group consists of both simple molecules (alcohols, organic acids and sugars) and complex compounds such as flavonoids, terpenes, polyketides and non-ribosomal peptide compounds [81] (KEGG).

These plants are key resources of alkaloids and flavonoids, which constitute important medicinal properties. The biosynthesis pathways for these secondary metabolites are very convoluted and complex. In fact, the complete biosynthesis pathway for most of the secondary metabolites is not characterized yet and putatively hypothesized based on the existing information. In plants two major secondary metabolism pathways exist, i.e. mevalonate (isoprenoid pathway or HMG-CoA reductase pathway or MVA) and non-mevalonate pathways (2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate or MEP). The MVA pathway initiates with acetyl-CoA and exits with the

production of IPP and DMAPP. At present six enzymes are known to be responsible for the production of final product DMAPP through MEV pathway. Whereas, MEP pathway begins with pyruvate and glyceraldehyde 3-phosphate and ends with dimethylallyl pyrophosphate.

2.3 Materials and Methods

2.3.1 Medicinal plant species

We have selected two medicinal herbs (*A. heterophyllum* and *S. chirayita*) from North-Western Himalayas having high therapeutic values. The plants were collected from the Himalayan Forest Research Institute, Shimla, Himachal Pradesh and maintained under controlled environmental condition in the greenhouse of Jaypee university of Information Technology. Two tissues were selected on the basis of differential conditions of secondary metabolite biosynthesis and accumulation in *A. heterophyllum* whereas, for *S. chirayita* two tissues were selected based on differential modes of nutrition i.e. photoheterotrophs and photoautotrophs.

1.3.1 Transcriptomes generation

A total of four transcriptomes, two from each plant species (*S. chirayita* and *A. heterophyllum*) were generated and analyzed to decipher the molecular components associated with the secondary metabolite synthesis. For transcriptomes generation, root (AHSR) and shoot (AHSS) tissues of *A. heterophyllum*; and greenhouse grown (SCFG) and tissue cultured grown (SCTC) plants of *S. chirayita* were used in the current research work (Fig. 2.4).

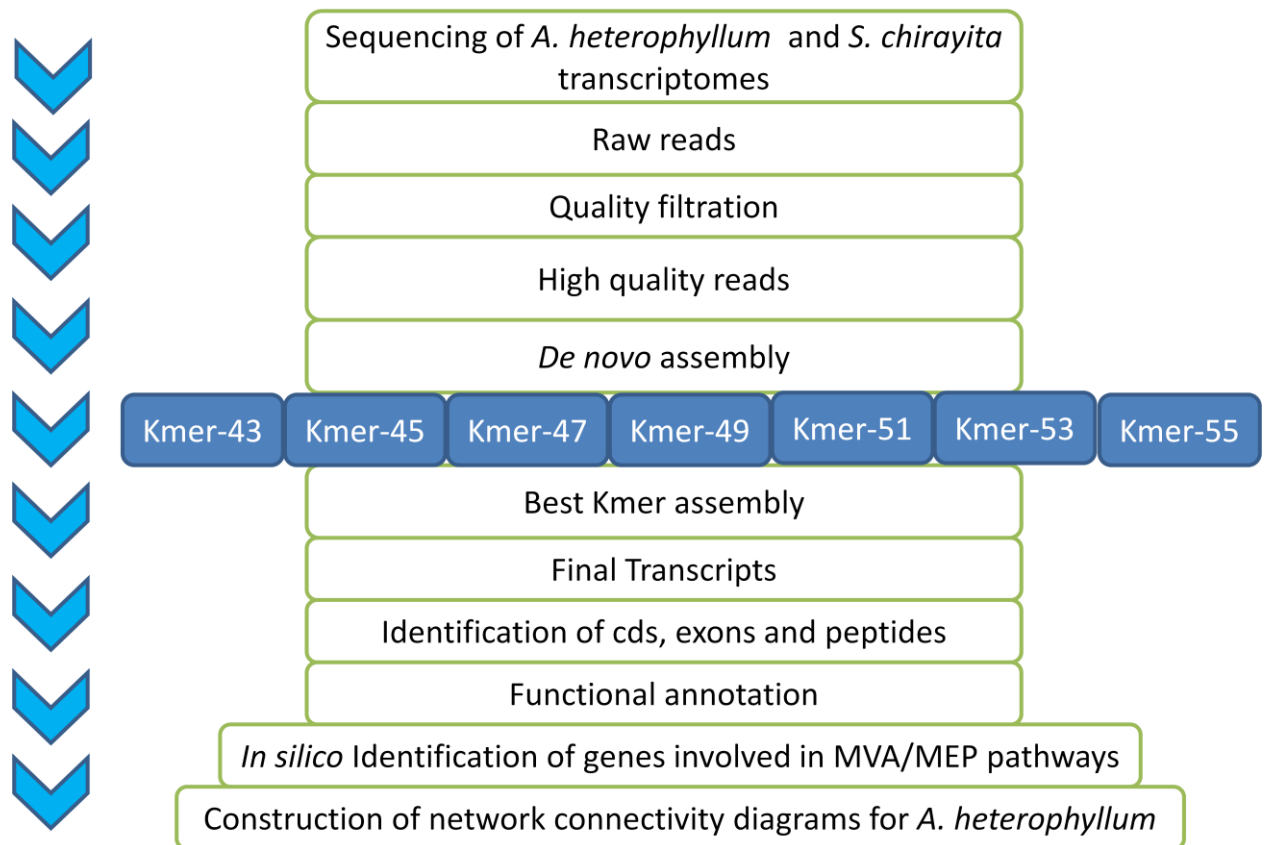


Fig. 2.4 Flow diagram of whole transcriptome sequencing, assembly, annotation and analysis for root (AHSR) and shoot (AHSS) transcriptomes of *A. heterophyllum* and tissue cultured (SCTC) and greenhouse grown (SCFG) plant transcriptomes of *S. chirayita*.

2.3.3 De novo assembly using Illumina HiSeq 2000 platform

Once the sequencing has been done, useful and meaningful information needs to be extracted from raw data. For quality filtration, the adaptors were trimmed and the raw-quality reads was filtered to retain only high quality reads using quality check software, Trimmomatic v0.32. High-quality reads with mean quality value $QV \geq 20$ were retained and trash reads were discarded. Velvet pipeline based on de Bruijn graph algorithm was employed to assemble all the four transcriptomes based on the k-mer value (k-mer range 31 to 63). The best k-mer from all these was chosen for each sample based on transcriptome length covered and N50 value, this best k-mer based assembly was further used for downstream analysis. The GENSCAN gene tool based on *Arabidopsis* model matrix was utilized to predict (CDS), exons and peptides from these transcripts. The dataset regarding both the plant species is available at our in-house developed website URL: <http://14.139.240.55/NGS/download.php>.

2.3.4 Functional annotation, GO mapping and COG analysis

The functional annotation analysis was carried out using homology search for all four datasets using NCBI non-redundant (nr) database (with significant E-value $<1e-5$). To further categorize the annotated sequences the BLAST2GO tool was used to map the assembled transcripts from all four transcriptomes to biological process, molecular function and cellular component ontologies [82]. Further, to classify the transcripts on the basis of orthologous genes in-house scripts were developed to align all the assembled transcripts from all the four transcriptomes to the COG database. The assembled transcripts were mapped against COG database (significant E-value $<1e-5$). From these the top scored transcripts in terms of E-values were further mapped to their respective COG IDs. The following code was used to predict and classify all four transcriptomes according to COG database.

```
open(F1,"check") or die "file not found";
open(F2,"whog.txt") or die "file not found";
while(<F1>)
{
chomp($_);
@a=split(>/>,$_);
push(@b,$a[1]);
}
while(<F2>)
{
    if($_ =~ /COG/)
    {

        @d=split(/\s/, $_);
        $count=1;

    }

    if ($count==1)
    {
        foreach $x (@b)
        {
            if (/ $x /)
            {
                print "$x\t$d[0]\n";
            }
        }
    }
}
}
```

2.3.5 Transcript abundance prediction using fragment mapping approach

Transcript abundance calculation for all four assembled transcriptomes was carried out using RSEM v1.2.5 (RNA-Seq by Expectation Maximization) software package. RSEM estimates transcript abundance by mapping RNA-Seq reads to the assembled transcriptome. It calculates posterior mean estimates, maximum likelihood abundance estimates and 95% credibility intervals for genes as well as isoforms. Since the sequencing was paired-end, FPKM (Fragments per kilobase per million) expression unit was used due to its sensitive nature for measuring expression level of even poorly expressed transcripts. Two commands of RSEM i.e. `rsem-prepare-reference` and `rsem-calculate-expression` were optimized to organize the reference sequences and the second to calculate the expression values using raw reads. The abundance data for all the four transcripts were compiled. The example of optimized commands used to gather information from RSEM are:

- 1) `./rsem-prepare-reference --no-polyA --bowtie-path /home/ngs_server/Project/bowtie-master AHSR_fasta Result`
- 2) `./rsem-calculate-expression --bowtie-path /home/ngs_server/Project/bowtie-0.12.8 --paired-end AHSR_pair_R1.fastq AHSR_pair_R2.fastq -p 3 Result Result_AHSR_qual`

2.3.6 Computational mining of transcriptomes for MVA/MEP pathway genes

All the fifteen genes of mevalonate (MVA) and non-mevalonate (MEP) pathways involved in the biosynthesis and accumulation of secondary metabolites in *A. heterophyllum* and *S. chirayita* were computationally mined using in-house developed perl script. The assembled transcripts from all the four transcriptomes (AHSR, AHSS, SCFG and SCTC) were scanned against NR protein database available at NCBI using BLASTX algorithm with E-value threshold of 10^{-5} . The identified fifteen genes/transcripts were *DXPS*, *DXPR*, *ISPD*, *ISPE*, *MECPS/ISPF*, *HDS/ISPG*, *ISPH*, *IPP*, *GDPS*, *ACTH*, *HMGs*, *HMGR*, *MVK*, *PMK* and *MVDD*. Further, *in silico* transcript abundance was quantified for these genes to assess the contribution and relevance of MVA/MEP in the biosynthesis of secondary metabolites in these plant species.

2.3.7 Domain prediction using Pfam database

Domains comprise of motifs and are the functional unit of the proteins [83]. They contain conserved region of the protein, which folds and can function independently. HMMER3.1b1 tool was used to predict all potential peptide/protein sequences by mapping them against Pfam domain database. The program pfam_scan.pl and pfam library of HMMs was employed for identification of common domains in these peptide/protein sequences (<http://pfam.janelia.org/>).

2.3.8 Pathway mapping using KEGG in *A. heterophyllum* and *S. Chirayita* transcriptomes

All the transcripts of the transcriptomes from both the plant species were mapped to their respective biological pathways using KEGG Automatic Annotation Server - KAAS. KAAS assigns “KEG orthology (KO) identifiers” to transcripts, which was further mapped to EC (enzyme commission) number. It carries out identification of homologs between the reference sequence set and the query sequence (KEGG gene database). The threshold of 60 value (default BLAST bit score) was used. The method used was Bi-directional best-hit method and only genes which has BHR greater than 0.95 were selected. Further, all the transcripts were classified into environmental and information processing (ABC transporters, Phosphatidylinositol signaling system), cellular processes (secretion system proteins, ion channels, GTP binding proteins) genetic information processing (translation factors, transcription factors and DNA replication) using BRITE hierarchies.

2.3.9 Biosystems classification and network connectivity diagrams

Functional classification was performed using NCBI Biosystems database. This database consists of group of biomolecules, which interact in a biological system. It stores and cross-links all popular existing biological systems databases including KEGG, BioCyc, Reactome and others. It maintains NCBI resource, which allows rapid classification of genes, proteins and other small molecules by metabolic pathway, biosystem type and their disease state. For further analysis, all the hits corresponding to transcripts (all the four transcriptomes) from the annotation file were mined and mapped onto their respective pathways using NCBI Biosystems database. All the transcripts from all four transcriptomes with significant matches were classified into five broad classes, i.e. genetic information processing, metabolism, cellular processes, organismal systems and environmental information processing. The largest class was observed to be metabolism

which was followed by genetic information processing (359, 41.74 %) cellular processes (38, 4.41 %), organismal systems (32, 3.72 %) and environmental information processing.

Whereas, Network connectivity diagrams can be used to gain insight as to how the assembled transcripts from our transcriptome map to the known molecules that interact in a biological system. They can be used to decipher the interaction between expressed transcribed genes, to depict connectivity between distinct pathways and can also contribute in detecting all possible interactions between the pathways. In order to decipher the interactions and flux between secondary and primary metabolic pathways network connectivity diagrams were drawn for isoquinoline alkaloids biosynthesis pathway for root transcriptome of *A. heterophyllum* using Perl and using Perl-tk. The connections of the arrows were drawn using Perl package manager GD arrow.

```
use GD;
use GD::Arrow;
open(F4,"pathway_information.txt");
while(<F4>)
    {
        chomp($_);
        @w=split(/\t/, $_);
        $s{$w[0]}=$w[2];
    }
close(F4);

opendir F1, "./information/";
@wa=readdir F1;
closedir(F1);
mkdir "./information1_long";
foreach $x8(@wa)
    {
        if($x8 ne '.' && $x8 ne '..')
            {
                $count=0;
                undef %count1,%count2,%count7;
                open(F2,"./information/".$x8);
                open(F5,">./information1_long/".$x8.".jpeg");
                while(<F2>)
                    {
                        $count7{$count}=0;
                        chomp($_);
                        @w1=split(/\t/, $_);
                        foreach $x(@w1)
                            {
                                @w=split(/[/]/,$x);
```

```

if($count1{$count}{$w[1]} ne 1 )
{
$count1{$count}{$w[1]}=1;
$name{$count}{$count7{$count}}=$w[1];
$count7{$count}=$count7{$count}+1;
$count3{$count}{$w[1]}=0;
$count2{$count}{$w[1]}{$count3{$count}{$w[1]}}=$w[1];
$count5{$count}{$w[1]}{$count3{$count}{$w[1]}}=$w[0];
$count3{$count}{$w[1]}=$count3{$count}{$w[1]}+1;
}
else
{
$count2{$count}{$w[1]}{$count3{$count}{$w[1]}}=$w[1];
$count5{$count}{$w[1]}{$count3{$count}{$w[1]}}=$w[0];
$count3{$count}{$w[1]}=$count3{$count}{$w[1]}+1;
}
}
$count=$count+1;
}

open(F2,">positions.txt");
for($i=0;$i<$count;$i++)
{

for($i1=0;$i1<$count7{$i};$i1++)
{

$position1{$i}{$i1}{x}=50+($i*540);

$position1{$i}{$i1}{y}=50+($i1*80);
print F2 $i." ".$i1." ".$position1{$i}{$i1}{x}." ".$position1{$i}{$i1}{y}."\n";

$position2{$i}{$i1}{x}=50+($i*540);

```

```

$position2{$i}{$i1}{y}=$position1{$i}{$i1}{y}+25;

$position3{$i}{$i1}{x}=$position1{$i}{$i1}{x}+400;

$position3{$i}{$i1}{y}=$position1{$i}{$i1}{y}+25;

print F2 $i." ".$i1." ".$position1{$i}{$i1}{x}." ".$position1{$i}{$i1}{y}."
".$position2{$i}{$i1}{x}." ".$position2{$i}{$i1}{y}." ".$position3{$i}{$i1}{x}."
".$position3{$i}{$i1}{y}."\n";
}
}

close(F2);

```

```

# create a new image
$im = new GD::Image(6500,3000);

# allocate some colors
$white = $im->colorAllocate(255,255,255);
$black = $im->colorAllocate(0,0,0);
$red = $im->colorAllocate(255,0,0);
$blue = $im->colorAllocate(0,0,255);

# make the background transparent and interlaced
$im->transparent($white);
$im->interlaced('true');

for($i1=0;$i1<$count;$i1++)
{
    for($i2=0;$i2<$count7{$i1};$i2++)
    {
        $x1=$position1{$i1}{$i2}{x};
        $y1=$position1{$i1}{$i2}{y};
        $x2=$x1+400;
        $y2=$y1+50;

        if($x1 eq 50 && $y1 eq 50 )
        {

$im>rectangle($x1,$y1,$x2,$y2,$red);
        }
        else
        {

$im>rectangle($x1,$y1,$x2,$y2,$black);

```

```

    }

    $im>string(gdMediumBoldFont,($x1+5),($y1+5),$s{$name{$i1}{$i2}},$red);
    }

    }

my $width = 1;
my @arrow;
$k=0;
for($i=1;$i<$count;$i++)
    {
        for($i1=0;$i1<$count7{$i};$i1++)
            {
                for($i2=0;$i2<$count3{$i}{$name{$i1}{$i1}};$i2++)
                    {
                        for($i3=0;$i3<$count7{($i-1)};$i3++)
                            {
                                for($i4=0;$i4<$count3{($i-
1)}{$name{($i-1)}{$i3}};$i4++)
                                    {

                                        if($count5{$i}{$name{$i1}{$i1}}{$i2} eq $count2{$i-1}{$name{$i-
1}{$i3}}{$i4} )
                                            {
#print $count5{$i}{$name{$i1}{$i1}}{$i2}." ".$count2{$i-1}{$name{$i-
1}{$i3}}{$i4}."\n";

                                }

                                $x2=$position3{($i-1)}{$i3}{x};

                                $y2=$position3{($i-1)}{$i3}{y};

                                $x1=$position2{$i}{$i1}{x};

                                $y1=$position2{$i}{$i1}{y};

                                $arrow[$k] = GD::Arrow::Full->new(

```

```
-X1 => $x1,
-Y1 => $y1,
-X2 => $x2,
-Y2 => $y2,
-WIDTH => $width,
);
```

```
$im->filledPolygon($arrow[$k],$blue);
```

```
$k=$k+1;
```

```
}
```

```
}
```

```
}
```

```
}
```

```
}
```

```
}
```

```
binmode F5;
```

```
print F5 $im->jpeg;
```

```
close(F5);
```

```
undef
```

```
%count7,%count1,%count2,%count3,%count5,%name,%position,%position1,%position2,%position3,$im;
```

```
}
```

```
}
```

2.3.10 Data availability

All the data related to the four transcriptomes generated in this study can be downloaded from URL: <http://14.139.240.55/NGS/download.php>. The website has been developed using PHP and HTML in combination with customized Java scripts. The website is hosted on DELL PowerEdge™ T410 server, which has 16 cores 2.67 GHz Intel R Xenon processors.

2.4 Results and Discussion

The transcriptome data provides a valuable resource for rapid elucidation of pathways, identification and characterization of biosynthesis pathway genes, transcription factors pathway mapping and construction of graphical connectivity diagrams, etc. [84, 85]. The transcriptomes for both these plant species have been assembled and characterized using Illumina platform, which generated millions of paired-end sequencing reads. The current study deciphered molecular components implicated in differential tissues (root versus shoot) and differential modes of nutrition (photoautotrophic versus photoheterotrophic) vis-à-vis secondary metabolites production in these plant species.

2.4.1 *De novo* sequencing using Illumina platform

In order to obtain a global overview of all the four transcriptomes, paired end sequencing data was generated for each sample using the Illumina HiSeq 2000 platform. Independent reads of 2×90 -bp (PE) were generated for each sample using standard Illumina protocols. In total 49,131,411 and 30,641,740 raw sequencing reads were produced for root (AHSR) and shoot (AHSS) *A. heterophyllum* samples, respectively. The quality filtering was performed on the raw reads to produce 46,612,687 and 28,777,415 good quality reads for AHSR and AHSS, respectively. The *de novo* assembly was performed for both the samples using Velvet pipeline with optimized parameters. To select the best k-mer size based on N50 and transcriptome length covered, the assembler was run at different k-mer sizes ranging between 31 to 63 mers. In case of AHSR the best K-mer size was observed to be 51, whereas for AHSS it was found to be 43. The assembler statistics generated a total of 75,548 transcripts with the length ranging from 200 to 12,376 bp, GC content 42% and N50 of 1059 bp for AHSR sample. Similarly, 39,100 assembled transcript sequences were generated for AHSS with the length ranging from 200 to 19,757 bp, GC content 42% and N50 of 1059 bp (Table 2.8).

In case of *S. chirayita* 43,306,144 and 23,075,416 raw sequencing reads were generated for for greenhouse and tissue cultured *S. chirayita* samples. After quality checking 41,031,326 and 21,859,688 high quality reads were observed for greenhouse and tissue cultured *S. chirayita* samples. The *denovo* assembly was performed using Velvet assembler with k-mer sizes ranging between 31 to 63 mers. The k-mer of 51 for SCFG and 47 for SCTC, were found to be best among all sets of k-mer. A total of 57,460 assembled transcript sequences were obtained for the SCFG sample with the length ranging from 200 to 10,838 bp, GC content 41 % and N50 of 1700 bp. Whereas, tissue

cultured sample yielded 43,702 assembled transcripts with the length varying from 200 to 7,803 bp, GC content 41% and N50 of 1629 bp (Table 2.9).

The difference in the number of transcripts generated for AHSR and AHSS samples of *A. heterophyllum* may be due to the differential mode of biosynthesis and accumulation in root as compared to shoot sample. Whereas, in case of *S. chirayita* this difference may be due to their differential growth conditions, i.e. photoautotrophic versus photoheterotrophic modes of nutrition.

Table 2.8 Assembly statistics for transcriptomes from roots (AHSR) and shoots (AHSS) samples of *A. heterophyllum*

Description	Root Transcriptome	Shoot Transcriptome
Best k-mer	k-mer 51	k-mer 43
Number of assembled transcripts	75,548	39,100
Transcript N50	1,059	1,239
Maximum transcript length (bp)	12,376	19,757
Minimum transcript length (bp)	200	200
Guanine-cytosine content (%)	42	42

Table 2.9 Assembly statistics for transcriptomes from greenhouse (SCFG) and tissue cultured (SCTC) samples of *S. chirayita*

Description	SCFG Transcriptome	SCTC Transcriptome
Best k-mer	k-mer 51	k-mer 47
Number of assembled transcripts	57,460	43,702
Transcript N50	1,700	1,629
Maximum transcript length (bp)	10,838	7,803
Minimum transcript length (bp)	200	200
Guanine-cytosine content (%)	41	41

2.4.2 Functional annotation and classification of transcripts

The functional annotation for all the four transcriptomes was carried out using sequence similarity search with non-redundant (nr) database of NCBI with E-value threshold of $1e-5$. From a total of 75,548 high quality assembled transcripts 46,850 sequences were found to have significant hits with nr protein database (NCBI), whereas for 28,698 sequences no hit was observed in AHSR sample. Similarly, in AHSS sample, from 39,100 assembled sequences significant BLAST hit with the nr database was observed for 36,326 sequences while no BLAST hit was found for 2,774 transcripts. Performing BLAST analysis against the UniProt/Swissprot database, significant hits were observed for 36,217 transcripts while no hits were observed for 39,331 transcripts in AHSR; whereas significant hits were observed for 21,927 transcript sequences and no hits were observed for 17,173 transcript sequences in AHSS sample (E-value cut off $1e-5$). While annotation against the UniProtKB/Swissprot database yielded significant annotation of 36,217 transcripts while no hits were found for 39,331 transcripts in AHSR sample whereas for AHSS significant hits were observed for 21,927 sequences while no hit was observed for 17,173 (at E-value cut off $1e-5$). Maximum percentage of transcripts showed significant similarity mainly with *Vitis vinifera* species for both the samples (Fig. 2.5 and 2.6). For both the samples, the CDS, exons and peptides were identified using GenScan gene prediction tool based on *Arabidopsis* model matrix parameter (Table 2.10).

Table 2.10 Prediction statistics of CDS, exons and peptides from roots (AHSR) and shoots (AHSS) samples of *A. heterophyllum*

Description	Root transcriptome	Shoot transcriptome
CDS	34,424	23,149
Exons	41,700	27,906
Peptides	34,424	23,149

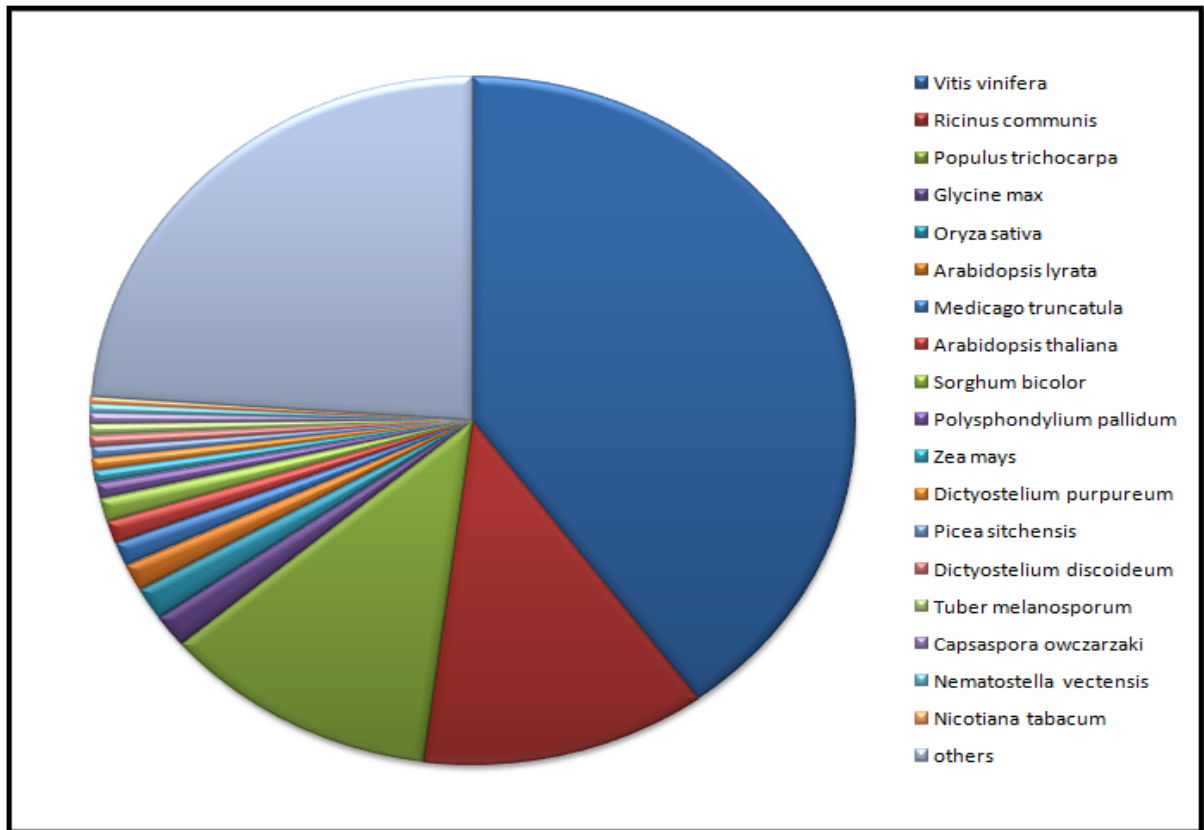


Fig. 2.5 Species distribution of the top Blastx hits against *A. heterophyllum* root transcriptome

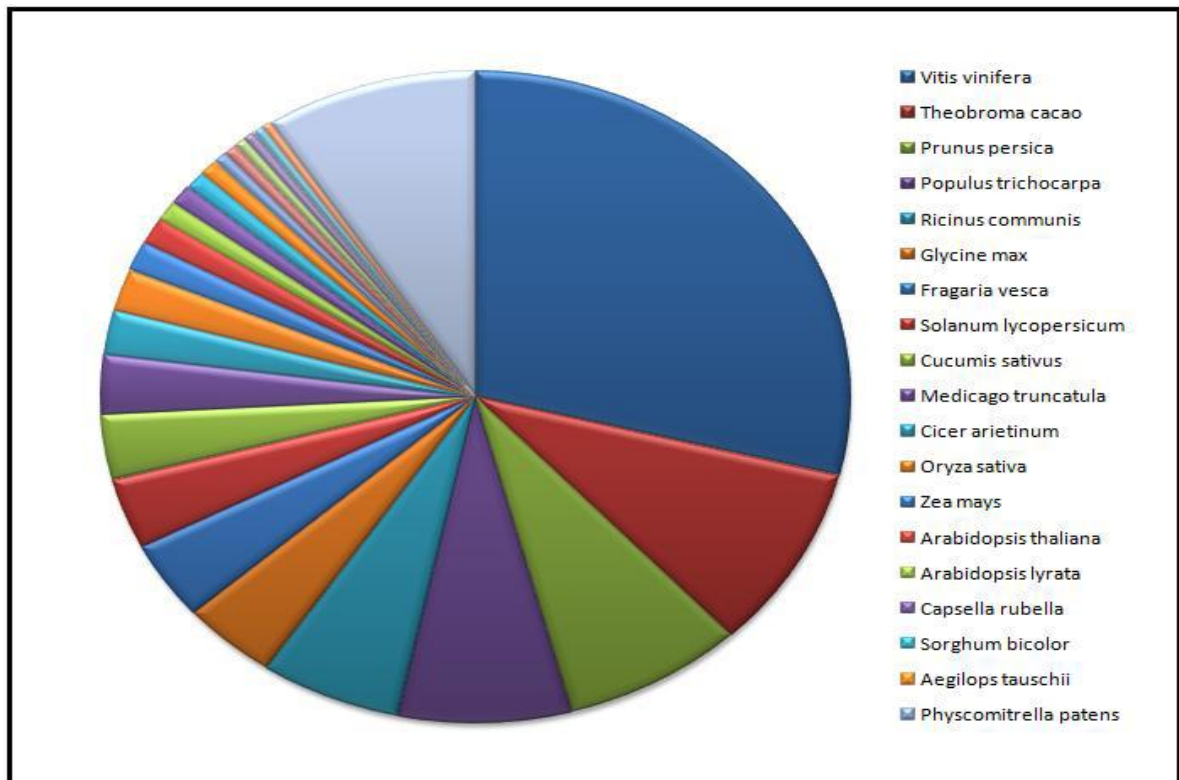


Fig. 2.6 Species distribution of the top Blastx hits against *A. heterophyllum* shoot transcriptome

For *Swertia* transcriptomes, Blastx resulted in the annotation of 50,795 transcripts with significant BLAST hits while no hits were observed for 6,665 transcripts out of total 57,460 assembled high quality transcripts for SCFG. For SCTC sequences with significant BLAST hit was observed to be 39,150 sequences, while for 4,552 transcripts no hits were found from a total of 43,702 assembled high quality sequences. For, species distribution the top Blastx hits were observed against *Solanum lycopersicum* for both the samples followed by *Vitis vinifera* and so forth (Figs. 2.7, 2.8). GenScan prediction tool, based on the *Arabidopsis* model matrix parameter, yielded 35,493 CDS, 44,277 exons and 34,493 peptides for SCFG and 26,349 CDS, 31,757 exons and 31,757 peptides for SCTC (Table 2.11). While the annotation against the UniProtKB/Swiss-Prot database yielded significant annotation of 30,903 from 57,460 assembled sequences in SCFG and 23,563 from 43,702 assembled sequences in SCTC transcriptomes, respectively, representing best possible hits.

Maximum number of the transcripts had a significant hit against unique known proteins from public database; this inferred that a sizeable fraction of unique genes from *A. heterophyllum* and *S. chirayita* transcriptomes were yielded through Illumina sequencing.

Table 2.11 Prediction statistics of CDS, exons and peptides from greenhouse (SCFG) and tissue cultured (SCTC) transcriptomes of *S. Chirayita*

Description	SCFG transcriptome	SCTC transcriptome
CDS	35, 493	26, 349
Exons	44, 277	31, 757
Peptides	35, 493	31, 757

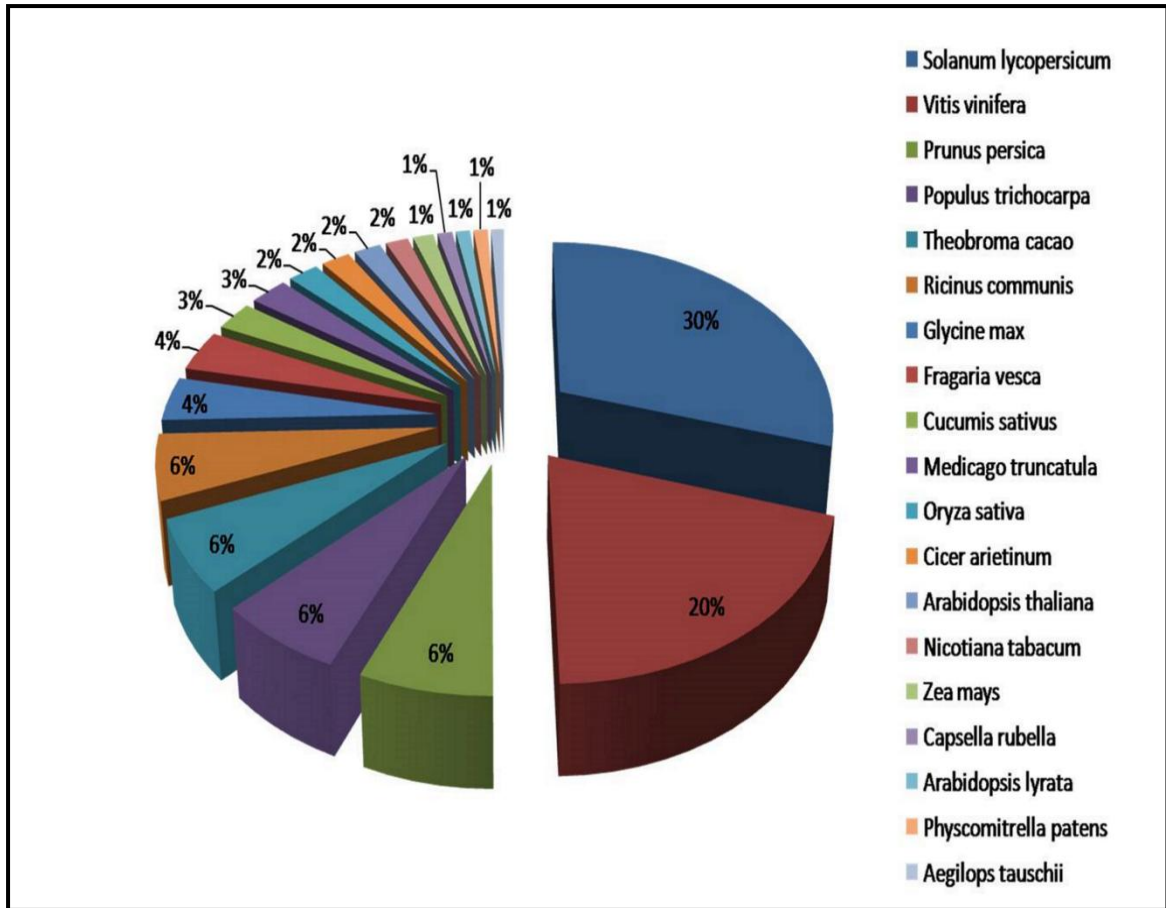


Fig. 2.7 Proportion of SCFG transcripts matching to different plant species

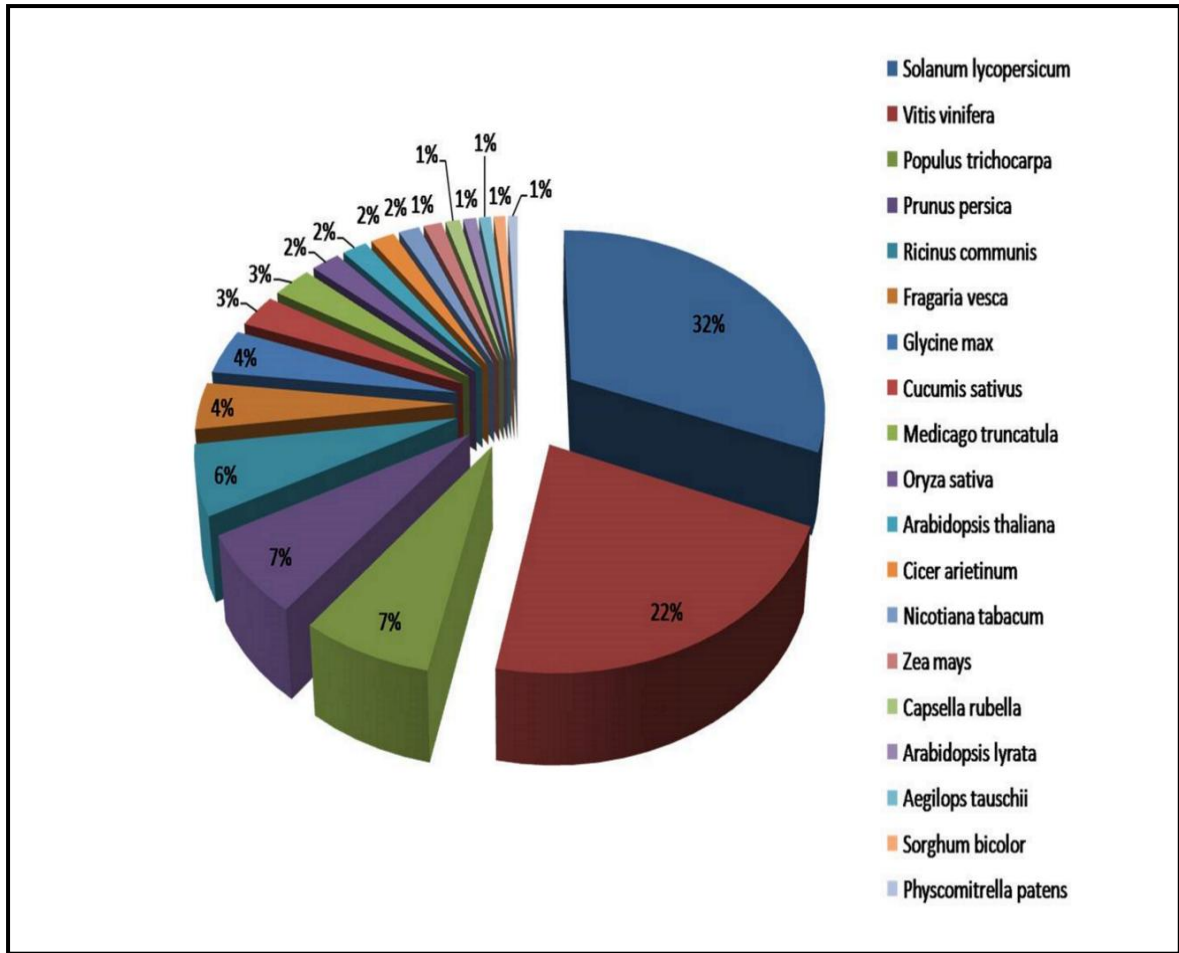


Fig. 2.8 Proportion of SCTC transcripts matching to different plant species

2.4.3 GO and COG functional classification

GO classification is based on orthology and direct experimental evidence with more detailed functional annotation and analysis for gene products (42,988 terms). Its terms are derived from ontologies, which can be used to describe the function of genes and their products. Go primarily classifies the data into three broad categories/ontologies i.e. biological process, molecular function and cellular component.

The GO was carried out using the BLAST2GO software. A total of 27,596 and 12,340 functional terms were assigned to gene ontology classes for AHSR and AHSS, respectively. The majority of the transcripts were aligned to molecular function class (11,996, 43.47 % for AHSR; 5665, 45.90 % for AHSS) as indicated in Fig. 2.9 and 2.10.

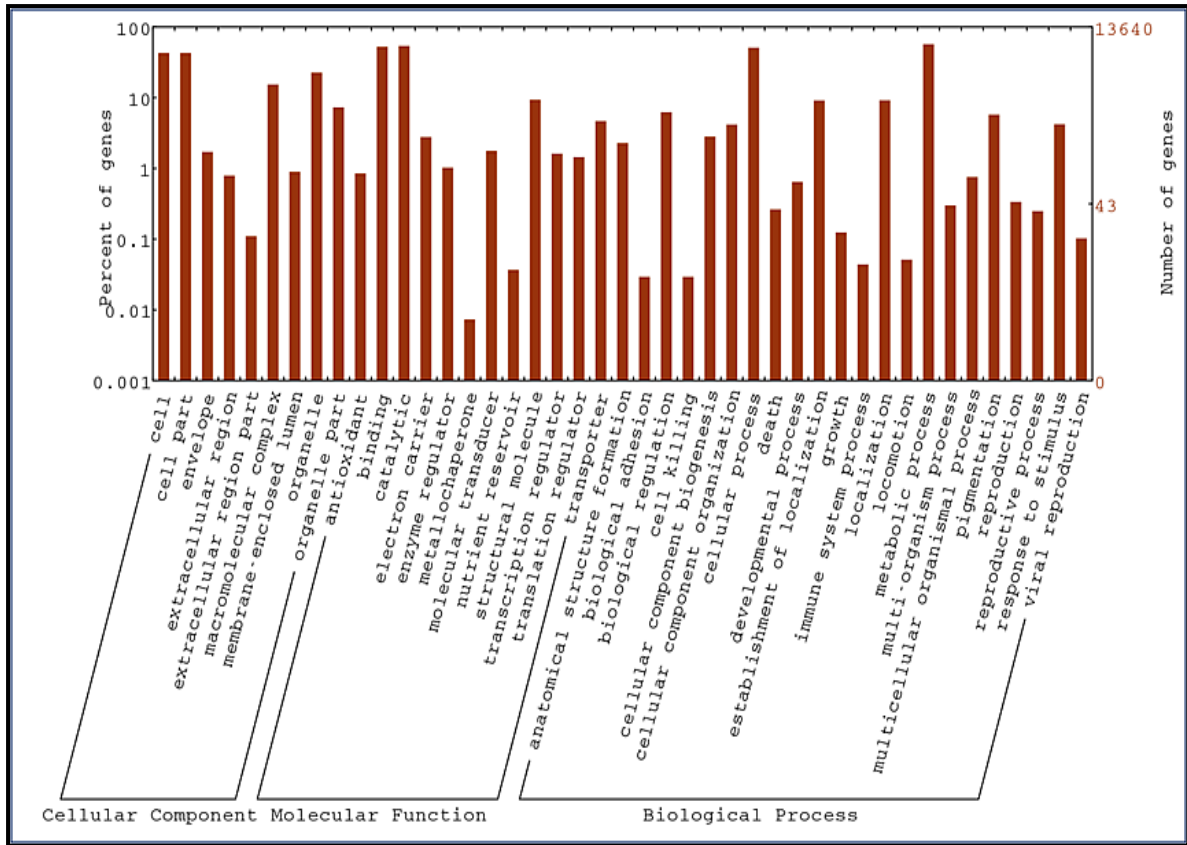


Fig. 2.9 Distribution of GO annotated transcripts for *A. heterophyllum* root transcriptomes

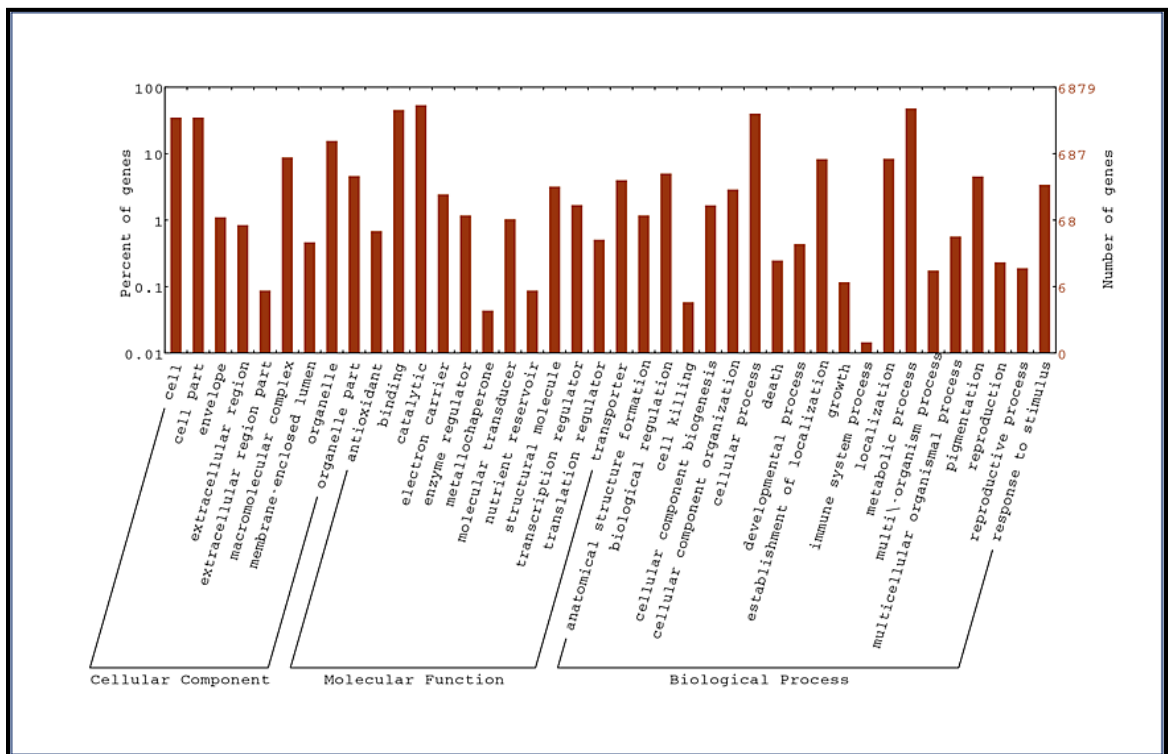


Fig. 2.10 Distribution of GO annotated transcripts for *A. heterophyllum* shoot transcriptomes

Similarly, for *Swertia* transcriptomes 18,090 and 2,102 functional terms were yielded for SCFG and SCTC, respectively. These transcripts were further classified into three major categories, from which the molecular function was observed to be the major class (8,278 and 45.76% for SCFG; 1,106 and 52.61% for SCTC) followed by the biological process (6,222 and 34.39% for SCFG; 560 and 26.64% for SCTC) and cellular component (3,590 and 19.84% for SCFG; 436 and 20.74% for SCTC) (Fig. 2.11 and 2.12). Among three GO categories, cell, catalytic and metabolic process were the most abundant classes in cellular component, molecular function and biological processes, respectively. These observations were in agreement with previously assigned GO terms in *Arabidopsis thaliana*, *Picrorhiza kurroa* and *Medicago truncatula* [85, 86].

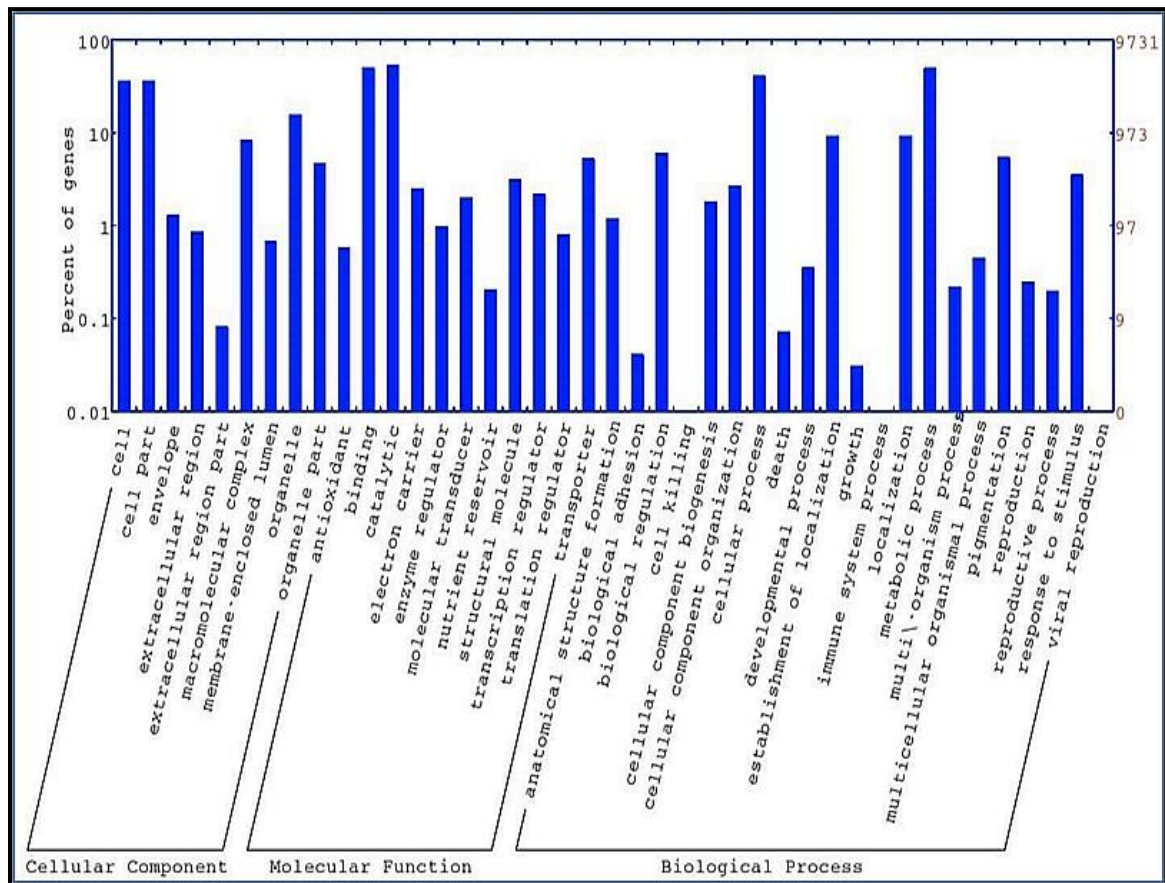


Fig. 2.11 Distribution of GO annotated transcripts in SCFG transcriptome of *S. chirayita*

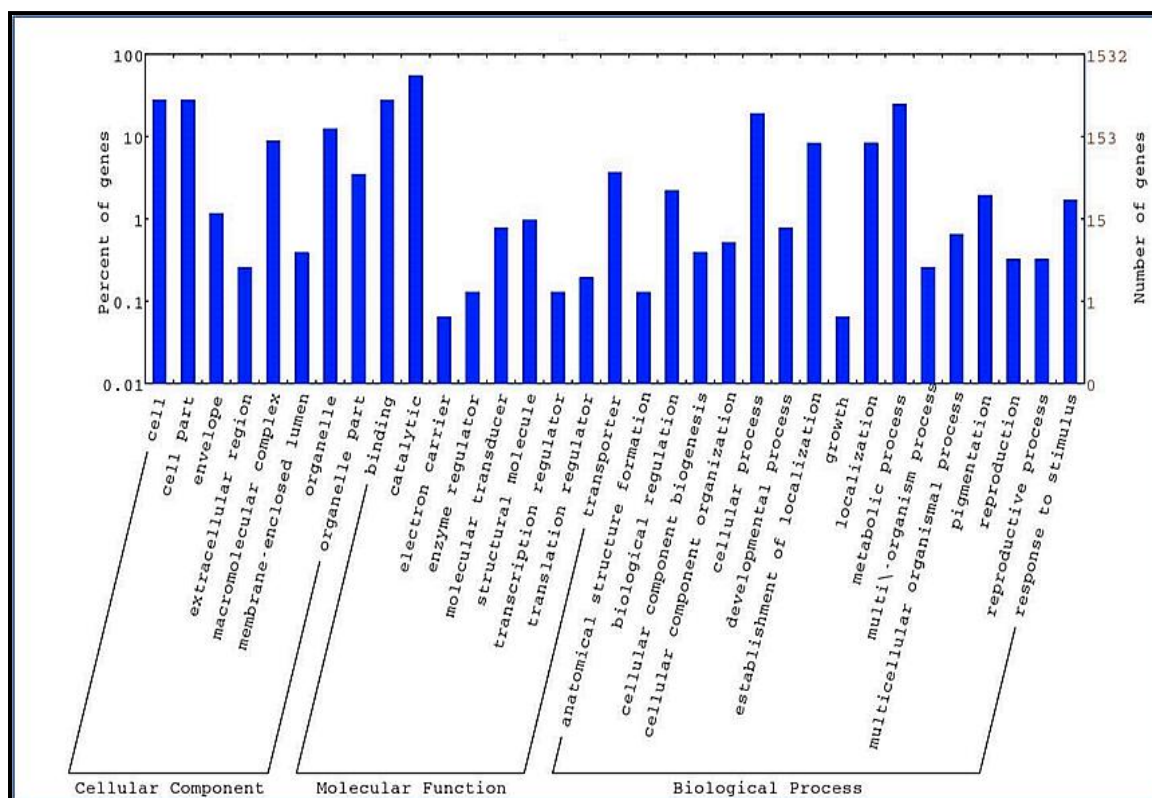


Fig. 2.12 Distribution of GO annotated transcripts in SCTC transcriptome of *S. chirayita*

However, the GO functional classification for all the four transcriptomes yielded maximum number of transcripts matching to molecular function category, which consists of these DNA binding, catalytic and transferase activity, etc. indicating control of gene regulation, enzymatically active and signal transduction processes. This was followed by biological process category, which pointed out that the plant is carrying out an extensive metabolic activity and undergoing rapid growth/development.

The Clusters of Orthologous Groups of proteins (COG) database is an attempt to phylogenetically classify the proteins encoded in a genome. To classify for plausible functions, all the transcripts of the *A. heterophyllum* transcriptomes were aligned against the COG database. The transcripts showing significant similarity (at an E-value < 1e-5) with those in the database were classified into the respective functional classes. Overall 16,604 and 9,398 transcripts of AHSR and AHSS, respectively, were assigned to COG classifications. From the 24 COG IDs, the largest group in both the tissues was observed to be the general function prediction (2,870, 17.28 % in AHSR and 1,786, 19.00 % in AHSS) (Fig. 2.13). The secondary metabolism class was represented by 563 (3.39 %) and

385 (4.09 %) transcripts from AHSR and AHSS samples of *A. heterophyllum*, respectively.

In the similar manner, we categorized *Swertia* transcripts into 24 functional classes, where the most frequent functional category was observed to be “general function prediction (symbol R)” for both tissue samples (2,405 in SCFG and 1,820 in SCTC) followed by “posttranslational modification, protein turnover, chaperones (symbol O)” (1,624 in SCFG and 1,164 in SCTC), “translation, ribosomal structure and biogenesis (symbol J)” (1,184 in SCFG and 907 in SCTC), “carbohydrate transport and metabolism (symbol G)” (783 in SCFG and 622 in SCTC), etc., (Fig. 2.14), etc., which was similar to the study done on *Taxodium* [45]. Furthermore, 528 and 313 transcripts of SCFG and SCTC, respectively, were classified into “secondary metabolites biosynthesis, transport and catabolism (category symbol Q)” in *S. chirayita*.

For all the four transcriptomes “general function prediction” was observed to be the major class which means it contains a set of poorly characterized proteins and the matching transcripts might be specific to these species only or these transcripts have not yet been classified in any other species moreover, to ascertain their exact role there is a need for further validation of these transcripts. Since the COG was updated with eukaryotes datasets recently in 2003, it might be possible that a very less number of genes might have been added to these groups for eukaryotes, which resulted in very less number of hits.

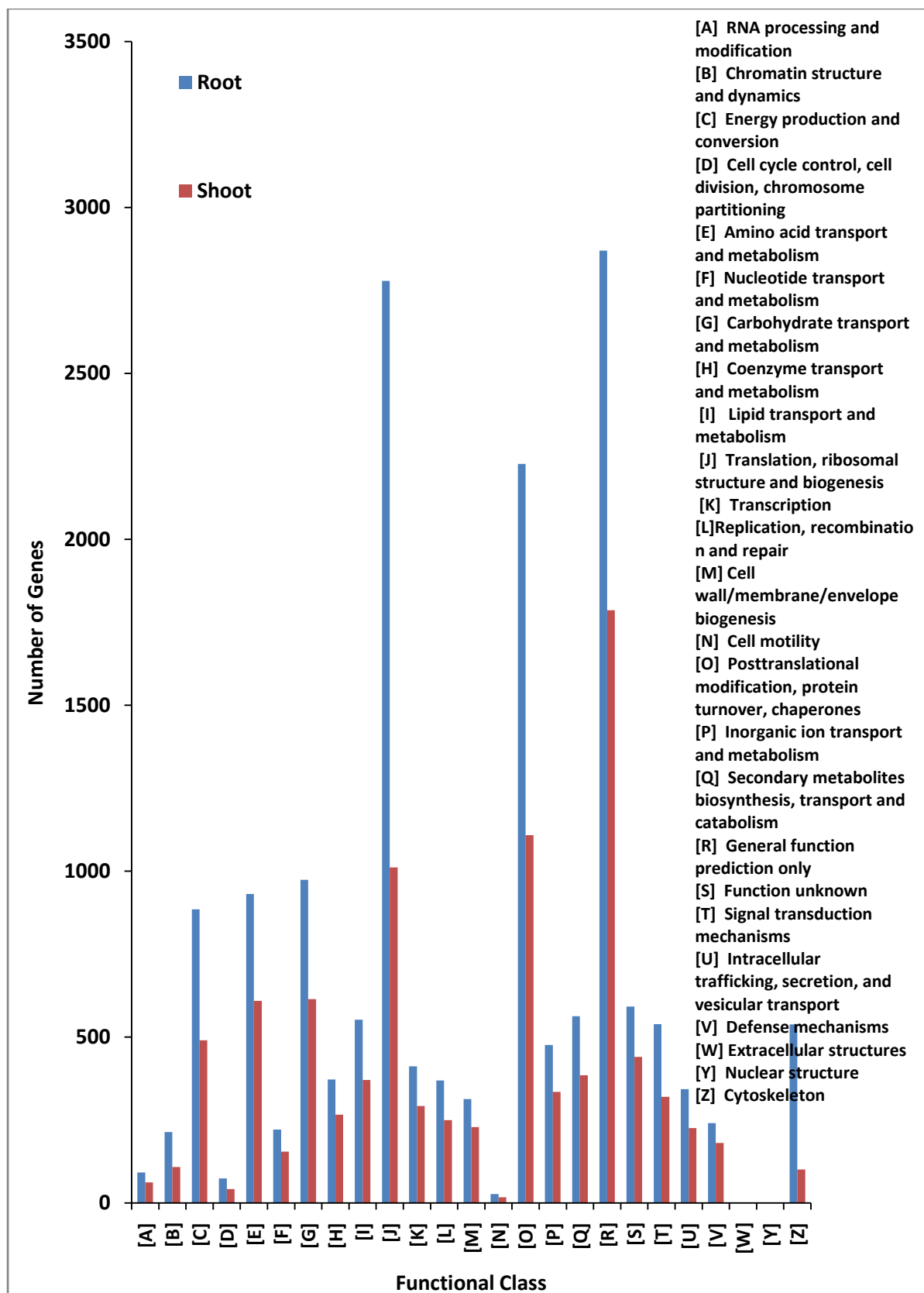


Fig. 2.13 Distribution of COG classified transcripts of root and shoot transcriptomes of *A. heterophyllum*

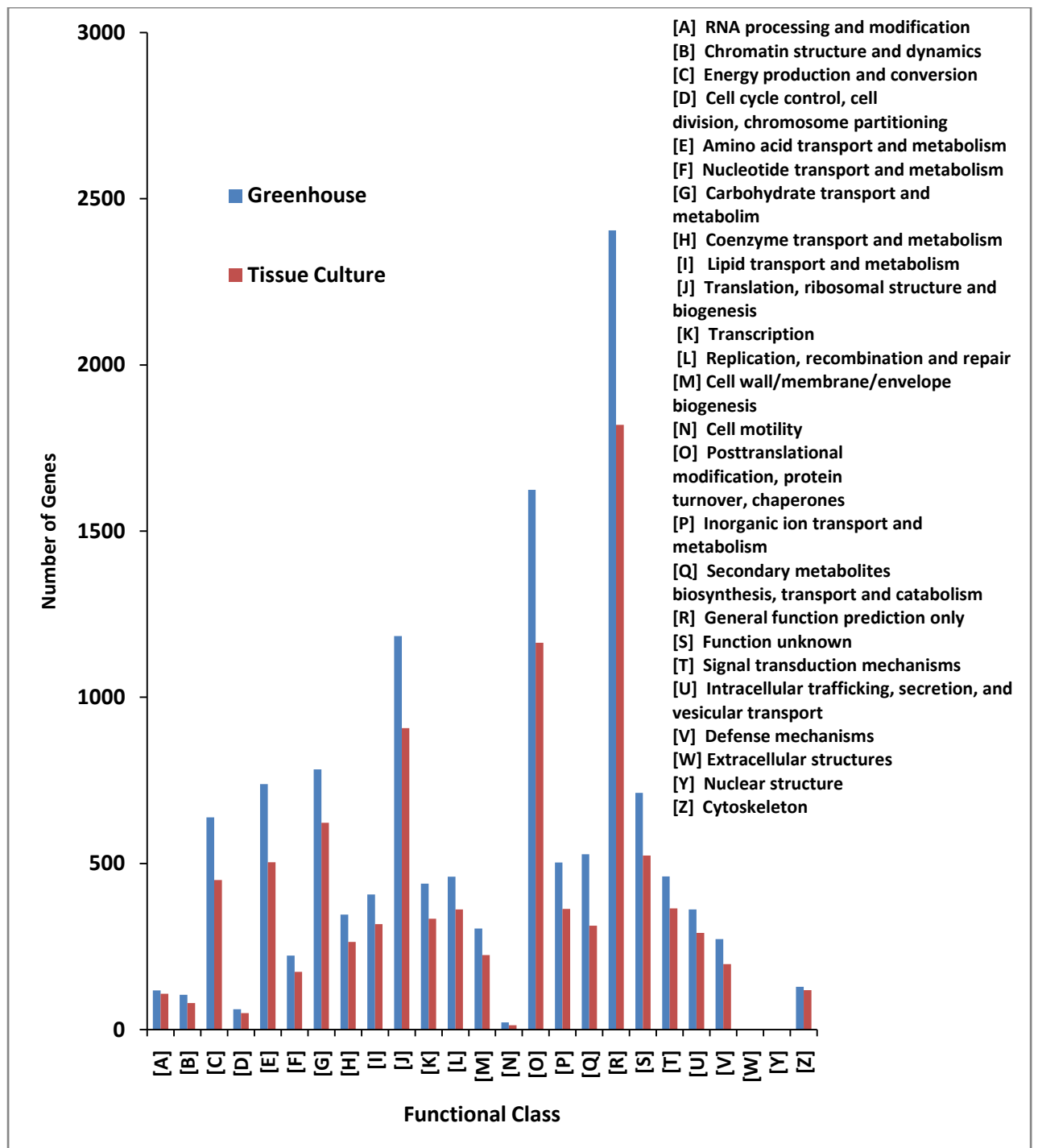


Fig. 2.14 Distribution of COG classified transcripts of greenhouse and tissue cultured transcriptomes of *S. chirayita*

2.4.4 Identification of domains

The domains from all the transcriptomes were mined in both the plant species using Pfam domain database. A total of 29,371 and 21,558 domains were identified against *A. heterophyllum* peptide/protein sequences. Whereas, for *S. chirayita* 39,374 and 28,261 domains were identified for SCFG and SCTC transcriptomes, which could be implicated

in differential modes of nutrition (photoautotrophic versus photoheterotrophic) for production of secondary metabolites in this plant species. From these, domains of different molecular components contributing to secondary metabolism such as 234, 165, 299 and 207 domains of ABC-type transporters (including ABC_membrane_2 domains, PDR_assoc and Cytochrom_C_asm family) were identified from AHSR, AHSS, SCFG and SCTC transcriptomes, respectively.

2.4.5 Mining transcriptomes for genes involved in MVA/MEP biosynthesis pathways

The potential genes of mevalonate (MVA) and non-mevalonate (MEP) pathways implicated in the biosynthesis and accumulation of aconites in *A. heterophyllum* and the secondary metabolites (for both the photoautotrophic and photoheterotrophic modes of nutrition) production in *S. chirayita* [87] were mined on the basis of similarity search using in-house Perl script. A total of 15 enzymes were identified from in-house generated four transcriptomes (AHSR, AHSS, SCFG and SCTC), namely *DXPS*, *DXPR*, *ISPD*, *ISPE*, *MECPS/ISPF*, *HDS/ISPG*, *ISPH*, *ACTH*, *HMGs*, *HMGR*, *MVK*, *PMK*, *MVDD*, *IPP* and *GDPS*, involved in MVA and MEP pathways, from nr database using BLASTX. The analysis of gene expression for all the four transcriptomes was estimated using FPKM approach (Tables 2.12, Fig. 2.15).

Table 2.12 *In silico* transcript quantification for MVA and MEP pathway genes predicted in *A. heterophyllum* root and shoot transcriptomes

Sr. No	Gene Name	EC. No.	AHSR Transcript ID(s)	AHSS Transcript ID(s)	FPKM Root	FPKM Shoot
1.	<i>DXPS</i>	2.2.1.7	Transcript_5409	Transcript_1017	57.68	176.67
2.	<i>DXPR</i>	1.1.1.267	Transcript_1136	Transcript_7612	48.54	102.02
3.	<i>ISPD</i>	2.7.7.60	Transcript_10174	Transcript_27344	23.7	96.94
4.	<i>ISPE</i>	2.7.1.148	Transcript_64620	Transcript_11462	19.13	38.73
5.	<i>MECPS</i>	4.6.1.12	Transcript_1824	Transcript_3605	115.13	403.23
6.	<i>HDS</i>	1.17.7.1	Transcript_8398	Transcript_7021	105.58	90.12
7.	<i>ISPH</i>	1.17.1.2	Transcript_30192	Transcript_4360	20.09	296.29
8.	<i>ACTH</i>	2.3.1.9	Transcript_51038	Transcript_11170	21.43	22.22

9.	<i>HMGS</i>	2.3.3.10	Transcript_5690	Transcript_5417	16.25	84.79
10.	<i>HMGR</i>	1.1.1.34	Transcript_3795	Transcript_2381	143.06	48.26
11.	<i>MVK</i>	2.7.1.36	Transcript_19736	Transcript_14392	14.12	1.92
12.	<i>PMK</i>	2.7.4.2	Transcript_6636	Transcript_4131	4.31	26.54
13.	<i>MVDD</i>	4.1.1.33	Transcript_63336	Transcript_33574	14.31	12.28
14.	<i>IPP</i>	5.3.3.2	Transcript_10661	Transcript_1345	112.15	154.8
15.	<i>GDPS</i>	2.5.1.84	Transcript_52592	Transcript_5178	8.47	66.33

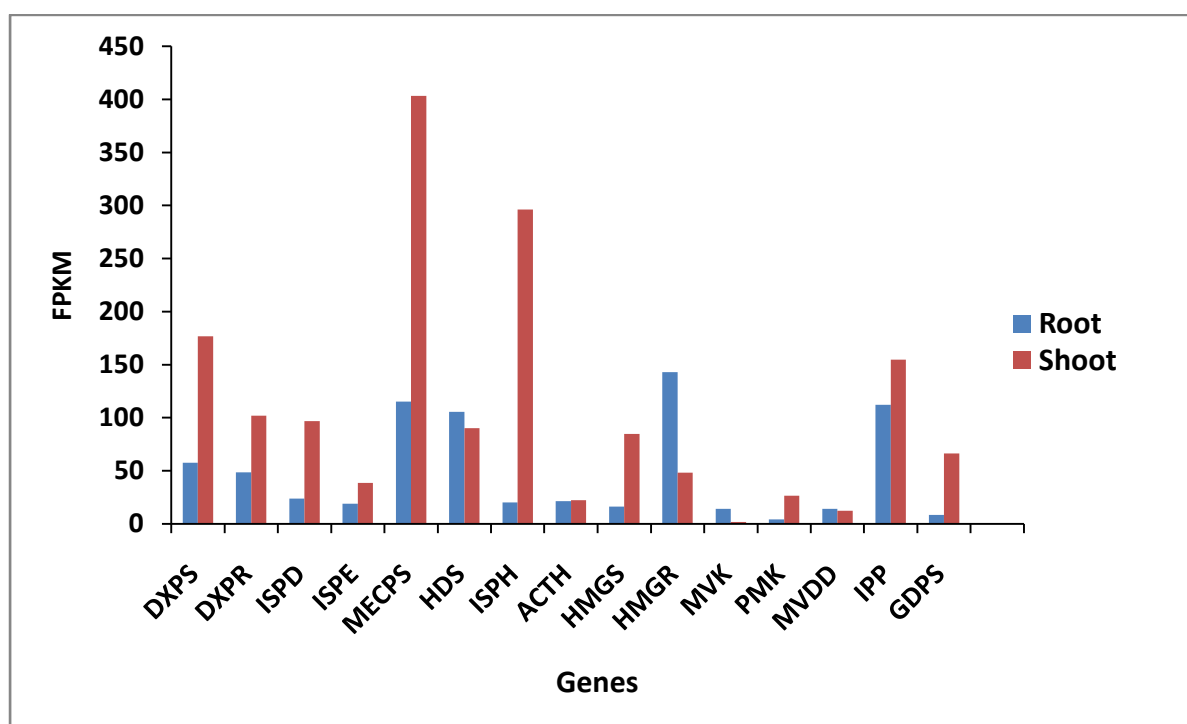


Fig. 2.15 Graphical representation of *in silico* transcript quantification for secondary metabolism (MVA and MEP) pathway genes in root versus shoot transcriptomes

Similarly, in case of *S. chirayita*, SCFG and SCTC transcriptomes were mined for genes involved in the secondary metabolites biosynthesis pathways. The transcript abundance analysis for MVA/MEP revealed that most of the genes (9 genes) showed higher *in silico* expression in SCFG compared to SCTC transcriptomes (Table 2.13, Fig. 2.16).

Table 2.13 *In silico* transcript quantification of secondary metabolism (MVA and MEP) pathway genes in greenhouse versus tissue cultured transcriptomes

Sr. No	Gene Name	E.C. No.	SCFG Transcripts ID(s)	SCTC Transcript ID(s)	FPKM SCFG	FPKM SCTC
1.	<i>AACT</i>	2.3.1.9	transcript_54602	transcript_32680	76.94	82.11
2.	<i>HMGS</i>	2.3.3.10	transcript_520	transcript_4233	63.52	54.79
3.	<i>HMGR</i>	1.1.1.88	transcript_43136	transcript_15869	3.99	11.27
4.	<i>MVK</i>	2.7.1.36	transcript_30528	transcript_41397	18.71	10.96
5.	<i>PMK</i>	2.7.4.2	transcript_30845	transcript_20230	18.22	2.46
6.	<i>MVDD</i>	4.1.1.33	transcript_6471	transcript_39675	37.89	19.7
7.	<i>DXS</i>	2.2.1.7	transcript_5820	transcript_32850	22.66	49.05
8.	<i>DXR</i>	1.1.1.267	transcript_1841	transcript_3020	28.3	160.35
9.	<i>ISPD</i>	2.7.7.60	transcript_3471	transcript_14543	51.25	12.76
10.	<i>ISPE</i>	2.7.1.148	transcript_898	transcript_5328	109.59	36.53
11.	<i>ISPF</i>	4.6.1.12	transcript_1885	transcript_32702	450.55	155.22
12.	<i>ISPG</i>	1.17.7.1	transcript_907	transcript_24152	277.77	300.43
13.	<i>ISPH</i>	1.17.1.2	transcript_827	transcript_32372	11.68	401.33
14.	<i>IPPI</i>	5.3.3.2	transcript_1912	transcript_14601	196.98	133.16
15.	<i>GDPS</i>	2.5.1.1	transcript_9738	transcript_18602	14.52	14.06

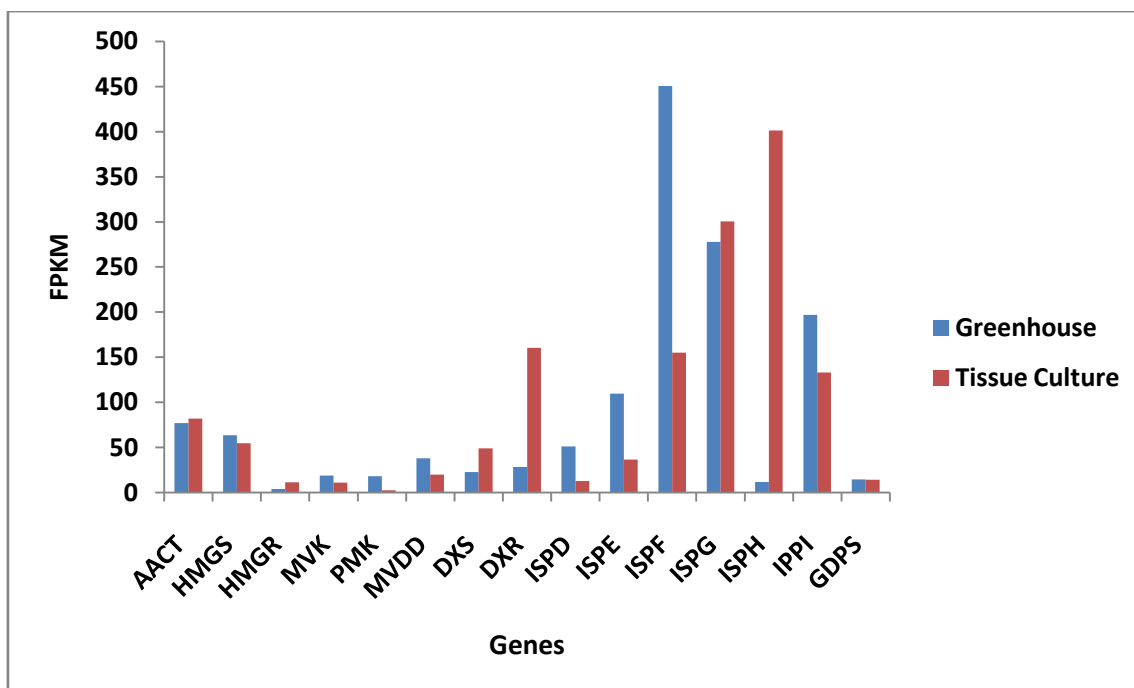


Fig. 2.16 Graphical representation of *in silico* transcript quantification for secondary metabolism (MVA and MEP) pathway genes in greenhouse versus tissue cultured transcriptomes

The transcript abundance analysis of MVA/MEP in *A. heterophyllum* revealed that four genes, namely *HDS*, *HMGR*, *MVK* and *MVDD* showed higher *in silico* expression in AHSR as compared to AHSS. It has been observed previously that phytosterols biosynthesis through the MVA pathway of isoprenogenesis is possible by regulating *HMGR* gene [88], also in *Arnebia euchroma* species this gene was known to be involved in shikonin plastidial monoterpenes biosynthesis [89]. Whereas, in *Catharanthus roseus* species for terpenoid indole alkaloids production genes such as *MVK* and *HDS* have played significant role [90, 91]. These observations support production of aconites biosynthesis in *A. heterophyllum* involving multiple genes of MVA/MEP pathways [87]. The rest of the genes showed higher *in silico* expression (FPKM) in the AHSS sample as compared AHSR sample, which could be due to their involvement in some other biological process, which has to be further functionally validated. These results are expected to further explore major genes for important agronomic traits in *A. heterophyllum*, and further understanding their regulatory mechanisms, especially for the production of medicinally important secondary metabolites.

In case of *S. chirayita* higher *in silico* expression was observed in SCFG compared to SCTC transcriptomes, which is obvious from the fact that they differ in mode of nutrition

i.e photoautotrophic versus photoheterotrophic mode of nutrition and since photoautotrophic mode of nutrition is combination of both soil and photosynthesis and thus photosynthesis plays an important role in maintaining the carbon pool of photosynthetic organisms, which regulates secondary metabolites and has been established in *Hypericum perforatum* [92].

Moreover, the genes showing higher transcript abundance could be the suitable potential targets for any genetic intervention strategies aimed towards enhancement of metabolite contents in *S. chirayita*. The FPKM based transcript abundance method has already been used in many plant species, including *Podophyllum hexandrum*, *Picrorhiza kurroa*, *Malus domestica* and *Camellia sinensis* etc. for determining the relative contribution of pathway genes in secondary metabolite production [93-95].

2.4.6 Mapping of transcriptomes on KEGG pathways

KEGG automatic annotation server (KAAS) was employed to map transcripts onto their biological pathways. BLAST bit score with a cut-off value of 60 (default) was used to identify homologs between the reference sequence and query sequence. KEGG orthology (KO) assignments were carried out based on the bi-directional best-hit method and genes whose BHR was greater than 0.95 (default) were selected. Each orthologous group was allocated a score so as to assign best k number to query genes. A total of 3,487 transcript sequences from AHSR and 3,177 transcript sequences from AHSS sample were assigned EC number using KAAS. These transcripts were further mapped to 337 unique pathways in AHSR transcriptomes as well as 333 unique pathways in AHSS transcriptome (Appendices Table A2, A3).

Whereas, in case of *S. chirayita* from 57,460 transcripts in SCFG, 8,690 transcripts sequences were mapped to KO which were further assigned to 342 KEGG pathways. Out of 8,690 transcripts, 830 (12.65%) were related to metabolic pathways, 351(5.35%) to the biosynthesis of secondary metabolites (Appendix Table A4). Similarly, in SCTC, out of 43,702 assembled transcripts, KO was assigned to 6,991 transcripts, which were mapped to 341 KEGG pathways. Out of 6,991 transcripts, 796 (12.86%) were related to metabolic pathways, 341 (5.50%) to the biosynthesis of secondary metabolites, 22 (0.355%) to citrate cycle (TCA cycle), 17 (0.274%) to the pentose phosphate pathway and 32 (0.51%) to glycolysis/gluconeogenesis (Appendix Table A5). The BRITE functional hierarchy

categorized the transcripts hierarchically thereby linking them with biological systems such as metabolism, genetic information processing and cellular processes in *S. chirayita*.

Moreover, the comparative KEGG analysis for within *A. heterophyllum* and *S. chirayita* datasets will help to unravel biological information (metabolic pathways, biochemical reactions, cellular processes, pathway interactions, etc.) hidden in the mass of data and will assist in gaining insight into the biology for both the species.

2.4.7 NCBI Biosystems for functional classification

All the transcripts from both the transcriptomes were further analyzed using NCBI Biosystems database. These transcripts were classified into five major categories, namely; genetic information processing, metabolism, cellular processes, organismal systems and environmental information processing. For AHSR, the highest number of transcripts belonged to the genetic information processing category (1821, 48.07%), which included transcription, translation, replication and repair, etc (Appendix Table A6). In case of AHSS sample metabolism was observed to be the biggest category (415, 48.25%), which included classes such as starch and sucrose metabolism, pyruvate metabolism, inositol phosphate metabolism, etc. (Appendix Table A7).

Highest number of transcripts belonged to the metabolism category (349, 46.65%) in SCFG transcriptome (Appendix Table A8). On the other hand, for SCTC, genetic information processing formed the biggest category SCTC (280, 47.61%) (Appendix Table A9). The flow diagram used for constructing network connectivity diagram in *A. heterophyllum* is given in Fig. 2.17. The in-house generated Perl code was employed for the construction of isoquinoline alkaloid biosynthesis network connectivity diagram in the root transcriptome of *A. heterophyllum* (Fig. 2.18).

In biology, biological networks have played a significant role to decipher important biological processes undertaking in an organism. Computational mapping of proteins onto a network can show the prevailing interconnections between them and can help to suggest the probable biological function [60]. Construction of isoquinoline alkaloids biosynthesis network connectivity diagram was carried out in the root transcriptome of *A. heterophyllum*. As reported by Malhotra in 2014, that the roots of *A. heterophyllum* are primary site for aconite biosynthesis and associated with its growth and development, the results of this study were found to be in agreement in with those. More number of genes

were mapped to the NCBI Biosystems pathways concluding that *A. heterophyllum* transcriptome generated in the study will act as a valuable source in near future.

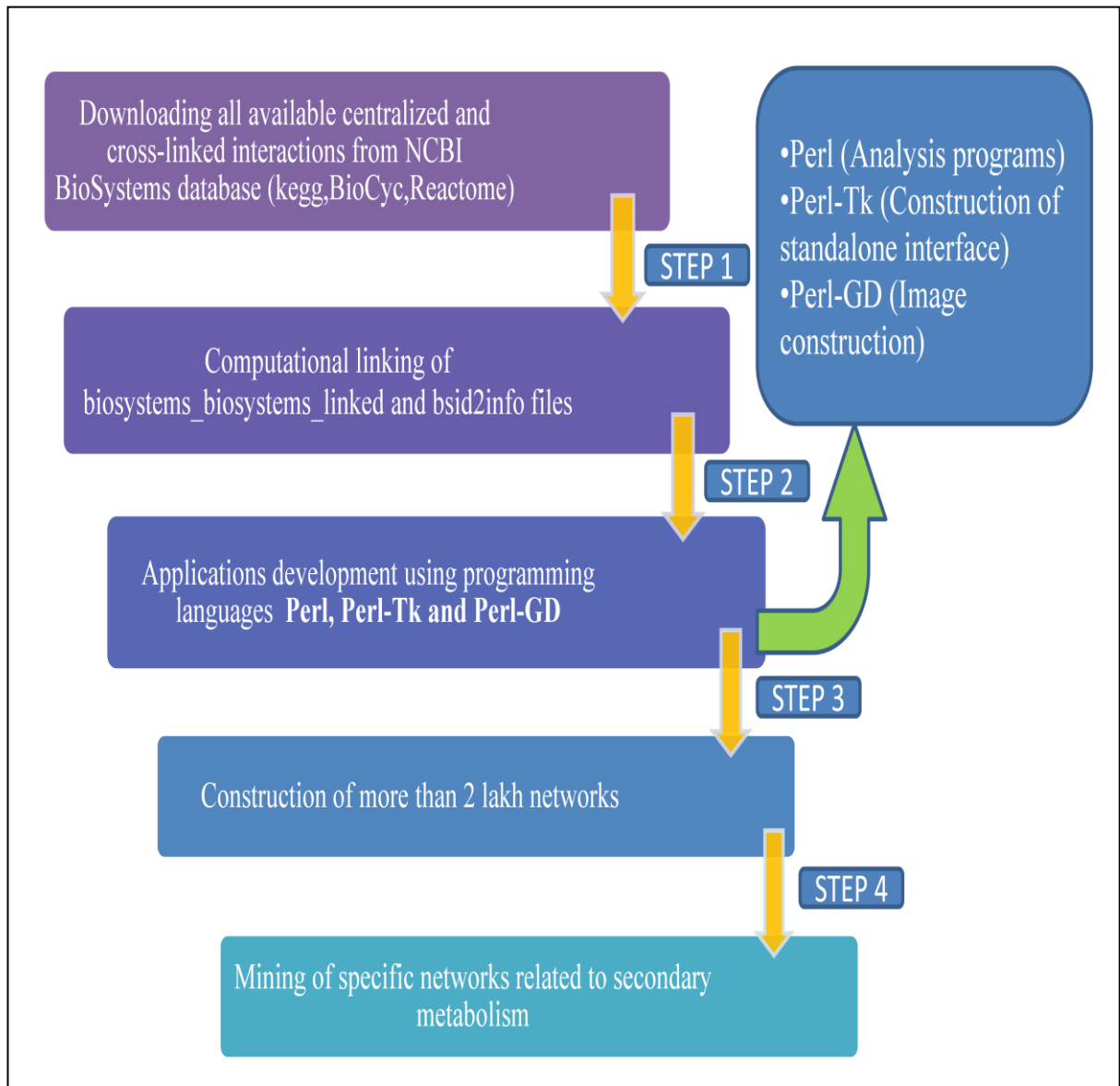


Fig. 2.17 Flow diagram used for constructing network connectivity diagram



Fig. 2.18 Construction of isoquinoline alkaloids biosynthesis network connectivity diagram in root transcriptome of *A. heterophyllum*

2.4.8 Conclusion

Although the knowledge available on atisine (aconites) biosynthesis is limited or incomplete, the results obtained from *A. heterophyllum* root and shoot transcriptomes will be of immense value in future. The roots versus shoots data corresponding to genes involved in MVA/MEP pathways which are uniquely present or abundant in both root and shoot transcriptomes can be used to plan a genetic intervention strategy. The substantial amount of transcripts obtained will certainly accelerate the understanding of the plant growth and development mechanism, along with providing new insights to increase the biomass yield. Furthermore, the construction of network connectivity diagrams will be the next big step to create a pipeline to investigate the regulation of various biological processes in *A. heterophyllum*. All the generated data sets need to be experimentally validated to ascertain the biological functions assigned through computational annotations. To the best of our knowledge, this is the first attempt to assemble and characterize the transcriptomes of *A. heterophyllum* using NGS technique.

REFERENCES

- [1] U. Schippmann, D. J. Leaman, and A. B. Cunningham, "Impact of cultivation and gathering of medicinal plants on biodiversity: global trends and issues," *Biodiversity and the Ecosystem Approach in Agriculture, Forestry and Fisheries*, 2002.
- [2] S. W. Pelletier, R. Aneja, and K. W. Gopinath, "The alkaloids of *Aconitum heterophyllum* Wall.: isolation and characterization," *Phytochemistry*, vol. 7, pp. 625-635, 1968.
- [3] Z. Wang, J. Wen, J. Xing, and Y. He, "Quantitative determination of diterpenoid alkaloids in four species of *Aconitum* by HPLC," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 40, pp. 1031-1034, 2006.
- [4] S. K. Mitra, A. Sachan, V. Udupa, S. J. Seshadri, and K. Jayakumar, "Histological changes in intestine in semichronic diarrhoea induced by lactose enriched diet in rats: Effect of Diarex-vet," vol. 78, pp. 212-216, 2003.
- [5] U. M. Thatte, N. N. Rege, S. D. Phatak, and S. A. Dahanukar, "The flip side of Ayurveda," *Journal of Postgraduate Medicine*, vol. 39, pp. 179, 1993.
- [6] B. P. Nautiyal, V. Prakash, R. Bahuguna, U. Maithani, H. Bisht, and M. C. Nautiyal, "Population study for monitoring the status of rarity of three Aconite species in Garhwal Himalaya," *Tropical Ecology*, vol. 43, pp. 297-303, 2002.
- [7] N. Srivastava, V. Sharma, A. K. Dobriyal, B. Kamal, S. Gupta, and V. S. Jadon, "Influence of pre-sowing treatments on in vitro seed germination of Ativisha (*Aconitum heterophyllum* Wall.) of Uttarakhand," *Biotechnology*, vol. 10, pp. 215-219, 2011.
- [8] M. Rodriguez-Concepcion and A. Boronat, "Elucidation of the methylerythritol phosphate pathway for isoprenoid biosynthesis in bacteria and plastids. A metabolic milestone achieved through genomics," *Plant Physiology*, vol. 130, pp. 1079-1089, 2002.
- [9] S. S. C. Iucn, "The IUCN Red List of Threatened Species, 1994-2007 version," ed: Switzerland, 2008.
- [10] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proceedings of the National Academy of Sciences*, vol. 74, pp. 560-564, 1977.

- [11] W. Ansorge, A. Rosenthal, B. Sproat, C. Schwager, J. Stegemann, and H. Voss, "Non-radioactive automated sequencing of oligonucleotides by chemical degradation," *Nucleic Acids Research*, vol. 16, pp. 2203-2206, 1988.
- [12] L. M. Smith, S. Fung, M. W. Hunkapiller, T. J. Hunkapiller, and L. E. Hood, "The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis," *Nucleic Acids Research*, vol. 13, pp. 2399-2412, 1985.
- [13] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, et al., "The sequence of the human genome," *Science*, vol. 291, pp. 1304-1351, 2001.
- [14] C. International Human Genome Sequencing, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931-945, 2004.
- [15] S. Tripathi, J. S. Jadaun, M. Chandra, and N. S. Sangwan, "Medicinal plant transcriptomes: the new gateways for accelerated understanding of plant secondary metabolism," *Plant Genetic Resources*, vol. 14, pp. 1-14, 2016.
- [16] Editorial, "Method of the Year," *Nature Methods*, vol. 5, pp. 1-1, 2008.
- [17] M. Barba, H. Czosnek, and A. Hadidi, "Historical perspective, development and applications of next-generation sequencing in plant virology," *Viruses*, vol. 6, pp. 106-136, 2014.
- [18] Roche, "History," Available: <http://sequencing.roche.com/about-us/history.html>, 2017.
- [19] Illumina, "History of Illumina Sequencing," Available: <https://www.illumina.com/technology/next-generation-sequencing/solexa-technology.html>, 2017.
- [20] T. F. Scientific, "Life Technologies Archive Details," Available: <http://ir.thermofisher.com/investors/news-and-events/news-releases/life-technologies-archive/life-technologies-archive-details/2011/Life-Technologies-Makes-DNA-Sequencing-More-Accessible-to-Laboratories-of-All-Sizes-Around-the-World/default.aspx>, 2017.
- [21] P. Biosciences, "About us," Available: <http://www.pacb.com/company/about-us/>, 2017.
- [22] O. Nanopore, "About us," Available: <https://nanoporetech.com/about-us>, 2017.
- [23] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt, "Assembling millions of short DNA sequences using SSAKE," *Bioinformatics*, vol. 23, pp. 500-501, 2007.

- [24] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, et al., "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler," *Gigascience*, vol. 1, pp. 18, 2012.
- [25] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and A. N. Birol, "ABySS: A parallel assembler for short read sequence data," *Genome Research*, vol. 19, pp. 1117-1123, 2009.
- [26] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome Research*, vol. 18, pp. 821-829, 2008.
- [27] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces using Phred. I. Accuracy assessment," *Genome Research*, vol. 8, pp. 175-185, 1998.
- [28] D. S. Horner, G. Pavesi, T. Castrignano, P. D. O. De Meo, S. Liuni, M. Sammeth, et al., "Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing," *Briefings in Bioinformatics*, vol. 11, pp. 181-197, 2009.
- [29] B. Bioinformatics, "FastQC A quality control tool for high throughput sequence data," *Cambridge, UK: Babraham Institute*, 2011.
- [30] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, pp. 2114-2120, 2014.
- [31] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, et al., "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology*, vol. 29, pp. 644-652, 2011.
- [32] W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen, "A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies," *Plos One*, vol. 6, pp. 1-12, 2011.
- [33] N. R. Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 44, pp. D7-D19, 2016.
- [34] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Research*, vol. 28, pp. 45-48, 2000.
- [35] C. Wu and D. W. Nebert, "Update on genome completion and annotations: Protein Information Resource," *Human Genomics*, vol. 1, pp. 229, 2004.
- [36] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, pp. 235-242, 2000.

- [37] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciuffo, D. Haddad, R. McVeigh, et al., "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation," *Nucleic Acids Research*, vol. 44, pp. D733-D745, 2016.
- [38] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, et al., "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25-29, 2000.
- [39] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, "The COG database: a tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, pp. 33-36, 2000.
- [40] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, et al., "The COG database: an updated version includes eukaryotes," *BMC Bioinformatics*, vol. 4, pp. 41-41, 2003.
- [41] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57-63, 2009.
- [42] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, et al., "De novo assembly and analysis of RNA-seq data," *Nature Methods*, vol. 7, pp. 909-912, 2010.
- [43] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, pp. R106, 2010.
- [44] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139-140, 2010.
- [45] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, et al., "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs," *Nature Biotechnology*, vol. 28, pp. 503-510, 2010.
- [46] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, pp. 511-515, 2010.
- [47] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, p. 323, 2011.

- [48] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, pp. 621-628, 2008.
- [49] C. H. Schilling and B. O. Palsson, "The underlying pathway structure of biochemical reaction networks," *Proceedings of the National Academy of Sciences*, vol. 95, pp. 4193-4198, 1998.
- [50] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, pp. D353-D361, 2016.
- [51] R. Caspi, R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, et al., "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 44, pp. D471-D480, 2016.
- [52] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, et al., "The Reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 42, pp. D472-D477, 2014.
- [53] M. Kutmon, A. Riutta, N. Nunes, K. Hanspers, Egon L. Willighagen, A. Bohler, et al., "WikiPathways: capturing the full diversity of pathway knowledge," *Nucleic Acids Research*, vol. 44, pp. D488-D494, 2016.
- [54] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Research*, vol. 42, pp. D199-D205, 2014.
- [55] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa, "KAAS: an automatic genome annotation and pathway reconstruction server," *Nucleic Acids Research*, vol. 35, pp. W182-W185, 2007.
- [56] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, et al., "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 40, pp. D742-D753, 2012.
- [57] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, et al., "The Reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 42, pp. D472-D477, 2014.
- [58] L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, et al., "The NCBI Biosystems database," *Nucleic Acids Research*, vol. 38, pp. D492-D496, 2009.

- [59] A. Maayan, "Introduction to network analysis in systems biology," *Science Signaling*, vol. 4, pp. 190, 2011.
- [60] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular Systems Biology*, vol. 8, pp. 565, 2012.
- [61] A. Blais and B. D. Dynlacht, "Constructing transcriptional regulatory networks," *Genes & Development*, vol. 19, pp. 1499-1511, 2005.
- [62] D. Chasman, Y. H. Ho, D. B. Berry, C. M. Nemecek, M. E. MacGilvray, J. Hose, et al., "Pathway connectivity and signaling coordination in the yeast stress-activated signaling network," *Molecular Systems Biology*, vol. 10, pp. 759, 2014.
- [63] N. C. Shah, "Conservation aspects of *Aconitum* species in the Himalayas with special reference to Uttaranchal (India)," *Medicinal Plant Conservation*, vol. 11, pp. 9-15, 2005.
- [64] M. B. Nariya, P. Parmar, V. J. Shukla, and B. Ravishankar, "Toxicological study of Balacaturbhadrika churna," *Journal of Ayurveda and Integrative Medicine*, vol. 2, pp. 79, 2011.
- [65] S. K. Prasad, D. Jain, D. K. Patel, A. N. Sahu, and S. Hemalatha, "Antisecretory and antimotility activity of *Aconitum heterophyllum* and its significance in treatment of diarrhea," *Indian Journal of Pharmacology*, vol. 46, pp. 82, 2014.
- [66] Y. M. Sinam, S. Kumar, S. Hajare, S. Gautam, G. A. S. Devi, and A. Sharma, "Antibacterial property of *Aconitum heterophyllum* root alkaloid," *International Journal*, vol. 2, pp. 839-844, 2014.
- [67] S. K. Prasad, R. Kumar, D. K. Patel, A. N. Sahu, and S. Hemalatha, "Physicochemical standardization and evaluation of in-vitro antioxidant activity of *Aconitum heterophyllum* Wall," *Asian Pacific Journal of Tropical Biomedicine*, vol. 2, pp. S526-S531, 2012.
- [68] M. D. Ukani, N. K. Mehta, and D. D. Nanavati, "*Aconitum heterophyllum* (ativisha) in ayurveda," *Ancient Science of Life*, vol. 16, pp. 166, 1996.
- [69] P. A. Lone and A. K. Bhardwaj, "Potent medicinal herbs used traditionally for the treatment of Arthritis in Bandipora, Kashmir," *International Journal of Recent Scientific Research*, vol. 4, pp. 1766-1770, 2013.
- [70] A. K. Subash and A. Augustine, "Hypolipidemic effect of methanol fraction of *Aconitum heterophyllum* wall ex Royle and the mechanism of action in diet-induced

- obese rats," *Journal of Advanced Pharmaceutical Technology & Research*, vol. 3, pp. 224, 2012.
- [71] P. Joshi and V. Dhawan, "Swertia chirayita- an overview," *Current Science-Bangalore-*, vol. 89, pp. 635, 2005.
- [72] J. A. Bhat, M. Kumar, A. K. Negi, and N. P. Todaria, "Informants consensus on ethnomedicinal plants in Kedarnath Wildlife Sanctuary of Indian Himalayas," *Journal of Medicinal Plants Research*, vol. 7, pp. 148-154, 2013.
- [73] V. Kumar and J. Van Staden, "A review of *Swertia chirayita* (Gentianaceae) as a traditional medicinal plant," *Frontiers in Pharmacology*, vol. 6, pp. 1-14, 2015.
- [74] D. Pal, S. Sur, S. Mandal, A. Das, A. Roy, S. Das, et al., "Prevention of liver carcinogenesis by amarogentin through modulation of G1/S cell cycle check point and induction of apoptosis," *Carcinogenesis*, vol. 33, pp. 2424- 2431, 2012.
- [75] R. Arya, S. K. Sharma, and S. Singh, "Antidiabetic effect of whole plant extract and fractions of *Swertia chirayita* Buch.-Ham," *Planta Medica*, vol. 77, pp. 138, 2011.
- [76] A. Laxmi, S. Siddhartha, and M. Archana, "Antimicrobial screening of methanol and aqueous extracts of *Swertia chirata*," *International Journal of Pharmacy and Pharmaceutical Sciences*, vol. 3, pp. 142-146, 2011.
- [77] M. Sarker, S. C. Das, S. K. Saha, Z. A. Mahmud, and S. C. Bachar, "Analgesic and Anti-inflammatory Activities of Flower Extracts of *Punica granatum* Linn.(Punicaceae)," vol. 2, pp. 133, 2012.
- [78] Y. Chen, B. Huang, J. He, L. Han, Y. Zhan, and Y. Wang, "In vitro and in vivo antioxidant effects of the ethanolic extract of *Swertia chirayita*," *Journal of Ethnopharmacology*, vol. 136, pp. 309-315, 2011.
- [79] H. Verma, P. R. Patil, R. M. Kolhapure, and V. Gopalkrishna, "Antiviral activity of the Indian medicinal plant extract, *Swertia chirata* against herpes simplex viruses: A study by in-vitro and molecular approach," *Indian Journal of Medical Microbiology*, vol. 26, p. 322, 2008.
- [80] A. Kossel, "Ueber die chemische Zusammensetzung der Zelle," *Du Bois-Reymonds Archiv Anatomie Physiologie Abt*, pp. 181-186, 1891.
- [81] A. G. Medentsev and V. K. Akimenko, "Naphthoquinone metabolites of the fungi," *Phytochemistry*, vol. 47, pp. 935-959, 1998.

- [82] A. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, pp. 3674-3676, 2005.
- [83] EMBL, "What are protein domains?," Available: <https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification/what-are-protein-domains>, 2017.
- [84] D. Bhattacharyya, R. Sinha, S. Hazra, R. Datta, and S. Chattopadhyay, "De novo transcriptome analysis using 454 pyrosequencing of the Himalayan Mayapple, *Podophyllum hexandrum*," *BMC Genomics*, vol. 14, pp. 748, 2013.
- [85] P. Gahlan, H. R. Singh, R. Shankar, N. Sharma, A. Kumari, V. Chawla, et al., "De novo sequencing and characterization of *Picrorhiza kurrooa* transcriptome at two temperatures showed major transcriptome adjustments," *BMC Genomics*, vol. 13, pp. 126, 2012.
- [86] S. Pradhan, N. Bandhiwal, N. Shah, C. Kant, R. Gaur, and S. Bhatia, "Global transcriptome analysis of developing chickpea (*Cicer arietinum* L.) seeds," *Frontiers in Plant Science*, vol. 5, pp. 698, 2014.
- [87] N. Malhotra, V. Kumar, H. Sood, T. R. Singh, and R. S. Chauhan, "Multiple genes of mevalonate and non-mevalonate pathways contribute to high aconites content in an endangered medicinal herb, *Aconitum heterophyllum* Wall," *Phytochemistry*, vol. 108, pp. 26-34, 2014.
- [88] I. Nogues, F. Brilli, and F. Loreto, "Dimethylallyl diphosphate and geranyl diphosphate pools of plant species characterized by different isoprenoid emissions," *Plant Physiology*, vol. 141, pp. 721-730, 2006.
- [89] R. S. Singh, R. K. Gara, P. K. Bhardwaj, A. Kaachra, S. Malik, R. Kumar, et al., "Expression of 3-hydroxy-3-methylglutaryl-CoA reductase, p-hydroxybenzoate-m-geranyltransferase and genes of phenylpropanoid pathway exhibits positive correlation with shikonins content in arnebia [*Arnebia euchroma* (Royle) Johnston]," *BMC Molecular Biology*, vol. 11, pp. 88, 2010.
- [90] O. Ginis, V. Courdavault, C. I. Melin, A. Lanoue, N. Giglioli-Guivarc'h, B. St-Pierre, et al., "Molecular cloning and functional characterization of *Catharanthus roseus* hydroxymethylbutenyl 4-diphosphate synthase gene promoter from the methyl erythritol phosphate pathway," *Molecular Biology Reports*, vol. 39, pp. 5433-5447, 2012.

- [91] A. E. Schulte, E. M. L. Duran, R. van der Heijden, and R. Verpoorte, "Mevalonate kinase activity in *Catharanthus roseus* plants and suspension cultured cells," *Plant Science*, vol. 150, pp. 59-69, 2000.
- [92] K. Mosaleeyanon, S. M. A. Zobayed, F. Afreen, and T. Kozai, "Relationships between net photosynthetic rate and secondary metabolite contents in St. John's wort," *Plant Science*, vol. 169, pp. 523-531, 2005.
- [93] P. Kumar, T. Pal, N. Sharma, V. Kumar, H. Sood, and R. S. Chauhan, "Expression analysis of biosynthetic pathway genes vis-à-vis podophyllotoxin content in *Podophyllum hexandrum* Royle," *Protoplasma*, vol. 252, pp. 1253-1262, 2015.
- [94] A. Paul, A. Jha, S. Bhardwaj, S. Singh, R. Shankar, and S. Kumar, "RNA-seq-mediated transcriptome analysis of actively growing and winter dormant shoots identifies non-deciduous habit of evergreen tree tea during winters," *Scientific Reports*, vol. 4, pp. 5932, 2014.
- [95] A. K. Singh, V. Sharma, A. K. Pal, V. Acharya, and P. S. Ahuja, "Genome-wide organization and expression profiling of the NAC transcription factor family in potato (*Solanum tuberosum* L.)," *DNA Research*, vol. 20, pp. 403-423, 2013.

CONCLUSION AND FUTURE PROSPECTS

The current study has two major outcomes; firstly, the developed tool has bridged the research gap by providing a set of influential features for predicting *R* proteins through feature extraction method. The tool DRPPP is robust and considers functionally validated disease resistance proteins and randomly generated negative dataset for its model building. The tool DRPPP has produced high accuracy (91.11%) as compared to other existing tools in this domain justifying its high reliability.

Secondly, generation and computational analysis of these two important endangered plant species, i.e. *A. heterophyllum* and *S. chirayita* transcriptomes has predicted the candidate genes involved in the production of secondary metabolites. The *in silico* expression profiling for these 15 genes in *A. heterophyllum* has identified four genes, namely *HDS*, *HMGR*, *MVK* and *MVDD* with higher expression in AHSR (root) as compared to AHSS (shoot) transcriptomes. The pathway analysis performed for both the tissues of *Aconitum* suggested that 341 number of mapped Kos in AHSR and 329 for AHSS are responsible for secondary metabolism, which attributes medicinal value to this plant. In total 77 interacting pathways associated with isoquinoline alkaloids biosynthesis were identified in root transcriptome of *A. heterophyllum* indicating how important primary and secondary metabolic pathways are connected with each other. Whereas, in case of *S. chirayita* higher *in silico* transcript abundance was observed for nine genes in SCFG as compared to SCTC transcriptomes suggesting them as suitable target for planning genetic intervention strategies aimed towards enhancement of metabolite contents in *S. chirayita*. The total number of 351 transcripts was related to biosynthesis of secondary metabolites in SCFG transcriptome, whereas 341 in SCTC transcriptomes, which will be helpful in understating the complex mechanism, associated with biosynthesis in differential conditions.

The future prospect associated with the above tool is that the predicted *R* proteins can accelerate the process of genetic improvement and breeding programs for the development of disease-resistant varieties in plants. The used method is robust enough to accommodate large datasets. Alternatively, methods such as HMM, ANN, Decision trees, Random forests and Bayesian networks can also be used in future to efficiently predict *R* proteins. Whereas, the genes predicted from the transcriptome analysis can be functionally validated aimed for enhancing the production of secondary metabolites in these two plant species. This current

study, is first *in silico* report on deciphering the important molecular components implicated in differential conditions or differential modes of nutrition required for secondary metabolites production in these plant species.

APPENDICES

Table A1 Reference *R* genes (manually curated) of PRGDB

Sr. no.	PRGID	Name	Species	Class	GenBank ID	GenBank Locus
1	PRGDB000000 29	Asc-1	<i>Solanum lycopersicum</i>	Other	7688741	AF198177
2	PRGDB000000 30	Bs2	<i>Capsicum chacoense</i>	CNL	6456754	AF202179
3	PRGDB000000 31	Bs4	<i>Solanum lycopersicum</i>	TNL	38489218	AY438027
4	PRGDB000000 32	Cf-2	<i>Solanum pimpinellifolium</i>	RLP	1184074	U42444
5	PRGDB000000 33	Cf-4	<i>Solanum habrochaites</i>	RLP	2808679	AJ002235
6	PRGDB000000 34	Cf4A	<i>Solanum habrochaites</i>	RLP	2808679	AJ002235
7	PRGDB000000 35	Cf-5	<i>Solanum lycopersicum</i> <i>var. cerasiforme</i>	RLP	3894382	AF053993
8	PRGDB000000 36	Cf-9	<i>Solanum pimpinellifolium</i>	RLP	2792183	AJ002236
9	PRGDB000000 37	Cf9B	<i>Solanum pimpinellifolium</i>	RLP	2792183	AJ002236
10	PRGDB000000 38	Gpa2	<i>Solanum tuberosum</i>	CNL	6164968	AF195939
11	PRGDB000000 39	Gro1.4	<i>Solanum tuberosum</i>	TNL	37781225	AY196151
12	PRGDB000000 40	Hero	<i>Solanum lycopersicum</i>	CNL	23095860	AJ457052
13	PRGDB000000 41	I-2	<i>Solanum lycopersicum</i>	NL	4689222	AF118127
14	PRGDB000000 42	Mi1.2	<i>Solanum lycopersicum</i>	CNL	3449379	AF039682
15	PRGDB000000 43	Pto	<i>Solanum pimpinellifolium</i>	Kinase	430991	U02271
16	PRGDB000000 44	N	<i>Nicotiana glutinosa</i>	TNL	558886	U15605

17	PRGDB000000 45	Prf	<i>Solanum pimpinellifolium</i>	CNL	8547226	AF220602
18	PRGDB000000 46	R1	<i>Solanum demissum</i>	CNL	17432422	AF447489
19	PRGDB000000 47	R3a	<i>Solanum tuberosum</i>	NL	57233496	AY849382
20	PRGDB000000 48	Rpi-blb1	<i>Solanum bulbocastanum</i>	CNL	32693280	AY336128
21	PRGDB000000 49	Rpi-blb2	<i>Solanum bulbocastanum</i>	CNL	74040323	DQ122125
22	PRGDB000000 50	Rx	<i>Solanum tuberosum</i>	CNL	5524753	AJ011801
23	PRGDB000000 51	Rx2	<i>Solanum acaule</i>	CNL	5918253	AJ249448
24	PRGDB000000 52	RY-1	<i>Solanum tuberosum subsp andigena</i>	TNL	16944810	AJ300266
25	PRGDB000000 53	Sw-5	<i>Solanum lycopersicum</i>	CNL	15418708	AY007366
26	PRGDB000000 54	Tm-2a	<i>Solanum lycopersicum</i>	CNL	33330975	AF536201
27	PRGDB000000 55	Ve1	<i>Solanum lycopersicum</i>	RLP	14279669	AF272367
28	PRGDB000000 56	Ve2	<i>Solanum lycopersicum</i>	RLP	14269076	AF365929
29	PRGDB000000 57	Tm-2	<i>Solanum lycopersicum</i>	CNL	33621254	AF536200
30	PRGDB000004 80	Dm3 (RGC2B)	<i>Lactuca sativa</i>	CNL	4106969	AH007213
31	PRGDB000004 81	At1	<i>Cucumis melo</i>	Other	18032027	AY066012
32	PRGDB000004 82	At2	<i>Cucumis melo</i>	Other	18158220	AF461048
33	PRGDB000004 83	Hm1	<i>Zea mays</i>	Other	16246322 7	NM_0011124 50
34	PRGDB000004 84	HRT	<i>Arabidopsis thaliana</i>	CNL	7110564	AF234174

35	PRGDB000004 85	Hs1	<i>Beta vulgaris</i>	Other	1850967	U79733
36	PRGDB000004 86	L6	<i>Linum usitatissimum</i>	TNL	862903	U27081
37	PRGDB000004 87	M	<i>Linum usitatissimum</i>	TNL	1842250	U73916
38	PRGDB000004 88	Pi-ta	<i>Oryza sativa</i>	CNL	28629808	AY196754
39	PRGDB000004 89	RCY1	<i>Arabidopsis thaliana</i>	CNL	29603481	AB087829
40	PRGDB000004 90	RFO1	<i>Arabidopsis thaliana</i>	RLK	14533776 6	NM_106616
41	PRGDB000004 91	RPM1	<i>Arabidopsis thaliana</i>	CNL	30680118	NM_111584
42	PRGDB000004 92	RPP13	<i>Arabidopsis thaliana</i>	CNL	30692728	NM_114520
43	PRGDB000004 93	RPP5	<i>Arabidopsis thaliana</i>	TNL	18651193 8	NM_117798
44	PRGDB000004 94	RPP8	<i>Arabidopsis thaliana</i>	CNL	14535880 7	NM_123713
45	PRGDB000004 95	Rps2	<i>Arabidopsis thaliana</i>	CNL	30687102	NM_118742
46	PRGDB000004 96	Rps4	<i>Arabidopsis thaliana</i>	TNL	18422530	NM_123893
47	PRGDB000004 97	RPS5	<i>Arabidopsis thaliana</i>	CNL	14533542 0	NM_101094
48	PRGDB000004 98	xa21	<i>Oryza sativa</i>	RLK	94481122	AB212799
49	PRGDB000004 99	Hm2	<i>Zea mays</i>	Other	16587542 7	EU367521
50	PRGDB000005 00	Rps1-k- 2	<i>Glycine max</i>	NL	18339649 6	EU450800
51	PRGDB000005 01	Rps1-k- 1	<i>Glycine max</i>	NL	18339649 6	EU450800
52	PRGDB000357 12	LeEIX1	<i>Solanum lycopersicum</i>	RLP	39577519	AY359965
53	PRGDB000357 13	LeEIX2	<i>Solanum lycopersicum</i>	RLP	39577521	AY359966

54	PRGDB000357 14	Bs3	<i>Capsicum annuum</i>	Other	15885151 6	EU078684
55	PRGDB000357 15	Bs3-E	<i>Capsicum annuum</i>	Other	15885151 2	EU078683
56	PRGDB000357 18	FLS2	<i>Arabidopsis thaliana</i>	RLK	42568348	NM_124003
57	PRGDB000357 19	EFR	<i>Arabidopsis thaliana</i>	RLK	14535826 9	NM_122055
58	PRGDB000357 20	PEPR1	<i>Arabidopsis thaliana</i>	RLK	18410260	NM_105966
59	PRGDB000357 21	ER - Erecta	<i>Arabidopsis thaliana</i>	RLK	30683082	NM_128190
60	PRGDB000357 22	PGIP	<i>Phaseolus vulgaris</i>	RLP	21028	X64769
61	PRGDB000357 23	Mlo	<i>Hordeum vulgare</i>	Other	1877220	Z83834
62	PRGDB000357 24	P2	<i>Linum usitatissimum</i>	TNL	13517467	AF310960
63	PRGDB000357 25	MLA10	<i>Hordeum vulgare</i>	CNL	33943719	AY266445
64	PRGDB000357 26	RPG1	<i>Hordeum vulgare</i>	RLK	11762192 5	DQ854803
65	PRGDB000509 57	PIB	<i>Oryza sativa</i>	CNL	4689079	AB013449
66	PRGDB000509 58	XA1	<i>Oryza sativa</i>	CNL	2943741	AB002266
67	PRGDB000509 59	RPP1	<i>Arabidopsis thaliana</i>	TNL	30692150	NM_114316
68	PRGDB000509 60	RPP4	<i>Arabidopsis thaliana</i>	TNL	42566890	NM_117790
69	PRGDB000509 61	RPW8.1	<i>Arabidopsis thaliana</i>	Other	12958161	AF273059
70	PRGDB000509 62	RPW8.2	<i>Arabidopsis thaliana</i>	Other	12958161	AF273059
71	PRGDB000509 63	RRS1	<i>Arabidopsis thaliana</i>	TNL	14533473 8	NM_0010852 46
72	PRGDB000509 64	RTM1	<i>Arabidopsis thaliana</i>	Other	30679421	NM_100456

73	PRGDB000509 65	RTM2	<i>Arabidopsis thaliana</i>	Other	30680686	NM_120571
74	PRGDB000614 32	MLA1	<i>Hordeum vulgare</i>	CNL	27026780 4	GU245961
75	PRGDB000614 33	Mla6	<i>Hordeum vulgare subsp. vulgare</i>	CNL	12957125	AJ302293
76	PRGDB000614 34	Mla12	<i>Hordeum vulgare subsp. vulgare</i>	CNL	28565621	AY196347
77	PRGDB000614 35	MLA13	<i>Hordeum vulgare</i>	CNL	27464252	AF523683
78	PRGDB000614 36	Pi36	<i>Oryza sativa Indica Group</i>	CNL	11432951 7	DQ900896
79	PRGDB000614 37	Rp1-D	<i>Zea mays</i>	CNL	5702195	AF107293
80	PRGDB000614 38	Pm3	<i>Triticum aestivum</i>	CNL	37624723	AY325736
81	PRGDB000614 39	Lr10	<i>Triticum aestivum</i>	CNL	33302326	AY270157
82	PRGDB000614 40	PI8	<i>Helianthus annuus</i>	CNL	47777762	AY490793
83	PRGDB000614 41	SSI4	<i>Arabidopsis thaliana</i>	TNL	27466163	AY179750
84	PRGDB000614 42	Pi9	<i>Oryza sativa Indica Group</i>	CNL	83571777	DQ285630
85	PRGDB000614 43	Piz-t	<i>Oryza sativa Japonica Group</i>	CNL	85682843	DQ352040
86	PRGDB000614 44	Pi2	<i>Oryza sativa Indica Group</i>	CNL	86361422	DQ352453
87	PRGDB000614 46	Cre1	<i>Aegilops tauschii</i>	CNL	22252945	AY124651
88	PRGDB000614 47	RPP27	<i>Arabidopsis thaliana</i>	RLP	42516773	AJ585978
89	PRGDB000614 48	Xa26	<i>Oryza sativa Japonica Group</i>	RLK	90018762	DQ426646
90	PRGDB000614 49	Xa5	<i>Oryza sativa Indica Group</i>	Other	55585038	AY643716

91	PRGDB000614 50	xa27	<i>Oryza sativa</i> <i>Indica Group</i>	Other	66735941	AY986491
92	PRGDB000614 51	Xa13	<i>Oryza sativa</i> <i>Indica Group</i>	Other	89892339	DQ421396
93	PRGDB000614 52	IVR	<i>Nicotiana</i> <i>tabacum</i>	Other	3355639	AJ009684
94	PRGDB000614 53	Pikm1- TS	<i>Oryza sativa</i> <i>Japonica Group</i>	CNL	20736732 9	AB462324
95	PRGDB000614 54	Pikm2- TS	<i>Oryza sativa</i> <i>Japonica Group</i>	CNL	20736733 1	AB462325
96	PRGDB000614 55	Pid2	<i>Oryza sativa</i> <i>Indica Group</i>	Other	23782412 9	FJ915121
97	PRGDB000614 56	Rdg2a	<i>Hordeum</i> <i>vulgare subsp.</i> <i>vulgare</i>	CNL	30101547 9	HM124452
98	PRGDB000614 57	Pid3	<i>Oryza sativa</i> <i>Japonica Group</i>	CNL	22503080 1	FJ773286
99	PRGDB000614 58	Pi5-1	<i>Oryza sativa</i> <i>Japonica Group</i>	CNL	21395860 0	EU869185
100	PRGDB000614 59	Pi5-2	<i>Oryza sativa</i> <i>Japonica Group</i>	CNL	21395860 2	EU869186
101	PRGDB000614 60	Pit	<i>Oryza sativa</i> <i>Japonica Group</i>	CNL	22244646 3	AB379816
102	PRGDB000614 61	Serk3A	<i>Nicotiana</i> <i>benthamiana</i>	RLK	30975475 8	HQ332144
103	PRGDB000614 62	Serk3B	<i>Nicotiana</i> <i>benthamiana</i>	RLK	30975476 0	HQ332145
104	PRGDB000614 63	Pikp-2	<i>Oryza sativa</i> <i>Japonica Group</i>	CNL	31965578 1	HM035360
105	PRGDB000614 64	KR1	<i>Glycine max</i>	TNL	18033110	AF327903
106	PRGDB000614 65	FOM-2	<i>Cucumis melo</i>	CNL	82794017	DQ287965
107	PRGDB000614 66	RLM3	<i>Arabidopsis</i> <i>thaliana</i>	TN	79325134	NM_0010365 75
108	PRGDB000614 67	RAC1	<i>Arabidopsis</i> <i>thaliana</i>	TNL	41387773	AY522496
109	PRGDB000614	Lr21	<i>Triticum</i>	CNL	22642482	FJ876280

	68		<i>aestivum</i>		5	
110	PRGDB000614 69	Lr1	<i>Triticum aestivum</i>	CNL	15206078 5	EF439840
111	PRGDB000614 70	Lr34	<i>Triticum aestivum</i>	Other	30113079 4	HM775493
112	PRGDB000614 71	VAT	<i>Cucumis melo</i>	CNL		

Table A2 KOs associated with secondary metabolism in AHSR

Sr. no.	KO associated enzymes
1	"ko:K00001 E1.1.1.1; alcohol dehydrogenase [EC:1.1.1.1]"
2	"ko:K00003 E1.1.1.3; homoserine dehydrogenase [EC:1.1.1.3]"
3	"ko:K00012 UGDH; UDPglucose 6-dehydrogenase [EC:1.1.1.22]"
4	"ko:K00013 hisD; histidinol dehydrogenase [EC:1.1.1.23]"
5	"ko:K00016 LDH; L-lactate dehydrogenase [EC:1.1.1.27]"
6	"ko:K00021 HMGCR; hydroxymethylglutaryl-CoA reductase (NADPH) [EC:1.1.1.34]"
7	"ko:K00024 mdh; malate dehydrogenase [EC:1.1.1.37]"
8	"ko:K00025 MDH1; malate dehydrogenase [EC:1.1.1.37]"
9	"ko:K00026 MDH2; malate dehydrogenase [EC:1.1.1.37]"
10	"ko:K00030 IDH3; isocitrate dehydrogenase (NAD+) [EC:1.1.1.41]"
11	"ko:K00031 IDH1; isocitrate dehydrogenase [EC:1.1.1.42]"
12	"ko:K00033 PGD; 6-phosphogluconate dehydrogenase [EC:1.1.1.44]"
13	"ko:K00036 G6PD; glucose-6-phosphate 1-dehydrogenase [EC:1.1.1.49]"
14	"ko:K00052 leuB; 3-isopropylmalate dehydrogenase [EC:1.1.1.85]"
15	"ko:K00053 ilvC; ketol-acid reductoisomerase [EC:1.1.1.86]"
16	"ko:K00083 E1.1.1.195; cinnamyl-alcohol dehydrogenase [EC:1.1.1.195]"
17	"ko:K00088 guaB; IMP dehydrogenase [EC:1.1.1.205]"
18	"ko:K00099 dxr; 1-deoxy-D-xylulose-5-phosphate reductoisomerase [EC:1.1.1.267]"
19	"ko:K00106 XDH; xanthine dehydrogenase/oxidase [EC:1.1.1.4 1.17.3.2]"
20	"ko:K00121 frmA; S-(hydroxymethyl)glutathione dehydrogenase / alcohol

	dehydrogenase [EC:1.1.1.284 1.1.1.1]"
21	"ko:K00128 E1.2.1.3; aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]"
22	"ko:K00133 asd; aspartate-semialdehyde dehydrogenase [EC:1.2.1.11]"
23	"ko:K00134 GAPDH; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12]"
24	"ko:K00145 argC; N-acetyl-gamma-glutamyl-phosphate reductase [EC:1.2.1.38]"
25	"ko:K00147 proA; glutamate-5-semialdehyde dehydrogenase [EC:1.2.1.41]"
26	"ko:K00161 PDHA; pyruvate dehydrogenase E1 component alpha subunit [EC:1.2.4.1]"
27	"ko:K00162 PDHB; pyruvate dehydrogenase E1 component beta subunit [EC:1.2.4.1]"
28	"ko:K00164 OGDH; 2-oxoglutarate dehydrogenase E1 component [EC:1.2.4.2]"
29	"ko:K00166 BCKDHA; 2-oxoisovalerate dehydrogenase E1 component alpha subunit [EC:1.2.4.4]"
30	"ko:K00167 BCKDHB; 2-oxoisovalerate dehydrogenase E1 component beta subunit [EC:1.2.4.4]"
31	"ko:K00213 DHCR7; 7-dehydrocholesterol reductase [EC:1.3.1.21]"
32	"ko:K00215 dapB; 4-hydroxy-tetrahydrodipicolinate reductase [EC:1.17.1.8]"
33	"ko:K00218 E1.3.1.33; protochlorophyllide reductase [EC:1.3.1.33]"
34	"ko:K00222 TM7SF2; delta14-sterol reductase [EC:1.3.1.70]"
35	"ko:K00225 GLDH; L-galactono-1,4-lactone dehydrogenase [EC:1.3.2.3]"
36	"ko:K00227 SC5DL; lathosterol oxidase [EC:1.14.21.6]"
37	"ko:K00228 CPOX; coproporphyrinogen III oxidase [EC:1.3.3.3]"
38	"ko:K00231 PPOX; oxygen-dependent protoporphyrinogen oxidase [EC:1.3.3.4]"
39	"ko:K00234 SDHA; succinate dehydrogenase (ubiquinone) flavoprotein subunit [EC:1.3.5.1]"
40	"ko:K00235 SDHB; succinate dehydrogenase (ubiquinone) iron-sulfur subunit [EC:1.3.5.1]"
41	"ko:K00249 ACADM; acyl-CoA dehydrogenase [EC:1.3.8.7]"
42	"ko:K00257 E1.3.99.-"
43	"ko:K00263 E1.4.1.9; leucine dehydrogenase [EC:1.4.1.9]"

44	"ko:K00264 GLT1; glutamate synthase (NADPH/NADH) [EC:1.4.1.13 1.4.1.14]"
45	"ko:K00265 gltB; glutamate synthase (NADPH/NADH) large chain [EC:1.4.1.13 1.4.1.14]"
46	"ko:K00266 gltD; glutamate synthase (NADPH/NADH) small chain [EC:1.4.1.13 1.4.1.14]"
47	"ko:K00276 AOC3; primary-amine oxidase [EC:1.4.3.21]"
48	"ko:K00286 E1.5.1.2; pyrroline-5-carboxylate reductase [EC:1.5.1.2]"
49	"ko:K00318 PRODH; proline dehydrogenase [EC:1.5.-.-]"
50	"ko:K00326 E1.6.2.2; cytochrome-b5 reductase [EC:1.6.2.2]"
51	"ko:K00382 DLD; dihydrolipoamide dehydrogenase [EC:1.8.1.4]"
52	"ko:K00430 E1.11.1.7; peroxidase [EC:1.11.1.7]"
53	"ko:K00475 E1.14.11.9; naringenin 3-dioxygenase [EC:1.14.11.9]"
54	"ko:K00487 CYP73A; trans-cinnamate 4-monooxygenase [EC:1.14.13.11]"
55	"ko:K00511 SQLE; squalene monooxygenase [EC:1.14.13.132]"
56	"ko:K00514 ZDS; zeta-carotene desaturase [EC:1.3.5.6]"
57	"ko:K00517 E1.14.-.-"
58	"ko:K00547 mmuM; homocysteine S-methyltransferase [EC:2.1.1.10]"
59	"ko:K00549 metE; 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase [EC:2.1.1.14]"
60	"ko:K00559 E2.1.1.41; sterol 24-C-methyltransferase [EC:2.1.1.41]"
61	"ko:K00588 E2.1.1.104; caffeoyl-CoA O-methyltransferase [EC:2.1.1.104]"
62	"ko:K00591 COQ3; hexaprenyldihydroxybenzoate methyltransferase [EC:2.1.1.114]"
63	"ko:K00600 glyA; glycine hydroxymethyltransferase [EC:2.1.2.1]"
64	"ko:K00601 E2.1.2.2; phosphoribosylglycinamide formyltransferase [EC:2.1.2.2]"
65	"ko:K00602 purH; phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohydrolase [EC:2.1.2.3 3.5.4.10]"
66	"ko:K00606 panB; 3-methyl-2-oxobutanoate hydroxymethyltransferase [EC:2.1.2.11]"
67	"ko:K00611 OTC; ornithine carbamoyltransferase [EC:2.1.3.3]"

68	"ko:K00615 E2.2.1.1; transketolase [EC:2.2.1.1]"
69	"ko:K00616 E2.2.1.2; transaldolase [EC:2.2.1.2]"
70	"ko:K00620 argJ; glutamate N-acetyltransferase / amino-acid N-acetyltransferase [EC:2.3.1.35 2.3.1.1]"
71	"ko:K00621 GNPAT1; glucosamine-phosphate N-acetyltransferase [EC:2.3.1.4]"
72	"ko:K00626 E2.3.1.9; acetyl-CoA C-acetyltransferase [EC:2.3.1.9]"
73	"ko:K00627 DLAT; pyruvate dehydrogenase E2 component (dihydrolipoamide acetyltransferase) [EC:2.3.1.12]"
74	"ko:K00632 E2.3.1.16; acetyl-CoA acyltransferase [EC:2.3.1.16]"
75	"ko:K00658 DLST; 2-oxoglutarate dehydrogenase E2 component (dihydrolipoamide succinyltransferase) [EC:2.3.1.61]"
76	"ko:K00660 CHS; chalcone synthase [EC:2.3.1.74]"
77	"ko:K00688 E2.4.1.1; starch phosphorylase [EC:2.4.1.1]"
78	"ko:K00700 glgB; 1,4-alpha-glucan branching enzyme [EC:2.4.1.18]"
79	"ko:K00703 E2.4.1.21; starch synthase [EC:2.4.1.21]"
80	"ko:K00760 hprT; hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]"
81	"ko:K00764 purF; amidophosphoribosyltransferase [EC:2.4.2.14]"
82	"ko:K00765 hisG; ATP phosphoribosyltransferase [EC:2.4.2.17]"
83	"ko:K00766 trpD; anthranilate phosphoribosyltransferase [EC:2.4.2.18]"
84	"ko:K00787 FDPS; farnesyl diphosphate synthase [EC:2.5.1.1 2.5.1.10]"
85	"ko:K00789 metK; S-adenosylmethionine synthetase [EC:2.5.1.6]"
86	"ko:K00791 miaA; tRNA dimethylallyltransferase [EC:2.5.1.75]"
87	"ko:K00800 aroA; 3-phosphoshikimate 1-carboxyvinyltransferase [EC:2.5.1.19]"
88	"ko:K00801 FDFT1; farnesyl-diphosphate farnesyltransferase [EC:2.5.1.21]"
89	"ko:K00811 ASP5; aspartate aminotransferase, chloroplastic [EC:2.6.1.1]"
90	"ko:K00815 TAT; tyrosine aminotransferase [EC:2.6.1.5]"
91	"ko:K00817 hisC; histidinol-phosphate aminotransferase [EC:2.6.1.9]"
92	"ko:K00818 E2.6.1.11; acetylornithine aminotransferase [EC:2.6.1.11]"
93	"ko:K00819 E2.6.1.13; ornithine--oxo-acid transaminase [EC:2.6.1.13]"
94	"ko:K00820 glmS; glucosamine--fructose-6-phosphate aminotransferase

	(isomerizing) [EC:2.6.1.16]"
95	"ko:K00826 E2.6.1.42; branched-chain amino acid aminotransferase [EC:2.6.1.42]"
96	"ko:K00830 AGXT; alanine-glyoxylate transaminase / serine-glyoxylate transaminase / serine-pyruvate transaminase [EC:2.6.1.44 2.6.1.45 2.6.1.51]"
97	"ko:K00844 HK; hexokinase [EC:2.7.1.1]"
98	"ko:K00850 pfkA; 6-phosphofructokinase 1 [EC:2.7.1.11]"
99	"ko:K00851 E2.7.1.12; gluconokinase [EC:2.7.1.12]"
100	"ko:K00869 E2.7.1.36; mevalonate kinase [EC:2.7.1.36]"
101	"ko:K00873 PK; pyruvate kinase [EC:2.7.1.40]"
102	"ko:K00891 E2.7.1.71; shikimate kinase [EC:2.7.1.71]"
103	"ko:K00919 ispE; 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase [EC:2.7.1.148]"
104	"ko:K00927 PGK; phosphoglycerate kinase [EC:2.7.2.3]"
105	"ko:K00928 lysC; aspartate kinase [EC:2.7.2.4]"
106	"ko:K00930 argB; acetylglutamate kinase [EC:2.7.2.8]"
107	"ko:K00938 E2.7.4.2; phosphomevalonate kinase [EC:2.7.4.2]"
108	"ko:K00939 adk; adenylate kinase [EC:2.7.4.3]"
109	"ko:K00940 ndk; nucleoside-diphosphate kinase [EC:2.7.4.6]"
110	"ko:K00948 PRPS; ribose-phosphate pyrophosphokinase [EC:2.7.6.1]"
111	"ko:K00963 UGP2; UTP--glucose-1-phosphate uridylyltransferase [EC:2.7.7.9]"
112	"ko:K00965 galT; UDPglucose--hexose-1-phosphate uridylyltransferase [EC:2.7.7.12]"
113	"ko:K00966 GMPP; mannose-1-phosphate guanylyltransferase [EC:2.7.7.13]"
114	"ko:K00972 UAP1; UDP-N-acetylglucosamine/UDP-N-acetylgalactosamine diphosphorylase [EC:2.7.7.23 2.7.7.83]"
115	"ko:K00973 E2.7.7.24; glucose-1-phosphate thymidylyltransferase [EC:2.7.7.24]"
116	"ko:K00975 glgC; glucose-1-phosphate adenidylyltransferase [EC:2.7.7.27]"
117	"ko:K00991 ispD; 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase [EC:2.7.7.60]"
118	"ko:K01057 PGLS; 6-phosphogluconolactonase [EC:3.1.1.31]"

119	"ko:K01061 E3.1.1.45; carboxymethylenebutenolidase [EC:3.1.1.45]"
120	"ko:K01068 ACOT1_2_4; acyl-coenzyme A thioesterase 1/2/4 [EC:3.1.2.2]"
121	"ko:K01081 E3.1.3.5; 5'-nucleotidase [EC:3.1.3.5]"
122	"ko:K01091 E3.1.3.18; phosphoglycolate phosphatase [EC:3.1.3.18]"
123	"ko:K01092 E3.1.3.25; myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]"
124	"ko:K01183 E3.2.1.14; chitinase [EC:3.2.1.14]"
125	"ko:K01188 E3.2.1.21; beta-glucosidase [EC:3.2.1.21]"
126	"ko:K01209 E3.2.1.55; alpha-N-arabinofuranosidase [EC:3.2.1.55]"
127	"ko:K01438 argE; acetylornithine deacetylase [EC:3.5.1.16]"
128	"ko:K01476 E3.5.3.1; arginase [EC:3.5.3.1]"
129	"ko:K01490 AMPD; AMP deaminase [EC:3.5.4.6]"
130	"ko:K01568 E4.1.1.1; pyruvate decarboxylase [EC:4.1.1.1]"
131	"ko:K01580 E4.1.1.15; glutamate decarboxylase [EC:4.1.1.15]"
132	"ko:K01581 E4.1.1.17; ornithine decarboxylase [EC:4.1.1.17]"
133	"ko:K01586 lysA; diaminopimelate decarboxylase [EC:4.1.1.20]"
134	"ko:K01590 hdc; histidine decarboxylase [EC:4.1.1.22]"
135	"ko:K01592 E4.1.1.25; tyrosine decarboxylase [EC:4.1.1.25]"
136	"ko:K01593 DDC; aromatic-L-amino-acid decarboxylase [EC:4.1.1.28]"
137	"ko:K01597 MVD; diphosphomevalonate decarboxylase [EC:4.1.1.33]"
138	"ko:K01599 hemE; uroporphyrinogen decarboxylase [EC:4.1.1.37]"
139	"ko:K01609 trpC; indole-3-glycerol phosphate synthase [EC:4.1.1.48]"
140	"ko:K01610 E4.1.1.49; phosphoenolpyruvate carboxykinase (ATP) [EC:4.1.1.49]"
141	"ko:K01620 ltaE; threonine aldolase [EC:4.1.2.5]"
142	"ko:K01623 ALDO; fructose-bisphosphate aldolase, class I [EC:4.1.2.13]"
143	"ko:K01624 FBA; fructose-bisphosphate aldolase, class II [EC:4.1.2.13]"
144	"ko:K01626 E2.5.1.54; 3-deoxy-7-phosphoheptulonate synthase [EC:2.5.1.54]"
145	"ko:K01640 E4.1.3.4; hydroxymethylglutaryl-CoA lyase [EC:4.1.3.4]"
146	"ko:K01641 E2.3.3.10; hydroxymethylglutaryl-CoA synthase [EC:2.3.3.10]"
147	"ko:K01647 CS; citrate synthase [EC:2.3.3.1]"

148	"ko:K01648 ACLY; ATP citrate (pro-S)-lyase [EC:2.3.3.8]"
149	"ko:K01649 leuA; 2-isopropylmalate synthase [EC:2.3.3.13]"
150	"ko:K01652 E2.2.1.6L; acetolactate synthase I/II/III large subunit [EC:2.2.1.6]"
151	"ko:K01653 E2.2.1.6S; acetolactate synthase I/III small subunit [EC:2.2.1.6]"
152	"ko:K01657 trpE; anthranilate synthase component I [EC:4.1.3.27]"
153	"ko:K01658 trpG; anthranilate synthase component II [EC:4.1.3.27]"
154	"ko:K01661 menB; naphthoate synthase [EC:4.1.3.36]"
155	"ko:K01662 dxs; 1-deoxy-D-xylulose-5-phosphate synthase [EC:2.2.1.7]"
156	"ko:K01663 HIS7; glutamine amidotransferase / cyclase [EC:2.4.2.- 4.1.3.-]"
157	"ko:K01679 E4.2.1.2B; fumarate hydratase, class II [EC:4.2.1.2]"
158	"ko:K01681 ACO; aconitate hydratase [EC:4.2.1.3]"
159	"ko:K01687 ilvD; dihydroxy-acid dehydratase [EC:4.2.1.9]"
160	"ko:K01689 ENO; enolase [EC:4.2.1.11]"
161	"ko:K01693 hisB; imidazoleglycerol-phosphate dehydratase [EC:4.2.1.19]"
162	"ko:K01695 trpA; tryptophan synthase alpha chain [EC:4.2.1.20]"
163	"ko:K01696 trpB; tryptophan synthase beta chain [EC:4.2.1.20]"
164	"ko:K01698 hemB; porphobilinogen synthase [EC:4.2.1.24]"
165	"ko:K01703 leuC; 3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit [EC:4.2.1.33 4.2.1.35]"
166	"ko:K01704 leuD; 3-isopropylmalate/(R)-2-methylmalate dehydratase small subunit [EC:4.2.1.33 4.2.1.35]"
167	"ko:K01714 dapA; 4-hydroxy-tetrahydrodipicolinate synthase [EC:4.3.3.7]"
168	"ko:K01735 aroB; 3-dehydroquinate synthase [EC:4.2.3.4]"
169	"ko:K01736 aroC; chorismate synthase [EC:4.2.3.5]"
170	"ko:K01739 metB; cystathionine gamma-synthase [EC:2.5.1.48]"
171	"ko:K01749 hemC; hydroxymethylbilane synthase [EC:2.5.1.61]"
172	"ko:K01754 E4.3.1.19; threonine dehydratase [EC:4.3.1.19]"
173	"ko:K01755 argH; argininosuccinate lyase [EC:4.3.2.1]"
174	"ko:K01756 purB; adenylosuccinate lyase [EC:4.3.2.2]"
175	"ko:K01760 metC; cystathionine beta-lyase [EC:4.4.1.8]"

176	"ko:K01762 ACS; 1-aminocyclopropane-1-carboxylate synthase [EC:4.4.1.14]"
177	"ko:K01770 ispF; 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase [EC:4.6.1.12]"
178	"ko:K01772 hemH; ferrochelatase [EC:4.99.1.1]"
179	"ko:K01778 dapF; diaminopimelate epimerase [EC:5.1.1.7]"
180	"ko:K01783 rpe; ribulose-phosphate 3-epimerase [EC:5.1.3.1]"
181	"ko:K01784 galE; UDP-glucose 4-epimerase [EC:5.1.3.2]"
182	"ko:K01785 galM; aldose 1-epimerase [EC:5.1.3.3]"
183	"ko:K01792 E5.1.3.15; glucose-6-phosphate 1-epimerase [EC:5.1.3.15]"
184	"ko:K01803 TPI; triosephosphate isomerase (TIM) [EC:5.3.1.1]"
185	"ko:K01807 rpiA; ribose 5-phosphate isomerase A [EC:5.3.1.6]"
186	"ko:K01809 manA; mannose-6-phosphate isomerase [EC:5.3.1.8]"
187	"ko:K01810 GPI; glucose-6-phosphate isomerase [EC:5.3.1.9]"
188	"ko:K01814 hisA; phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase [EC:5.3.1.16]"
189	"ko:K01817 trpF; phosphoribosylanthranilate isomerase [EC:5.3.1.24]"
190	"ko:K01823 idi; isopentenyl-diphosphate delta-isomerase [EC:5.3.3.2]"
191	"ko:K01824 EBP; cholesterol delta-isomerase [EC:5.3.3.5]"
192	"ko:K01834 PGAM; 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase [EC:5.4.2.11]"
193	"ko:K01835 pgm; phosphoglucomutase [EC:5.4.2.2]"
194	"ko:K01836 E5.4.2.3; phosphoacetylglucosamine mutase [EC:5.4.2.3]"
195	"ko:K01841 pepM; phosphoenolpyruvate phosphomutase [EC:5.4.2.9]"
196	"ko:K01845 hemL; glutamate-1-semialdehyde 2,1-aminomutase [EC:5.4.3.8]"
197	"ko:K01850 E5.4.99.5; chorismate mutase [EC:5.4.99.5]"
198	"ko:K01853 E5.4.99.8; cycloartenol synthase [EC:5.4.99.8]"
199	"ko:K01858 E5.5.1.4; myo-inositol-1-phosphate synthase [EC:5.5.1.4]"
200	"ko:K01859 E5.5.1.6; chalcone isomerase [EC:5.5.1.6]"
201	"ko:K01885 EARS; glutamyl-tRNA synthetase [EC:6.1.1.17]"
202	"ko:K01895 ACSS; acetyl-CoA synthetase [EC:6.2.1.1]"

203	"ko:K01899 LSC1; succinyl-CoA synthetase alpha subunit [EC:6.2.1.4 6.2.1.5]"
204	"ko:K01900 LSC2; succinyl-CoA synthetase beta subunit [EC:6.2.1.4 6.2.1.5]"
205	"ko:K01903 sucC; succinyl-CoA synthetase beta subunit [EC:6.2.1.5]"
206	"ko:K01904 4CL; 4-coumarate--CoA ligase [EC:6.2.1.12]"
207	"ko:K01918 panC; pantoate--beta-alanine ligase [EC:6.3.2.1]"
208	"ko:K01923 purC; phosphoribosylaminoimidazole-succinocarboxamide synthase [EC:6.3.2.6]"
209	"ko:K01933 purM; phosphoribosylformylglycinamide cyclo-ligase [EC:6.3.3.1]"
210	"ko:K01940 argG; argininosuccinate synthase [EC:6.3.4.5]"
211	"ko:K01945 purD; phosphoribosylamine--glycine ligase [EC:6.3.4.13]"
212	"ko:K01952 purL; phosphoribosylformylglycinamide synthase [EC:6.3.5.3]"
213	"ko:K01953 asnB; asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4]"
214	"ko:K01961 accC; acetyl-CoA carboxylase, biotin carboxylase subunit [EC:6.4.1.2 6.3.4.14]"
215	"ko:K01962 accA; acetyl-CoA carboxylase carboxyl transferase subunit alpha [EC:6.4.1.2]"
216	"ko:K01963 accD; acetyl-CoA carboxylase carboxyl transferase subunit beta [EC:6.4.1.2]"
217	"ko:K02160 accB; acetyl-CoA carboxylase biotin carboxyl carrier protein"
218	"ko:K02259 COX15; cytochrome c oxidase assembly protein subunit 15"
219	"ko:K02291 crtB; phytoene synthase [EC:2.5.1.32]"
220	"ko:K02293 PDS; 15-cis-phytoene desaturase [EC:1.3.5.5]"
221	"ko:K02437 gcvH; glycine cleavage system H protein"
222	"ko:K02492 hemA; glutamyl-tRNA reductase [EC:1.2.1.70]"
223	"ko:K02552 menF; menaquinone-specific isochorismate synthase [EC:5.4.4.2]"
224	"ko:K03183 ubiE; ubiquinone/menaquinone biosynthesis methyltransferase [EC:2.1.1.163 2.1.1.201]"
225	"ko:K03403 chlH; magnesium chelatase subunit H [EC:6.6.1.1]"
226	"ko:K03404 chlD; magnesium chelatase subunit D [EC:6.6.1.1]"
227	"ko:K03405 chlI; magnesium chelatase subunit I [EC:6.6.1.1]"
228	"ko:K03428 E2.1.1.11; magnesium-protoporphyrin O-methyltransferase"

	[EC:2.1.1.11]"
229	"ko:K03526 E1.17.7.1; (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase [EC:1.17.7.1]"
230	"ko:K03527 ispH; 4-hydroxy-3-methylbut-2-enyl diphosphate reductase [EC:1.17.1.2]"
231	"ko:K03781 katE; catalase [EC:1.11.1.6]"
232	"ko:K03787 surE; 5'-nucleotidase [EC:3.1.3.5]"
233	"ko:K03841 FBP; fructose-1,6-bisphosphatase I [EC:3.1.3.11]"
234	"ko:K04035 E1.14.13.81; magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase [EC:1.14.13.81]"
235	"ko:K04040 chlG; chlorophyll synthase [EC:2.5.1.62]"
236	"ko:K04120 E5.5.1.13; ent-copalyl diphosphate synthase [EC:5.5.1.13]"
237	"ko:K04121 E4.2.3.19; ent-kaurene synthase [EC:4.2.3.19]"
238	"ko:K04122 GA3; ent-kaurene oxidase [EC:1.14.13.78]"
239	"ko:K04123 KAO; ent-kaurenoic acid hydroxylase [EC:1.14.13.79]"
240	"ko:K04124 E1.14.11.15; gibberellin 3-beta-dioxygenase [EC:1.14.11.15]"
241	"ko:K04518 pheA2; prephenate dehydratase [EC:4.2.1.51]"
242	"ko:K05278 FLS; flavonol synthase [EC:1.14.11.23]"
243	"ko:K05280 E1.14.13.21; flavonoid 3'-monooxygenase [EC:1.14.13.21]"
244	"ko:K05349 bglX; beta-glucosidase [EC:3.2.1.21]"
245	"ko:K05350 bglB; beta-glucosidase [EC:3.2.1.21]"
246	"ko:K05359 ADT; arogenate/prephenate dehydratase [EC:4.2.1.91 4.2.1.51]"
247	"ko:K05917 CYP51; sterol 14-demethylase [EC:1.14.13.70]"
248	"ko:K05928 E2.1.1.95; tocopherol O-methyltransferase [EC:2.1.1.95]"
249	"ko:K05933 E1.14.17.4; aminocyclopropanecarboxylate oxidase [EC:1.14.17.4]"
250	"ko:K06001 trpB; tryptophan synthase beta chain [EC:4.2.1.20]"
251	"ko:K06125 COQ2; 4-hydroxybenzoate hexaprenyltransferase [EC:2.5.1.-]"
252	"ko:K06126 COQ6; ubiquinone biosynthesis monooxygenase Coq6 [EC:1.14.13.-]"
253	"ko:K06127 COQ5; ubiquinone biosynthesis methyltransferase [EC:2.1.1.201]"
254	"ko:K06443 lcyB; lycopene beta-cyclase [EC:5.5.1.19]"

255	"ko:K07385 TPS-Cin; 1,8-cineole synthase [EC:4.2.3.108]"
256	"ko:K07513 ACAA1; acetyl-CoA acyltransferase 1 [EC:2.3.1.16]"
257	"ko:K08081 TR1; Tropinone reductase 1 [EC:1.1.1.206]"
258	"ko:K08099 E3.1.1.14; chlorophyllase [EC:3.1.1.14]"
259	"ko:K08242 E2.1.1.143; 24-methylenesterol C-methyltransferase [EC:2.1.1.143]"
260	"ko:K08246 CPI1; cycloeucaleanol cycloisomerase [EC:5.5.1.9]"
261	"ko:K08683 HSD17B10; 3-hydroxyacyl-CoA dehydrogenase / 3-hydroxy-2-methylbutyryl-CoA dehydrogenase [EC:1.1.1.35 1.1.1.178]"
262	"ko:K09587 CYP90B1; steroid 22-alpha-hydroxylase [EC:1.14.13.-]"
263	"ko:K09588 CYP90A1; cytochrome P450, family 90, subfamily A, polypeptide 1 [EC:1.14.-.-]"
264	"ko:K09591 DET2; steroid 5-alpha-reductase [EC:1.3.1.22]"
265	"ko:K09699 DBT; 2-oxoisovalerate dehydrogenase E2 component (dihydrolipoyl transacylase) [EC:2.3.1.168]"
266	"ko:K09753 CCR; cinnamoyl-CoA reductase [EC:1.2.1.44]"
267	"ko:K09754 CYP98A3; coumaroylquininate(coumaroylshikimate) 3'-monooxygenase [EC:1.14.13.36]"
268	"ko:K09755 CYP84A; ferulate-5-hydroxylase [EC:1.14.-.-]"
269	"ko:K09828 DHCR24; delta24-sterol reductase [EC:1.3.1.72]"
270	"ko:K09832 CYP710A; cytochrome P450, family 710, subfamily A"
271	"ko:K09833 HPT; homogentisate phytyltransferase / homogentisate geranylgeranyltransferase [EC:2.5.1.115 2.5.1.116]"
272	"ko:K09834 VTE1; tocopherol cyclase [EC:5.5.1.24]"
273	"ko:K09835 crtISO; prolycopene isomerase [EC:5.2.1.13]"
274	"ko:K09837 LUT1; carotene epsilon-monooxygenase [EC:1.14.99.45]"
275	"ko:K09838 ZEP; zeaxanthin epoxidase [EC:1.14.13.90]"
276	"ko:K09839 VDE; violaxanthin de-epoxidase [EC:1.10.99.3]"
277	"ko:K09840 NCED; 9-cis-epoxycarotenoid dioxygenase [EC:1.13.11.51]"
278	"ko:K09841 ABA2; xanthoxin dehydrogenase [EC:1.1.1.288]"
279	"ko:K09842 AAO3; abscisic-aldehyde oxidase [EC:1.2.3.14]"
280	"ko:K10046 GME; GDP-D-mannose 3', 5'-epimerase [EC:5.1.3.18 5.1.3.-]"

281	"ko:K10047 VTC4; inositol-phosphate phosphatase / L-galactose 1-phosphate phosphatase [EC:3.1.3.25 3.1.3.93]"
282	"ko:K10206 E2.6.1.83; LL-diaminopimelate aminotransferase [EC:2.6.1.83]"
283	"ko:K10251 HSD17B12; 17beta-estradiol 17-dehydrogenase / very-long-chain 3-oxoacyl-CoA reductase [EC:1.1.1.62 1.1.1.330]"
284	"ko:K10258 TER; very-long-chain enoyl-CoA reductase [EC:1.3.1.93]"
285	"ko:K10703 PHS1; very-long-chain (3R)-3-hydroxyacyl-CoA dehydratase [EC:4.2.1.134]"
286	"ko:K10720 PHM; CYP306A1; ecdysteroid 25-hydroxylase"
287	"ko:K10757 E2.4.1.91; flavonol 3-O-glucosyltransferase [EC:2.4.1.91]"
288	"ko:K10760 IPT; adenylate isopentenyltransferase (cytokinin synthase)"
289	"ko:K10775 E4.3.1.24; phenylalanine ammonia-lyase [EC:4.3.1.24]"
290	"ko:K10960 ch1P; geranylgeranyl reductase [EC:1.3.1.83]"
291	"ko:K11188 PRDX6; peroxiredoxin 6, 1-Cys peroxiredoxin [EC:1.11.1.7 1.11.1.15 3.1.1.-]"
292	"ko:K11517 HAO; (S)-2-hydroxy-acid oxidase [EC:1.1.3.15]"
293	"ko:K11755 hisIE; phosphoribosyl-ATP pyrophosphohydrolase / phosphoribosyl-AMP cyclohydrolase [EC:3.6.1.31 3.5.4.19]"
294	"ko:K11778 DHDDS; ditrans, polycis-polyprenyl diphosphate synthase [EC:2.5.1.87]"
295	"ko:K11808 ADE2; phosphoribosylaminoimidazole carboxylase [EC:4.1.1.21]"
296	"ko:K12373 HEXA_B; hexosaminidase [EC:3.2.1.52]"
297	"ko:K12446 E2.7.1.46; L-arabinokinase [EC:2.7.1.46]"
298	"ko:K12447 USP; UDP-sugar pyrophosphorylase [EC:2.7.7.64]"
299	"ko:K12448 UXE; UDP-arabinose 4-epimerase [EC:5.1.3.5]"
300	"ko:K12449 AXS; UDP-apiose/xylose synthase"
301	"ko:K12450 RHM; UDP-glucose 4,6-dehydratase [EC:4.2.1.76]"
302	"ko:K12451 UER1; 3,5-epimerase/4-reductase [EC:5.1.3.- 1.1.1.-]"
303	"ko:K12502 VTE3; MPBQ/MSBQ methyltransferase [EC:2.1.1.295]"
304	"ko:K12504 PDSS1; decaprenyl-diphosphate synthase subunit 1 [EC:2.5.1.91]"
305	"ko:K12524 thrA; bifunctional aspartokinase / homoserine dehydrogenase 1 [EC:2.7.2.4 1.1.1.3]"

306	"ko:K12637 CYP90C1; 3-epi-6-deoxocathasterone 23-monooxygenase [EC:1.14.13.112]"
307	"ko:K13065 E2.3.1.133; shikimate O-hydroxycinnamoyltransferase [EC:2.3.1.133]"
308	"ko:K13066 E2.1.1.68; caffeic acid 3-O-methyltransferase [EC:2.1.1.68]"
309	"ko:K13071 PAO; pheophorbide a oxygenase [EC:1.14.12.20]"
310	"ko:K13427 NOA1; nitric-oxide synthase, plant [EC:1.14.13.39]"
311	"ko:K13545 ACD2; red chlorophyll catabolite reductase [EC:1.3.1.80]"
312	"ko:K13600 CAO; chlorophyllide a oxygenase [EC:1.14.13.122]"
313	"ko:K13606 NOL; chlorophyll(ide) b reductase [EC:1.1.1.294]"
314	"ko:K13648 GAUT; alpha-1,4-galacturonosyltransferase [EC:2.4.1.43]"
315	"ko:K13789 GGPS; geranylgeranyl diphosphate synthase, type II [EC:2.5.1.1 2.5.1.10 2.5.1.29]"
316	"ko:K13832 aroDE; 3-dehydroquininate dehydratase / shikimate dehydrogenase [EC:4.2.1.10 1.1.1.25]"
317	"ko:K14066 GPS; geranyl diphosphate synthase [EC:2.5.1.1]"
318	"ko:K14085 ALDH7A1; aldehyde dehydrogenase family 7 member A1 [EC:1.2.1.31 1.2.1.8 1.2.1.3]"
319	"ko:K14190 VTC2_5; GDP-L-galactose phosphorylase [EC:2.7.7.69]"
320	"ko:K14272 GGAT; glutamate--glyoxylate aminotransferase [EC:2.6.1.4 2.6.1.2 2.6.1.44 2.6.1.-]"
321	"ko:K14454 GOT1; aspartate aminotransferase, cytoplasmic [EC:2.6.1.1]"
322	"ko:K14455 GOT2; aspartate aminotransferase, mitochondrial [EC:2.6.1.1]"
323	"ko:K14677 ACY1; aminoacylase [EC:3.5.1.14]"
324	"ko:K14682 argAB; amino-acid N-acetyltransferase [EC:2.3.1.1]"
325	"ko:K14759 PHYLL0; isochorismate synthase / 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase / 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase / O-succinylbenzoate synthase [EC:5.4.4.2 2.2.1.9 4.2.99.20 4.2.1.113]"
326	"ko:K14985 CYP18A1; 26-hydroxylase [EC:1.14.-.-]"
327	"ko:K15095 E1.1.1.208; (+)-neomenthol dehydrogenase [EC:1.1.1.208]"
328	"ko:K15397 KCS; 3-ketoacyl-CoA synthase [EC:2.3.1.199]"
329	"ko:K15404 CER1; aldehyde decarbonylase [EC:4.1.99.5]"

330	"ko:K15633 gpmI; 2,3-bisphosphoglycerate-independent phosphoglycerate mutase [EC:5.4.2.12]"
331	"ko:K15634 gpmB; probable phosphoglycerate mutase [EC:5.4.2.12]"
332	"ko:K15746 crtZ; beta-carotene 3-hydroxylase [EC:1.14.13.129]"
333	"ko:K15747 LUT5; beta-ring hydroxylase [EC:1.14.-.-]"
334	"ko:K15891 FLDH; farnesol dehydrogenase [EC:1.1.1.216]"
335	"ko:K15918 GLYK; D-glycerate 3-kinase [EC:2.7.1.31]"
336	"ko:K15919 HPR2; hydroxypyruvate reductase 2"
337	"ko:K17497 PMM; phosphomannomutase [EC:5.4.2.8]"
338	"ko:K17744 GalDH; L-galactose dehydrogenase [EC:1.1.1.316]"
339	"ko:K17989 SDS; L-serine/L-threonine ammonia-lyase [EC:4.3.1.17 4.3.1.19]"
340	"ko:K18368 CSE; caffeoylshikimate esterase [EC:3.1.1.-]"
341	"ko:K18649 IMPL2; inositol-phosphate phosphatase / L-galactose 1-phosphate phosphatase / histidinol-phosphatase [EC:3.1.3.25 3.1.3.93 3.1.3.15]"

Table A3 KOs associated with secondary metabolism in AHSS

Sr. no.	KO associated enzymes
1	"ko:K00001 E1.1.1.1; alcohol dehydrogenase [EC:1.1.1.1]"
2	"ko:K00003 E1.1.1.3; homoserine dehydrogenase [EC:1.1.1.3]"
3	"ko:K00012 UGDH; UDPglucose 6-dehydrogenase [EC:1.1.1.22]"
4	"ko:K00013 hisD; histidinol dehydrogenase [EC:1.1.1.23]"
5	"ko:K00016 LDH; L-lactate dehydrogenase [EC:1.1.1.27]"
6	"ko:K00021 HMGCR; hydroxymethylglutaryl-CoA reductase (NADPH) [EC:1.1.1.34]"
7	"ko:K00025 MDH1; malate dehydrogenase [EC:1.1.1.37]"
8	"ko:K00026 MDH2; malate dehydrogenase [EC:1.1.1.37]"
9	"ko:K00030 IDH3; isocitrate dehydrogenase (NAD+) [EC:1.1.1.41]"
10	"ko:K00031 IDH1; isocitrate dehydrogenase [EC:1.1.1.42]"
11	"ko:K00033 PGD; 6-phosphogluconate dehydrogenase [EC:1.1.1.44]"
12	"ko:K00036 G6PD; glucose-6-phosphate 1-dehydrogenase [EC:1.1.1.49]"

13	"ko:K00052 leuB; 3-isopropylmalate dehydrogenase [EC:1.1.1.85]"
14	"ko:K00053 ilvC; ketol-acid reductoisomerase [EC:1.1.1.86]"
15	"ko:K00083 E1.1.1.195; cinnamyl-alcohol dehydrogenase [EC:1.1.1.195]"
16	"ko:K00088 guaB; IMP dehydrogenase [EC:1.1.1.205]"
17	"ko:K00099 dxr; 1-deoxy-D-xylulose-5-phosphate reductoisomerase [EC:1.1.1.267]"
18	"ko:K00106 XDH; xanthine dehydrogenase/oxidase [EC:1.17.1.4 1.17.3.2]"
19	"ko:K00121 frmA; S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]"
20	"ko:K00128 E1.2.1.3; aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]"
21	"ko:K00133 asd; aspartate-semialdehyde dehydrogenase [EC:1.2.1.11]"
22	"ko:K00134 GAPDH; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12]"
23	"ko:K00145 argC; N-acetyl-gamma-glutamyl-phosphate reductase [EC:1.2.1.38]"
24	"ko:K00161 PDHA; pyruvate dehydrogenase E1 component alpha subunit [EC:1.2.4.1]"
25	"ko:K00162 PDHB; pyruvate dehydrogenase E1 component beta subunit [EC:1.2.4.1]"
26	"ko:K00164 OGDH; 2-oxoglutarate dehydrogenase E1 component [EC:1.2.4.2]"
27	"ko:K00166 BCKDHA; 2-oxoisovalerate dehydrogenase E1 component alpha subunit [EC:1.2.4.4]"
28	"ko:K00167 BCKDHB; 2-oxoisovalerate dehydrogenase E1 component beta subunit [EC:1.2.4.4]"
29	"ko:K00213 DHCR7; 7-dehydrocholesterol reductase [EC:1.3.1.21]"
30	"ko:K00215 dapB; 4-hydroxy-tetrahydrodipicolinate reductase [EC:1.17.1.8]"
31	"ko:K00218 E1.3.1.33; protochlorophyllide reductase [EC:1.3.1.33]"
32	"ko:K00222 TM7SF2; delta14-sterol reductase [EC:1.3.1.70]"
33	"ko:K00225 GLDH; L-galactono-1,4-lactone dehydrogenase [EC:1.3.2.3]"
34	"ko:K00227 SC5DL; lathosterol oxidase [EC:1.14.21.6]"
35	"ko:K00228 CPOX; coproporphyrinogen III oxidase [EC:1.3.3.3]"
36	"ko:K00231 PPOX; oxygen-dependent protoporphyrinogen oxidase [EC:1.3.3.4]"
37	"ko:K00234 SDHA; succinate dehydrogenase (ubiquinone) flavoprotein subunit [EC:1.3.5.1]"

38	"ko:K00235 SDHB; succinate dehydrogenase (ubiquinone) iron-sulfur subunit [EC:1.3.5.1]"
39	"ko:K00249 ACADM; acyl-CoA dehydrogenase [EC:1.3.8.7]"
40	"ko:K00257 E1.3.99.-"
41	"ko:K00264 GLT1; glutamate synthase (NADPH/NADH) [EC:1.4.1.13 1.4.1.14]"
42	"ko:K00276 AOC3; primary-amine oxidase [EC:1.4.3.21]"
43	"ko:K00286 E1.5.1.2; pyrroline-5-carboxylate reductase [EC:1.5.1.2]"
44	"ko:K00318 PRODH; proline dehydrogenase [EC:1.5.-.-]"
45	"ko:K00326 E1.6.2.2; cytochrome-b5 reductase [EC:1.6.2.2]"
46	"ko:K00382 DLD; dihydrolipoamide dehydrogenase [EC:1.8.1.4]"
47	"ko:K00430 E1.11.1.7; peroxidase [EC:1.11.1.7]"
48	"ko:K00475 E1.14.11.9; naringenin 3-dioxygenase [EC:1.14.11.9]"
49	"ko:K00487 CYP73A; trans-cinnamate 4-monooxygenase [EC:1.14.13.11]"
50	"ko:K00511 SQLE; squalene monooxygenase [EC:1.14.13.132]"
51	"ko:K00514 ZDS; zeta-carotene desaturase [EC:1.3.5.6]"
52	"ko:K00517 E1.14.-.-"
53	"ko:K00547 mmuM; homocysteine S-methyltransferase [EC:2.1.1.10]"
54	"ko:K00549 metE; 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase [EC:2.1.1.14]"
55	"ko:K00559 E2.1.1.41; sterol 24-C-methyltransferase [EC:2.1.1.41]"
56	"ko:K00588 E2.1.1.104; caffeoyl-CoA O-methyltransferase [EC:2.1.1.104]"
57	"ko:K00600 glyA; glycine hydroxymethyltransferase [EC:2.1.2.1]"
58	"ko:K00601 E2.1.2.2; phosphoribosylglycinamide formyltransferase [EC:2.1.2.2]"
59	"ko:K00602 purH; phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohydrolase [EC:2.1.2.3 3.5.4.10]"
60	"ko:K00606 panB; 3-methyl-2-oxobutanoate hydroxymethyltransferase [EC:2.1.2.11]"
61	"ko:K00611 OTC; ornithine carbamoyltransferase [EC:2.1.3.3]"
62	"ko:K00615 E2.2.1.1; transketolase [EC:2.2.1.1]"
63	"ko:K00616 E2.2.1.2; transaldolase [EC:2.2.1.2]"

64	"ko:K00620 argJ; glutamate N-acetyltransferase / amino-acid N-acetyltransferase [EC:2.3.1.35 2.3.1.1]"
65	"ko:K00621 GNPAT1; glucosamine-phosphate N-acetyltransferase [EC:2.3.1.4]"
66	"ko:K00626 E2.3.1.9; acetyl-CoA C-acetyltransferase [EC:2.3.1.9]"
67	"ko:K00627 DLAT; pyruvate dehydrogenase E2 component (dihydrolipoamide acetyltransferase) [EC:2.3.1.12]"
68	"ko:K00658 DLST; 2-oxoglutarate dehydrogenase E2 component (dihydrolipoamide succinyltransferase) [EC:2.3.1.61]"
69	"ko:K00660 CHS; chalcone synthase [EC:2.3.1.74]"
70	"ko:K00688 E2.4.1.1; starch phosphorylase [EC:2.4.1.1]"
71	"ko:K00700 glgB; 1,4-alpha-glucan branching enzyme [EC:2.4.1.18]"
72	"ko:K00703 E2.4.1.21; starch synthase [EC:2.4.1.21]"
73	"ko:K00760 hprT; hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]"
74	"ko:K00764 purF; amidophosphoribosyltransferase [EC:2.4.2.14]"
75	"ko:K00765 hisG; ATP phosphoribosyltransferase [EC:2.4.2.17]"
76	"ko:K00766 trpD; anthranilate phosphoribosyltransferase [EC:2.4.2.18]"
77	"ko:K00787 FDPS; farnesyl diphosphate synthase [EC:2.5.1.1 2.5.1.10]"
78	"ko:K00789 metK; S-adenosylmethionine synthetase [EC:2.5.1.6]"
79	"ko:K00791 miaA; tRNA dimethylallyltransferase [EC:2.5.1.75]"
80	"ko:K00800 aroA; 3-phosphoshikimate 1-carboxyvinyltransferase [EC:2.5.1.19]"
81	"ko:K00801 FDFT1; farnesyl-diphosphate farnesyltransferase [EC:2.5.1.21]"
82	"ko:K00806 uppS; undecaprenyl diphosphate synthase [EC:2.5.1.31]"
83	"ko:K00811 ASP5; aspartate aminotransferase, chloroplastic [EC:2.6.1.1]"
84	"ko:K00815 TAT; tyrosine aminotransferase [EC:2.6.1.5]"
85	"ko:K00817 hisC; histidinol-phosphate aminotransferase [EC:2.6.1.9]"
86	"ko:K00818 E2.6.1.11; acetylornithine aminotransferase [EC:2.6.1.11]"
87	"ko:K00819 E2.6.1.13; ornithine--oxo-acid transaminase [EC:2.6.1.13]"
88	"ko:K00820 glmS; glucosamine--fructose-6-phosphate aminotransferase (isomerizing) [EC:2.6.1.16]"
89	"ko:K00826 E2.6.1.42; branched-chain amino acid aminotransferase

	[EC:2.6.1.42]"
90	"ko:K00830 AGXT; alanine-glyoxylate transaminase / serine-glyoxylate transaminase / serine-pyruvate transaminase [EC:2.6.1.44 2.6.1.45 2.6.1.51]"
91	"ko:K00844 HK; hexokinase [EC:2.7.1.1]"
92	"ko:K00850 pfkA; 6-phosphofructokinase 1 [EC:2.7.1.11]"
93	"ko:K00851 E2.7.1.12; gluconokinase [EC:2.7.1.12]"
94	"ko:K00869 E2.7.1.36; mevalonate kinase [EC:2.7.1.36]"
95	"ko:K00873 PK; pyruvate kinase [EC:2.7.1.40]"
96	"ko:K00891 E2.7.1.71; shikimate kinase [EC:2.7.1.71]"
97	"ko:K00919 ispE; 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase [EC:2.7.1.148]"
98	"ko:K00927 PGK; phosphoglycerate kinase [EC:2.7.2.3]"
99	"ko:K00928 lysC; aspartate kinase [EC:2.7.2.4]"
100	"ko:K00930 argB; acetylglutamate kinase [EC:2.7.2.8]"
101	"ko:K00938 E2.7.4.2; phosphomevalonate kinase [EC:2.7.4.2]"
102	"ko:K00939 adk; adenylylate kinase [EC:2.7.4.3]"
103	"ko:K00940 ndk; nucleoside-diphosphate kinase [EC:2.7.4.6]"
104	"ko:K00948 PRPS; ribose-phosphate pyrophosphokinase [EC:2.7.6.1]"
105	"ko:K00963 UGP2; UTP--glucose-1-phosphate uridylyltransferase [EC:2.7.7.9]"
106	"ko:K00965 galT; UDPglucose--hexose-1-phosphate uridylyltransferase [EC:2.7.7.12]"
107	"ko:K00966 GMPP; mannose-1-phosphate guanylyltransferase [EC:2.7.7.13]"
108	"ko:K00972 UAP1; UDP-N-acetylglucosamine/UDP-N-acetylgalactosamine diphosphorylase [EC:2.7.7.23 2.7.7.83]"
109	"ko:K00975 glgC; glucose-1-phosphate adenylyltransferase [EC:2.7.7.27]"
110	"ko:K00991 ispD; 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase [EC:2.7.7.60]"
111	"ko:K01057 PGLS; 6-phosphogluconolactonase [EC:3.1.1.31]"
112	"ko:K01061 E3.1.1.45; carboxymethylenebutenolidase [EC:3.1.1.45]"
113	"ko:K01068 ACOT1_2_4; acyl-coenzyme A thioesterase 1/2/4 [EC:3.1.2.2]"
114	"ko:K01081 E3.1.3.5; 5'-nucleotidase [EC:3.1.3.5]"

115	"ko:K01091 E3.1.3.18; phosphoglycolate phosphatase [EC:3.1.3.18]"
116	"ko:K01092 E3.1.3.25; myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]"
117	"ko:K01183 E3.2.1.14; chitinase [EC:3.2.1.14]"
118	"ko:K01188 E3.2.1.21; beta-glucosidase [EC:3.2.1.21]"
119	"ko:K01209 E3.2.1.55; alpha-N-arabinofuranosidase [EC:3.2.1.55]"
120	"ko:K01438 argE; acetylornithine deacetylase [EC:3.5.1.16]"
121	"ko:K01476 E3.5.3.1; arginase [EC:3.5.3.1]"
122	"ko:K01490 AMPD; AMP deaminase [EC:3.5.4.6]"
123	"ko:K01568 E4.1.1.1; pyruvate decarboxylase [EC:4.1.1.1]"
124	"ko:K01580 E4.1.1.15; glutamate decarboxylase [EC:4.1.1.15]"
125	"ko:K01582 E4.1.1.18; lysine decarboxylase [EC:4.1.1.18]"
126	"ko:K01586 lysA; diaminopimelate decarboxylase [EC:4.1.1.20]"
127	"ko:K01588 purE; 5-(carboxyamino)imidazole ribonucleotide mutase [EC:5.4.99.18]"
128	"ko:K01590 hdc; histidine decarboxylase [EC:4.1.1.22]"
129	"ko:K01592 E4.1.1.25; tyrosine decarboxylase [EC:4.1.1.25]"
130	"ko:K01597 MVD; diphosphomevalonate decarboxylase [EC:4.1.1.33]"
131	"ko:K01599 hemE; uroporphyrinogen decarboxylase [EC:4.1.1.37]"
132	"ko:K01609 trpC; indole-3-glycerol phosphate synthase [EC:4.1.1.48]"
133	"ko:K01610 E4.1.1.49; phosphoenolpyruvate carboxykinase (ATP) [EC:4.1.1.49]"
134	"ko:K01620 ltaE; threonine aldolase [EC:4.1.2.5]"
135	"ko:K01623 ALDO; fructose-bisphosphate aldolase, class I [EC:4.1.2.13]"
136	"ko:K01626 E2.5.1.54; 3-deoxy-7-phosphoheptulonate synthase [EC:2.5.1.54]"
137	"ko:K01640 E4.1.3.4; hydroxymethylglutaryl-CoA lyase [EC:4.1.3.4]"
138	"ko:K01641 E2.3.3.10; hydroxymethylglutaryl-CoA synthase [EC:2.3.3.10]"
139	"ko:K01647 CS; citrate synthase [EC:2.3.3.1]"
140	"ko:K01648 ACLY; ATP citrate (pro-S)-lyase [EC:2.3.3.8]"
141	"ko:K01649 leuA; 2-isopropylmalate synthase [EC:2.3.3.13]"
142	"ko:K01652 E2.2.1.6L; acetolactate synthase I/II/III large subunit [EC:2.2.1.6]"

143	"ko:K01653 E2.2.1.6S; acetolactate synthase I/III small subunit [EC:2.2.1.6]"
144	"ko:K01657 trpE; anthranilate synthase component I [EC:4.1.3.27]"
145	"ko:K01658 trpG; anthranilate synthase component II [EC:4.1.3.27]"
146	"ko:K01661 menB; naphthoate synthase [EC:4.1.3.36]"
147	"ko:K01662 dxs; 1-deoxy-D-xylulose-5-phosphate synthase [EC:2.2.1.7]"
148	"ko:K01663 HIS7; glutamine amidotransferase / cyclase [EC:2.4.2.- 4.1.3.-]"
149	"ko:K01679 E4.2.1.2B; fumarate hydratase, class II [EC:4.2.1.2]"
150	"ko:K01681 ACO; aconitate hydratase [EC:4.2.1.3]"
151	"ko:K01687 ilvD; dihydroxy-acid dehydratase [EC:4.2.1.9]"
152	"ko:K01689 ENO; enolase [EC:4.2.1.11]"
153	"ko:K01693 hisB; imidazoleglycerol-phosphate dehydratase [EC:4.2.1.19]"
154	"ko:K01695 trpA; tryptophan synthase alpha chain [EC:4.2.1.20]"
155	"ko:K01696 trpB; tryptophan synthase beta chain [EC:4.2.1.20]"
156	"ko:K01698 hemB; porphobilinogen synthase [EC:4.2.1.24]"
157	"ko:K01703 leuC; 3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit [EC:4.2.1.33 4.2.1.35]"
158	"ko:K01704 leuD; 3-isopropylmalate/(R)-2-methylmalate dehydratase small subunit [EC:4.2.1.33 4.2.1.35]"
159	"ko:K01714 dapA; 4-hydroxy-tetrahydrodipicolinate synthase [EC:4.3.3.7]"
160	"ko:K01719 hemD; uroporphyrinogen-III synthase [EC:4.2.1.75]"
161	"ko:K01735 aroB; 3-dehydroquinate synthase [EC:4.2.3.4]"
162	"ko:K01736 aroC; chorismate synthase [EC:4.2.3.5]"
163	"ko:K01739 metB; cystathionine gamma-synthase [EC:2.5.1.48]"
164	"ko:K01749 hemC; hydroxymethylbilane synthase [EC:2.5.1.61]"
165	"ko:K01754 E4.3.1.19; threonine dehydratase [EC:4.3.1.19]"
166	"ko:K01755 argH; argininosuccinate lyase [EC:4.3.2.1]"
167	"ko:K01756 purB; adenylosuccinate lyase [EC:4.3.2.2]"
168	"ko:K01760 metC; cystathionine beta-lyase [EC:4.4.1.8]"
169	"ko:K01762 ACS; 1-aminocyclopropane-1-carboxylate synthase [EC:4.4.1.14]"
170	"ko:K01770 ispF; 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase

	[EC:4.6.1.12]"
171	"ko:K01772 hemH; ferrochelatase [EC:4.99.1.1]"
172	"ko:K01778 dapF; diaminopimelate epimerase [EC:5.1.1.7]"
173	"ko:K01783 rpe; ribulose-phosphate 3-epimerase [EC:5.1.3.1]"
174	"ko:K01784 galE; UDP-glucose 4-epimerase [EC:5.1.3.2]"
175	"ko:K01785 galM; aldose 1-epimerase [EC:5.1.3.3]"
176	"ko:K01792 E5.1.3.15; glucose-6-phosphate 1-epimerase [EC:5.1.3.15]"
177	"ko:K01803 TPI; triosephosphate isomerase (TIM) [EC:5.3.1.1]"
178	"ko:K01807 rpiA; ribose 5-phosphate isomerase A [EC:5.3.1.6]"
179	"ko:K01809 manA; mannose-6-phosphate isomerase [EC:5.3.1.8]"
180	"ko:K01810 GPI; glucose-6-phosphate isomerase [EC:5.3.1.9]"
181	"ko:K01814 hisA; phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase [EC:5.3.1.16]"
182	"ko:K01817 trpF; phosphoribosylanthranilate isomerase [EC:5.3.1.24]"
183	"ko:K01823 idi; isopentenyl-diphosphate delta-isomerase [EC:5.3.3.2]"
184	"ko:K01824 EBP; cholestenol delta-isomerase [EC:5.3.3.5]"
185	"ko:K01834 PGAM; 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase [EC:5.4.2.11]"
186	"ko:K01835 pgm; phosphoglucomutase [EC:5.4.2.2]"
187	"ko:K01836 E5.4.2.3; phosphoacetylglucosamine mutase [EC:5.4.2.3]"
188	"ko:K01845 hemL; glutamate-1-semialdehyde 2,1-aminomutase [EC:5.4.3.8]"
189	"ko:K01850 E5.4.99.5; chorismate mutase [EC:5.4.99.5]"
190	"ko:K01853 E5.4.99.8; cycloartenol synthase [EC:5.4.99.8]"
191	"ko:K01858 E5.5.1.4; myo-inositol-1-phosphate synthase [EC:5.5.1.4]"
192	"ko:K01859 E5.5.1.6; chalcone isomerase [EC:5.5.1.6]"
193	"ko:K01885 EARS; glutamyl-tRNA synthetase [EC:6.1.1.17]"
194	"ko:K01895 ACSS; acetyl-CoA synthetase [EC:6.2.1.1]"
195	"ko:K01899 LSC1; succinyl-CoA synthetase alpha subunit [EC:6.2.1.4 6.2.1.5]"
196	"ko:K01900 LSC2; succinyl-CoA synthetase beta subunit [EC:6.2.1.4 6.2.1.5]"
197	"ko:K01904 4CL; 4-coumarate--CoA ligase [EC:6.2.1.12]"

198	"ko:K01918 panC; pantoate--beta-alanine ligase [EC:6.3.2.1]"
199	"ko:K01923 purC; phosphoribosylaminoimidazole-succinocarboxamide synthase [EC:6.3.2.6]"
200	"ko:K01933 purM; phosphoribosylformylglycinamide cyclo-ligase [EC:6.3.3.1]"
201	"ko:K01940 argG; argininosuccinate synthase [EC:6.3.4.5]"
202	"ko:K01945 purD; phosphoribosylamine--glycine ligase [EC:6.3.4.13]"
203	"ko:K01952 purL; phosphoribosylformylglycinamide synthase [EC:6.3.5.3]"
204	"ko:K01953 asnB; asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4]"
205	"ko:K01961 accC; acetyl-CoA carboxylase, biotin carboxylase subunit [EC:6.4.1.2 6.3.4.14]"
206	"ko:K01962 accA; acetyl-CoA carboxylase carboxyl transferase subunit alpha [EC:6.4.1.2]"
207	"ko:K02160 accB; acetyl-CoA carboxylase biotin carboxyl carrier protein"
208	"ko:K02259 COX15; cytochrome c oxidase assembly protein subunit 15"
209	"ko:K02291 crtB; phytoene synthase [EC:2.5.1.32]"
210	"ko:K02293 PDS; 15-cis-phytoene desaturase [EC:1.3.5.5]"
211	"ko:K02437 gcvH; glycine cleavage system H protein"
212	"ko:K02492 hemA; glutamyl-tRNA reductase [EC:1.2.1.70]"
213	"ko:K02523 ispB; octaprenyl-diphosphate synthase [EC:2.5.1.90]"
214	"ko:K02548 menA; 1,4-dihydroxy-2-naphthoate octaprenyltransferase [EC:2.5.1.74 2.5.1.-]"
215	"ko:K02552 menF; menaquinone-specific isochorismate synthase [EC:5.4.4.2]"
216	"ko:K03183 ubiE; ubiquinone/menaquinone biosynthesis methyltransferase [EC:2.1.1.163 2.1.1.201]"
217	"ko:K03403 chlH; magnesium chelatase subunit H [EC:6.6.1.1]"
218	"ko:K03404 chlD; magnesium chelatase subunit D [EC:6.6.1.1]"
219	"ko:K03405 chlI; magnesium chelatase subunit I [EC:6.6.1.1]"
220	"ko:K03428 E2.1.1.11; magnesium-protoporphyrin O-methyltransferase [EC:2.1.1.11]"
221	"ko:K03526 E1.17.7.1; (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase [EC:1.17.7.1]"

222	"ko:K03527 ispH; 4-hydroxy-3-methylbut-2-enyl diphosphate reductase [EC:1.17.1.2]"
223	"ko:K03781 katE; catalase [EC:1.11.1.6]"
224	"ko:K03787 surE; 5'-nucleotidase [EC:3.1.3.5]"
225	"ko:K03841 FBP; fructose-1,6-bisphosphatase I [EC:3.1.3.11]"
226	"ko:K04035 E1.14.13.81; magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase [EC:1.14.13.81]"
227	"ko:K04040 chlG; chlorophyll synthase [EC:2.5.1.62]"
228	"ko:K04120 E5.5.1.13; ent-copalyl diphosphate synthase [EC:5.5.1.13]"
229	"ko:K04121 E4.2.3.19; ent-kaurene synthase [EC:4.2.3.19]"
230	"ko:K04122 GA3; ent-kaurene oxidase [EC:1.14.13.78]"
231	"ko:K04124 E1.14.11.15; gibberellin 3-beta-dioxygenase [EC:1.14.11.15]"
232	"ko:K04518 pheA2; prephenate dehydratase [EC:4.2.1.51]"
233	"ko:K05278 FLS; flavonol synthase [EC:1.14.11.23]"
234	"ko:K05280 E1.14.13.21; flavonoid 3'-monooxygenase [EC:1.14.13.21]"
235	"ko:K05349 bglX; beta-glucosidase [EC:3.2.1.21]"
236	"ko:K05350 bglB; beta-glucosidase [EC:3.2.1.21]"
237	"ko:K05356 SPS; all-trans-nonaprenyl-diphosphate synthase [EC:2.5.1.84 2.5.1.85]"
238	"ko:K05359 ADT; arogenate/prephenate dehydratase [EC:4.2.1.91 4.2.1.51]"
239	"ko:K05917 CYP51; sterol 14-demethylase [EC:1.14.13.70]"
240	"ko:K05928 E2.1.1.95; tocopherol O-methyltransferase [EC:2.1.1.95]"
241	"ko:K05933 E1.14.17.4; aminocyclopropanecarboxylate oxidase [EC:1.14.17.4]"
242	"ko:K06001 trpB; tryptophan synthase beta chain [EC:4.2.1.20]"
243	"ko:K06125 COQ2; 4-hydroxybenzoate hexaprenyltransferase [EC:2.5.1.-]"
244	"ko:K06126 COQ6; ubiquinone biosynthesis monooxygenase Coq6 [EC:1.14.13.-]"
245	"ko:K06127 COQ5; ubiquinone biosynthesis methyltransferase [EC:2.1.1.201]"
246	"ko:K06443 lcyB; lycopene beta-cyclase [EC:5.5.1.19]"
247	"ko:K06444 lcyE; lycopene epsilon-cyclase [EC:5.5.1.18]"
248	"ko:K07513 ACAA1; acetyl-CoA acyltransferase 1 [EC:2.3.1.16]"

249	"ko:K08081 TR1; Tropinone reductase 1 [EC:1.1.1.206]"
250	"ko:K08099 E3.1.1.14; chlorophyllase [EC:3.1.1.14]"
251	"ko:K08233 E3.1.1.78; polyneuridine-aldehyde esterase [EC:3.1.1.78]"
252	"ko:K08242 E2.1.1.143; 24-methylenesterol C-methyltransferase [EC:2.1.1.143]"
253	"ko:K09588 CYP90A1; cytochrome P450, family 90, subfamily A, polypeptide 1 [EC:1.14.-.-]"
254	"ko:K09591 DET2; steroid 5-alpha-reductase [EC:1.3.1.22]"
255	"ko:K09699 DBT; 2-oxoisovalerate dehydrogenase E2 component (dihydrolipoyl transacylase) [EC:2.3.1.168]"
256	"ko:K09753 CCR; cinnamoyl-CoA reductase [EC:1.2.1.44]"
257	"ko:K09754 CYP98A3; coumaroylquininate(coumaroylshikimate) 3'-monooxygenase [EC:1.14.13.36]"
258	"ko:K09755 CYP84A; ferulate-5-hydroxylase [EC:1.14.-.-]"
259	"ko:K09828 DHCR24; delta24-sterol reductase [EC:1.3.1.72]"
260	"ko:K09832 CYP710A; cytochrome P450, family 710, subfamily A"
261	"ko:K09833 HPT; homogentisate phytyltransferase / homogentisate geranylgeranyltransferase [EC:2.5.1.115 2.5.1.116]"
262	"ko:K09834 VTE1; tocopherol cyclase [EC:5.5.1.24]"
263	"ko:K09835 crtISO; prolycopene isomerase [EC:5.2.1.13]"
264	"ko:K09837 LUT1; carotene epsilon-monooxygenase [EC:1.14.99.45]"
265	"ko:K09838 ZEP; zeaxanthin epoxidase [EC:1.14.13.90]"
266	"ko:K09839 VDE; violaxanthin de-epoxidase [EC:1.10.99.3]"
267	"ko:K09840 NCED; 9-cis-epoxycarotenoid dioxygenase [EC:1.13.11.51]"
268	"ko:K09842 AAO3; abscisic-aldehyde oxidase [EC:1.2.3.14]"
269	"ko:K10046 GME; GDP-D-mannose 3', 5'-epimerase [EC:5.1.3.18 5.1.3.-]"
270	"ko:K10047 VTC4; inositol-phosphate phosphatase / L-galactose 1-phosphate phosphatase [EC:3.1.3.25 3.1.3.93]"
271	"ko:K10206 E2.6.1.83; LL-diaminopimelate aminotransferase [EC:2.6.1.83]"
272	"ko:K10251 HSD17B12; 17beta-estradiol 17-dehydrogenase / very-long-chain 3-oxoacyl-CoA reductase [EC:1.1.1.62 1.1.1.330]"
273	"ko:K10258 TER; very-long-chain enoyl-CoA reductase [EC:1.3.1.93]"
274	"ko:K10703 PHS1; very-long-chain (3R)-3-hydroxyacyl-CoA dehydratase"

	[EC:4.2.1.134]"
275	"ko:K10720 PHM; CYP306A1; ecdysteroid 25-hydroxylase"
276	"ko:K10757 E2.4.1.91; flavonol 3-O-glucosyltransferase [EC:2.4.1.91]"
277	"ko:K10775 E4.3.1.24; phenylalanine ammonia-lyase [EC:4.3.1.24]"
278	"ko:K10960 chlP; geranylgeranyl reductase [EC:1.3.1.83]"
279	"ko:K11188 PRDX6; peroxiredoxin 6, 1-Cys peroxiredoxin [EC:1.11.1.7 1.11.1.15 3.1.1.-]"
280	"ko:K11517 HAO; (S)-2-hydroxy-acid oxidase [EC:1.1.3.15]"
281	"ko:K11755 hisIE; phosphoribosyl-ATP pyrophosphohydrolase / phosphoribosyl- AMP cyclohydrolase [EC:3.6.1.31 3.5.4.19]"
282	"ko:K11778 DHDDS; ditrans, polycis-polyprenyl diphosphate synthase [EC:2.5.1.87]"
283	"ko:K11813 CYP79B3; tryptophan N-monooxygenase [EC:1.14.13.125]"
284	"ko:K11818 CYP83B1; cytochrome P450, family 83, subfamily B, polypeptide 1 [EC:1.14.-.-]"
285	"ko:K12373 HEXA_B; hexosaminidase [EC:3.2.1.52]"
286	"ko:K12446 E2.7.1.46; L-arabinokinase [EC:2.7.1.46]"
287	"ko:K12447 USP; UDP-sugar pyrophosphorylase [EC:2.7.7.64]"
288	"ko:K12448 UXE; UDP-arabinose 4-epimerase [EC:5.1.3.5]"
289	"ko:K12449 AXS; UDP-apiiose/xylose synthase"
290	"ko:K12450 RHM; UDP-glucose 4,6-dehydratase [EC:4.2.1.76]"
291	"ko:K12451 UER1; 3,5-epimerase/4-reductase [EC:5.1.3.- 1.1.1.-]"
292	"ko:K12502 VTE3; MPBQ/MSBQ methyltransferase [EC:2.1.1.295]"
293	"ko:K12524 thrA; bifunctional aspartokinase / homoserine dehydrogenase 1 [EC:2.7.2.4 1.1.1.3]"
294	"ko:K12640 CYP85A2; brassinosteroid-6-oxidase 2 [EC:1.14.-.-]"
295	"ko:K13065 E2.3.1.133; shikimate O-hydroxycinnamoyltransferase [EC:2.3.1.133]"
296	"ko:K13066 E2.1.1.68; caffeic acid 3-O-methyltransferase [EC:2.1.1.68]"
297	"ko:K13071 PAO; pheophorbide a oxygenase [EC:1.14.12.20]"
298	"ko:K13427 NOA1; nitric-oxide synthase, plant [EC:1.14.13.39]"

299	"ko:K13545 ACD2; red chlorophyll catabolite reductase [EC:1.3.1.80]"
300	"ko:K13600 CAO; chlorophyllide a oxygenase [EC:1.14.13.122]"
301	"ko:K13606 NOL; chlorophyll(ide) b reductase [EC:1.1.1.294]"
302	"ko:K13648 GAUT; alpha-1,4-galacturonosyltransferase [EC:2.4.1.43]"
303	"ko:K13789 GGPS; geranylgeranyl diphosphate synthase, type II [EC:2.5.1.1 2.5.1.10 2.5.1.29]"
304	"ko:K13832 aroDE; 3-dehydroquinate dehydratase / shikimate dehydrogenase [EC:4.2.1.10 1.1.1.25]"
305	"ko:K14066 GPS; geranyl diphosphate synthase [EC:2.5.1.1]"
306	"ko:K14085 ALDH7A1; aldehyde dehydrogenase family 7 member A1 [EC:1.2.1.31 1.2.1.8 1.2.1.3]"
307	"ko:K14190 VTC2_5; GDP-L-galactose phosphorylase [EC:2.7.7.69]"
308	"ko:K14272 GGAT; glutamate--glyoxylate aminotransferase [EC:2.6.1.4 2.6.1.2 2.6.1.44 2.6.1.-]"
309	"ko:K14454 GOT1; aspartate aminotransferase, cytoplasmic [EC:2.6.1.1]"
310	"ko:K14455 GOT2; aspartate aminotransferase, mitochondrial [EC:2.6.1.1]"
311	"ko:K14677 ACY1; aminoacylase [EC:3.5.1.14]"
312	"ko:K14682 argAB; amino-acid N-acetyltransferase [EC:2.3.1.1]"
313	"ko:K14759 PHYLLO; isochorismate synthase / 2-succinyl-5-enolpyruvyl-6- hydroxy-3-cyclohexene-1-carboxylate synthase / 2-succinyl-6-hydroxy-2,4- cyclohexadiene-1-carboxylate synthase / O-succinylbenzoate synthase [EC:5.4.4.2 2.2.1.9 4.2.99.20 4.2.1.113]"
314	"ko:K14760 AAE14; acyl-activating enzyme 14 [EC:6.2.1.26]"
315	"ko:K15095 E1.1.1.208; (+)-neomenthol dehydrogenase [EC:1.1.1.208]"
316	"ko:K15397 KCS; 3-ketoacyl-CoA synthase [EC:2.3.1.199]"
317	"ko:K15404 CER1; aldehyde decarbonylase [EC:4.1.99.5]"
318	"ko:K15633 gpmI; 2,3-bisphosphoglycerate-independent phosphoglycerate mutase [EC:5.4.2.12]"
319	"ko:K15634 gpmB; probable phosphoglycerate mutase [EC:5.4.2.12]"
320	"ko:K15746 crtZ; beta-carotene 3-hydroxylase [EC:1.14.13.129]"
321	"ko:K15747 LUT5; beta-ring hydroxylase [EC:1.14.-.-]"
322	"ko:K15891 FLDH; farnesol dehydrogenase [EC:1.1.1.216]"

323	"ko:K15893 HPR1; hydroxypyruvate reductase 1"
324	"ko:K15918 GLYK; D-glycerate 3-kinase [EC:2.7.1.31]"
325	"ko:K15919 HPR2; hydroxypyruvate reductase 2"
326	"ko:K17497 PMM; phosphomannomutase [EC:5.4.2.8]"
327	"ko:K17744 GalDH; L-galactose dehydrogenase [EC:1.1.1.316]"
328	"ko:K18368 CSE; caffeoylshikimate esterase [EC:3.1.1.-]"
329	"ko:K18649 IMPL2; inositol-phosphate phosphatase / L-galactose 1-phosphate phosphatase / histidinol-phosphatase [EC:3.1.3.25 3.1.3.93 3.1.3.15]"

Table A4 KOs associated pathways in SCFG

Sr. no.	Pathways	Transcripts mapped KO
1	"ko01100 Metabolic pathways"	830
2	"ko01110 Biosynthesis of secondary metabolites"	351
3	"ko01120 Microbial metabolism in diverse environments"	147
4	"ko01130 Biosynthesis of antibiotics"	188
5	"ko01200 Carbon metabolism"	89
6	"ko01210 2-Oxocarboxylic acid metabolism"	27
7	"ko01212 Fatty acid metabolism"	31
8	"ko01230 Biosynthesis of amino acids"	102
9	"ko01220 Degradation of aromatic compounds"	4
10	"ko00010 Glycolysis / Gluconeogenesis"	35
11	"ko00020 Citrate cycle (TCA cycle)"	20
12	"ko00030 Pentose phosphate pathway"	20
13	"ko00040 Pentose and glucuronate interconversions"	15
14	"ko00051 Fructose and mannose metabolism"	20
15	"ko00052 Galactose metabolism"	18
16	"ko00053 Ascorbate and aldarate metabolism"	15
17	"ko00500 Starch and sucrose metabolism"	34
18	"ko00520 Amino sugar and nucleotide sugar"	39

	metabolism"	
19	"ko00620 Pyruvate metabolism"	35
20	"ko00630 Glyoxylate and dicarboxylate metabolism"	26
21	"ko00640 Propanoate metabolism"	14
22	"ko00650 Butanoate metabolism"	10
23	"ko00660 C5-Branched dibasic acid metabolism"	5
24	"ko00562 Inositol phosphate metabolism"	20
25	"ko00190 Oxidative phosphorylation"	78
26	"ko00195 Photosynthesis"	32
27	"ko00196 Photosynthesis - antenna proteins"	7
28	"ko00710 Carbon fixation in photosynthetic organisms"	26
29	"ko00720 Carbon fixation pathways in prokaryotes"	14
30	"ko00680 Methane metabolism"	19
31	"ko00910 Nitrogen metabolism"	13
32	"ko00920 Sulfur metabolism"	19
33	"ko00061 Fatty acid biosynthesis"	15
34	"ko00062 Fatty acid elongation"	9
35	"ko00071 Fatty acid degradation"	13
36	"ko00072 Synthesis and degradation of ketone bodies"	3
37	"ko00073 Cutin, suberine and wax biosynthesis"	5
38	"ko00100 Steroid biosynthesis"	20
39	"ko00140 Steroid hormone biosynthesis"	6
40	"ko00561 Glycerolipid metabolism"	25
41	"ko00564 Glycerophospholipid metabolism"	38
42	"ko00565 Ether lipid metabolism"	8
43	"ko00600 Sphingolipid metabolism"	17
44	"ko00590 Arachidonic acid metabolism"	9
45	"ko00591 Linoleic acid metabolism"	5
46	"ko00592 alpha-Linolenic acid metabolism"	12

47	"ko01040 Biosynthesis of unsaturated fatty acids"	15
48	"ko00230 Purine metabolism"	91
49	"ko00240 Pyrimidine metabolism"	71
50	"ko00250 Alanine, aspartate and glutamate metabolism"	27
51	"ko00260 Glycine, serine and threonine metabolism"	35
52	"ko00270 Cysteine and methionine metabolism"	33
53	"ko00280 Valine, leucine and isoleucine degradation"	20
54	"ko00290 Valine, leucine and isoleucine biosynthesis"	10
55	"ko00300 Lysine biosynthesis"	9
56	"ko00310 Lysine degradation"	12
57	"ko00330 Arginine and proline metabolism"	40
58	"ko00340 Histidine metabolism"	16
59	"ko00350 Tyrosine metabolism"	18
60	"ko00360 Phenylalanine metabolism"	15
61	"ko00380 Tryptophan metabolism"	13
62	"ko00400 Phenylalanine, tyrosine and tryptophan biosynthesis"	24
63	"ko00410 beta-Alanine metabolism"	17
64	"ko00430 Taurine and hypotaurine metabolism"	3
65	"ko00440 Phosphonate and phosphinate metabolism"	4
66	"ko00450 Selenocompound metabolism"	11
67	"ko00460 Cyanoamino acid metabolism"	8
68	"ko00471 D-Glutamine and D-glutamate metabolism"	2
69	"ko00480 Glutathione metabolism"	17
70	"ko00510 N-Glycan biosynthesis"	31
71	"ko00513 Various types of N-glycan biosynthesis"	22
72	"ko00512 Mucin type O-Glycan biosynthesis"	1
73	"ko00514 Other types of O-glycan biosynthesis"	2
74	"ko00532 Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate"	2

75	"ko00534 Glycosaminoglycan biosynthesis - heparan sulfate / heparin"	2
76	"ko00533 Glycosaminoglycan biosynthesis - keratan sulfate"	1
77	"ko00531 Glycosaminoglycan degradation"	5
78	"ko00563 Glycosylphosphatidylinositol (GPI)-anchor biosynthesis"	22
79	"ko00603 Glycosphingolipid biosynthesis - globo series"	3
80	"ko00604 Glycosphingolipid biosynthesis - ganglio series"	3
81	"ko00540 Lipopolysaccharide biosynthesis"	7
82	"ko00550 Peptidoglycan biosynthesis"	1
83	"ko00511 Other glycan degradation"	9
84	"ko00730 Thiamine metabolism"	10
85	"ko00740 Riboflavin metabolism"	10
86	"ko00750 Vitamin B6 metabolism"	7
87	"ko00760 Nicotinate and nicotinamide metabolism"	12
88	"ko00770 Pantothenate and CoA biosynthesis"	16
89	"ko00780 Biotin metabolism"	7
90	"ko00785 Lipoic acid metabolism"	2
91	"ko00790 Folate biosynthesis"	13
92	"ko00670 One carbon pool by folate"	10
93	"ko00830 Retinol metabolism"	9
94	"ko00860 Porphyrin and chlorophyll metabolism"	36
95	"ko00130 Ubiquinone and other terpenoid-quinone biosynthesis"	21
96	"ko00900 Terpenoid backbone biosynthesis"	30
97	"ko00902 Monoterpenoid biosynthesis"	3
98	"ko00909 Sesquiterpenoid and triterpenoid biosynthesis"	6
99	"ko00904 Diterpenoid biosynthesis"	7

100	"ko00906 Carotenoid biosynthesis"	18
101	"ko00905 Brassinosteroid biosynthesis"	5
102	"ko00908 Zeatin biosynthesis"	4
103	"ko00903 Limonene and pinene degradation"	2
104	"ko00281 Geraniol degradation"	1
105	"ko01051 Biosynthesis of ansamycins"	1
106	"ko00253 Tetracycline biosynthesis"	4
107	"ko00523 Polyketide sugar unit biosynthesis"	1
108	"ko01053 Biosynthesis of siderophore group nonribosomal peptides"	2
109	"ko00940 Phenylpropanoid biosynthesis"	17
110	"ko00945 Stilbenoid, diarylheptanoid and gingerol biosynthesis"	5
111	"ko00941 Flavonoid biosynthesis"	10
112	"ko00944 Flavone and flavonol biosynthesis"	3
113	"ko00942 Anthocyanin biosynthesis"	1
114	"ko00901 Indole alkaloid biosynthesis"	2
115	"ko00950 Isoquinoline alkaloid biosynthesis"	7
116	"ko00960 Tropane, piperidine and pyridine alkaloid biosynthesis"	9
117	"ko00232 Caffeine metabolism"	3
118	"ko00965 Betalain biosynthesis"	1
119	"ko00966 Glucosinolate biosynthesis"	1
120	"ko00521 Streptomycin biosynthesis"	4
121	"ko00524 Butirosin and neomycin biosynthesis"	1
122	"ko00401 Novobiocin biosynthesis"	2
123	"ko00254 Aflatoxin biosynthesis"	1
124	"ko00362 Benzoate degradation"	3
125	"ko00627 Aminobenzoate degradation"	3
126	"ko00364 Fluorobenzoate degradation"	1

127	"ko00625 Chloroalkane and chloroalkene degradation"	5
128	"ko00361 Chlorocyclohexane and chlorobenzene degradation"	2
129	"ko00623 Toluene degradation"	1
130	"ko00633 Nitrotoluene degradation"	1
131	"ko00643 Styrene degradation"	3
132	"ko00791 Atrazine degradation"	1
133	"ko00351 DDT degradation"	1
134	"ko00363 Bisphenol degradation"	1
135	"ko00626 Naphthalene degradation"	3
136	"ko00624 Polycyclic aromatic hydrocarbon degradation"	2
137	"ko00980 Metabolism of xenobiotics by cytochrome P450"	8
138	"ko00982 Drug metabolism - cytochrome P450"	8
139	"ko00983 Drug metabolism - other enzymes"	14
140	"ko03020 RNA polymerase"	28
141	"ko03022 Basal transcription factors"	30
142	"ko03040 Spliceosome"	100
143	"ko03010 Ribosome"	128
144	"ko00970 Aminoacyl-tRNA biosynthesis"	26
145	"ko03013 RNA transport"	95
146	"ko03015 mRNA surveillance pathway"	48
147	"ko03008 Ribosome biogenesis in eukaryotes"	56
148	"ko03060 Protein export"	26
149	"ko04141 Protein processing in endoplasmic reticulum"	77
150	"ko04130 SNARE interactions in vesicular transport"	17
151	"ko04120 Ubiquitin mediated proteolysis"	61
152	"ko04122 Sulfur relay system"	10
153	"ko03050 Proteasome"	34

154	"ko03018 RNA degradation"	49
155	"ko03030 DNA replication"	29
156	"ko03410 Base excision repair"	23
157	"ko03420 Nucleotide excision repair"	37
158	"ko03430 Mismatch repair"	20
159	"ko03440 Homologous recombination"	22
160	"ko03450 Non-homologous end-joining"	8
161	"ko03460 Fanconi anemia pathway"	24
162	"ko02010 ABC transporters"	5
163	"ko03070 Bacterial secretion system"	6
164	"ko02020 Two-component system"	15
165	"ko04014 Ras signaling pathway"	10
166	"ko04015 Rap1 signaling pathway"	9
167	"ko04010 MAPK signaling pathway"	14
168	"ko04013 MAPK signaling pathway - fly"	2
169	"ko04011 MAPK signaling pathway - yeast"	6
170	"ko04012 ErbB signaling pathway"	5
171	"ko04310 Wnt signaling pathway"	17
172	"ko04330 Notch signaling pathway"	7
173	"ko04340 Hedgehog signaling pathway"	3
174	"ko04350 TGF-beta signaling pathway"	10
175	"ko04390 Hippo signaling pathway"	8
176	"ko04391 Hippo signaling pathway - fly"	7
177	"ko04370 VEGF signaling pathway"	7
178	"ko04630 Jak-STAT signaling pathway"	2
179	"ko04064 NF-kappa B signaling pathway"	6
180	"ko04668 TNF signaling pathway"	5
181	"ko04066 HIF-1 signaling pathway"	15
182	"ko04068 FoxO signaling pathway"	20

183	"ko04020 Calcium signaling pathway"	10
184	"ko04070 Phosphatidylinositol signaling system"	18
185	"ko04071 Sphingolipid signaling pathway"	21
186	"ko04024 cAMP signaling pathway"	12
187	"ko04022 cGMP-PKG signaling pathway"	10
188	"ko04151 PI3K-Akt signaling pathway"	22
189	"ko04152 AMPK signaling pathway"	26
190	"ko04150 mTOR signaling pathway"	11
191	"ko04075 Plant hormone signal transduction"	39
192	"ko04080 Neuroactive ligand-receptor interaction"	3
193	"ko04144 Endocytosis"	44
194	"ko04145 Phagosome"	30
195	"ko04142 Lysosome"	34
196	"ko04146 Peroxisome"	38
197	"ko04140 Regulation of autophagy"	14
198	"ko04810 Regulation of actin cytoskeleton"	19
199	"ko04110 Cell cycle"	48
200	"ko04111 Cell cycle - yeast"	44
201	"ko04112 Cell cycle - Caulobacter"	4
202	"ko04113 Meiosis - yeast"	34
203	"ko04114 Oocyte meiosis"	29
204	"ko04210 Apoptosis"	5
205	"ko04115 p53 signaling pathway"	12
206	"ko04510 Focal adhesion"	10
207	"ko04520 Adherens junction"	8
208	"ko04530 Tight junction"	14
209	"ko04540 Gap junction"	6
210	"ko04550 Signaling pathways regulating pluripotency of stem cells"	5
211	"ko04611 Platelet activation"	5

212	"ko04620 Toll-like receptor signaling pathway"	6
213	"ko04621 NOD-like receptor signaling pathway"	5
214	"ko04622 RIG-I-like receptor signaling pathway"	7
215	"ko04623 Cytosolic DNA-sensing pathway"	15
216	"ko04650 Natural killer cell mediated cytotoxicity"	4
217	"ko04612 Antigen processing and presentation"	10
218	"ko04660 T cell receptor signaling pathway"	8
219	"ko04662 B cell receptor signaling pathway"	5
220	"ko04664 Fc epsilon RI signaling pathway"	5
221	"ko04666 Fc gamma R-mediated phagocytosis"	14
222	"ko04670 Leukocyte transendothelial migration"	6
223	"ko04062 Chemokine signaling pathway"	7
224	"ko04911 Insulin secretion"	1
225	"ko04910 Insulin signaling pathway"	19
226	"ko04922 Glucagon signaling pathway"	16
227	"ko04920 Adipocytokine signaling pathway"	7
228	"ko03320 PPAR signaling pathway"	8
229	"ko04912 GnRH signaling pathway"	7
230	"ko04913 Ovarian steroidogenesis"	4
231	"ko04915 Estrogen signaling pathway"	7
232	"ko04914 Progesterone-mediated oocyte maturation"	21
233	"ko04917 Prolactin signaling pathway"	6
234	"ko04921 Oxytocin signaling pathway"	15
235	"ko04918 Thyroid hormone synthesis"	5
236	"ko04919 Thyroid hormone signaling pathway"	17
237	"ko04916 Melanogenesis"	5
238	"ko04614 Renin-angiotensin system"	1
239	"ko04260 Cardiac muscle contraction"	13
240	"ko04261 Adrenergic signaling in cardiomyocytes"	11

241	"ko04270 Vascular smooth muscle contraction"	6
242	"ko04970 Salivary secretion"	2
243	"ko04971 Gastric acid secretion"	1
244	"ko04972 Pancreatic secretion"	7
245	"ko04976 Bile secretion"	4
246	"ko04973 Carbohydrate digestion and absorption"	3
247	"ko04974 Protein digestion and absorption"	3
248	"ko04975 Fat digestion and absorption"	4
249	"ko04977 Vitamin digestion and absorption"	1
250	"ko04978 Mineral absorption"	6
251	"ko04962 Vasopressin-regulated water reabsorption"	7
252	"ko04960 Aldosterone-regulated sodium reabsorption"	2
253	"ko04961 Endocrine and other factor-regulated calcium reabsorption"	6
254	"ko04964 Proximal tubule bicarbonate reclamation"	4
255	"ko04966 Collecting duct acid secretion"	11
256	"ko04724 Glutamatergic synapse"	11
257	"ko04727 GABAergic synapse"	10
258	"ko04725 Cholinergic synapse"	3
259	"ko04728 Dopaminergic synapse"	11
260	"ko04726 Serotonergic synapse"	5
261	"ko04720 Long-term potentiation"	8
262	"ko04730 Long-term depression"	5
263	"ko04723 Retrograde endocannabinoid signaling"	7
264	"ko04721 Synaptic vesicle cycle"	23
265	"ko04722 Neurotrophin signaling pathway"	16
266	"ko04744 Phototransduction"	3
267	"ko04745 Phototransduction - fly"	3
268	"ko04740 Olfactory transduction"	1
269	"ko04742 Taste transduction"	1

270	"ko04750 Inflammatory mediator regulation of TRP channels"	4
271	"ko04320 Dorso-ventral axis formation"	3
272	"ko04360 Axon guidance"	7
273	"ko04380 Osteoclast differentiation"	7
274	"ko04710 Circadian rhythm"	7
275	"ko04713 Circadian entrainment"	5
276	"ko04711 Circadian rhythm - fly"	1
277	"ko04712 Circadian rhythm - plant"	19
278	"ko04626 Plant-pathogen interaction"	25
279	"ko05200 Pathways in cancer"	27
280	"ko05230 Central carbon metabolism in cancer"	13
281	"ko05231 Choline metabolism in cancer"	11
282	"ko05202 Transcriptional misregulation in cancer"	11
283	"ko05206 MicroRNAs in cancer"	16
284	"ko05205 Proteoglycans in cancer"	15
285	"ko05204 Chemical carcinogenesis"	7
286	"ko05203 Viral carcinogenesis"	40
287	"ko05210 Colorectal cancer"	10
288	"ko05212 Pancreatic cancer"	5
289	"ko05214 Glioma"	6
290	"ko05216 Thyroid cancer"	3
291	"ko05221 Acute myeloid leukemia"	4
292	"ko05220 Chronic myeloid leukemia"	4
293	"ko05217 Basal cell carcinoma"	2
294	"ko05218 Melanoma"	4
295	"ko05211 Renal cell carcinoma"	8
296	"ko05219 Bladder cancer"	3
297	"ko05215 Prostate cancer"	11
298	"ko05213 Endometrial cancer"	6

299	"ko05222 Small cell lung cancer"	7
300	"ko05223 Non-small cell lung cancer"	5
301	"ko05322 Systemic lupus erythematosus"	8
302	"ko05323 Rheumatoid arthritis"	14
303	"ko05340 Primary immunodeficiency"	2
304	"ko05010 Alzheimer's disease"	56
305	"ko05012 Parkinson's disease"	54
306	"ko05014 Amyotrophic lateral sclerosis (ALS)"	9
307	"ko05016 Huntington's disease"	72
308	"ko05020 Prion diseases"	6
309	"ko05030 Cocaine addiction"	1
310	"ko05031 Amphetamine addiction"	4
311	"ko05032 Morphine addiction"	2
312	"ko05033 Nicotine addiction"	2
313	"ko05034 Alcoholism"	14
314	"ko05410 Hypertrophic cardiomyopathy (HCM)"	5
315	"ko05412 Arrhythmogenic right ventricular cardiomyopathy (ARVC)"	2
316	"ko05414 Dilated cardiomyopathy"	2
317	"ko05416 Viral myocarditis"	4
318	"ko04940 Type I diabetes mellitus"	2
319	"ko04930 Type II diabetes mellitus"	4
320	"ko04932 Non-alcoholic fatty liver disease (NAFLD)"	46
321	"ko05110 Vibrio cholerae infection"	19
322	"ko05120 Epithelial cell signaling in Helicobacter pylori infection"	17
323	"ko05130 Pathogenic Escherichia coli infection"	11
324	"ko05132 Salmonella infection"	13
325	"ko05131 Shigellosis"	14
326	"ko05133 Pertussis"	7

327	"ko05134 Legionellosis"	12
328	"ko05152 Tuberculosis"	23
329	"ko05100 Bacterial invasion of epithelial cells"	13
330	"ko05166 HTLV-I infection"	43
331	"ko05162 Measles"	12
332	"ko05164 Influenza A"	20
333	"ko05161 Hepatitis B"	12
334	"ko05160 Hepatitis C"	11
335	"ko05168 Herpes simplex infection"	30
336	"ko05169 Epstein-Barr virus infection"	63
337	"ko05146 Amoebiasis"	3
338	"ko05145 Toxoplasmosis"	10
339	"ko05140 Leishmaniasis"	5
340	"ko05142 Chagas disease (American trypanosomiasis)"	9
341	"ko05143 African trypanosomiasis"	2
342	"ko01502 Vancomycin resistance"	1

Table A5 KO associated pathways in SCTC

Sr. no.	Pathways	Transcripts mapped KO
1	"ko01100 Metabolic pathways"	796
2	"ko01110 Biosynthesis of secondary metabolites"	341
3	"ko01120 Microbial metabolism in diverse environments"	144
4	"ko01130 Biosynthesis of antibiotics"	184
5	"ko01200 Carbon metabolism"	88
6	"ko01210 2-Oxocarboxylic acid metabolism"	28
7	"ko01212 Fatty acid metabolism"	29
8	"ko01230 Biosynthesis of amino acids"	104
9	"ko01220 Degradation of aromatic compounds"	5
10	"ko00010 Glycolysis / Gluconeogenesis"	32

11	"ko00020 Citrate cycle (TCA cycle)"	22
12	"ko00030 Pentose phosphate pathway"	17
13	"ko00040 Pentose and glucuronate interconversions"	15
14	"ko00051 Fructose and mannose metabolism"	19
15	"ko00052 Galactose metabolism"	18
16	"ko00053 Ascorbate and aldarate metabolism"	16
17	"ko00500 Starch and sucrose metabolism"	33
18	"ko00520 Amino sugar and nucleotide sugar metabolism"	36
19	"ko00620 Pyruvate metabolism"	32
20	"ko00630 Glyoxylate and dicarboxylate metabolism"	24
21	"ko00640 Propanoate metabolism"	14
22	"ko00650 Butanoate metabolism"	11
23	"ko00660 C5-Branched dibasic acid metabolism"	5
24	"ko00562 Inositol phosphate metabolism"	20
25	"ko00190 Oxidative phosphorylation"	69
26	"ko00195 Photosynthesis"	37
27	"ko00196 Photosynthesis - antenna proteins"	11
28	"ko00710 Carbon fixation in photosynthetic organisms"	25
29	"ko00720 Carbon fixation pathways in prokaryotes"	18
30	"ko00680 Methane metabolism"	21
31	"ko00910 Nitrogen metabolism"	12
32	"ko00920 Sulfur metabolism"	15
33	"ko00061 Fatty acid biosynthesis"	16
34	"ko00062 Fatty acid elongation"	7
35	"ko00071 Fatty acid degradation"	13
36	"ko00072 Synthesis and degradation of ketone bodies"	3
37	"ko00073 Cutin, suberine and wax biosynthesis"	5
38	"ko00100 Steroid biosynthesis"	18
39	"ko00120 Primary bile acid biosynthesis"	1

40	"ko00140 Steroid hormone biosynthesis"	5
41	"ko00561 Glycerolipid metabolism"	26
42	"ko00564 Glycerophospholipid metabolism"	37
43	"ko00565 Ether lipid metabolism"	9
44	"ko00600 Sphingolipid metabolism"	15
45	"ko00590 Arachidonic acid metabolism"	10
46	"ko00591 Linoleic acid metabolism"	7
47	"ko00592 alpha-Linolenic acid metabolism"	13
48	"ko01040 Biosynthesis of unsaturated fatty acids"	12
49	"ko00230 Purine metabolism"	88
50	"ko00240 Pyrimidine metabolism"	67
51	"ko00250 Alanine, aspartate and glutamate metabolism"	26
52	"ko00260 Glycine, serine and threonine metabolism"	31
53	"ko00270 Cysteine and methionine metabolism"	32
54	"ko00280 Valine, leucine and isoleucine degradation"	20
55	"ko00290 Valine, leucine and isoleucine biosynthesis"	10
56	"ko00300 Lysine biosynthesis"	10
57	"ko00310 Lysine degradation"	12
58	"ko00330 Arginine and proline metabolism"	39
59	"ko00340 Histidine metabolism"	12
60	"ko00350 Tyrosine metabolism"	17
61	"ko00360 Phenylalanine metabolism"	14
62	"ko00380 Tryptophan metabolism"	12
63	"ko00400 Phenylalanine, tyrosine and tryptophan biosynthesis"	25
64	"ko00410 beta-Alanine metabolism"	15
65	"ko00430 Taurine and hypotaurine metabolism"	3
66	"ko00440 Phosphonate and phosphinate metabolism"	3
67	"ko00450 Selenocompound metabolism"	10
68	"ko00460 Cyanoamino acid metabolism"	8

69	"ko00471 D-Glutamine and D-glutamate metabolism"	1
70	"ko00473 D-Alanine metabolism"	1
71	"ko00480 Glutathione metabolism"	16
72	"ko00510 N-Glycan biosynthesis"	30
73	"ko00513 Various types of N-glycan biosynthesis"	21
74	"ko00514 Other types of O-glycan biosynthesis"	2
75	"ko00532 Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate"	2
76	"ko00534 Glycosaminoglycan biosynthesis - heparan sulfate / heparin"	2
77	"ko00531 Glycosaminoglycan degradation"	4
78	"ko00563 Glycosylphosphatidylinositol (GPI)-anchor biosynthesis"	22
79	"ko00603 Glycosphingolipid biosynthesis - globo series"	2
80	"ko00604 Glycosphingolipid biosynthesis - ganglio series"	2
81	"ko00540 Lipopolysaccharide biosynthesis"	6
82	"ko00550 Peptidoglycan biosynthesis"	2
83	"ko00511 Other glycan degradation"	8
84	"ko00730 Thiamine metabolism"	7
85	"ko00740 Riboflavin metabolism"	8
86	"ko00750 Vitamin B6 metabolism"	7
87	"ko00760 Nicotinate and nicotinamide metabolism"	12
88	"ko00770 Pantothenate and CoA biosynthesis"	16
89	"ko00780 Biotin metabolism"	8
90	"ko00785 Lipoic acid metabolism"	2
91	"ko00790 Folate biosynthesis"	12
92	"ko00670 One carbon pool by folate"	11
93	"ko00830 Retinol metabolism"	8
94	"ko00860 Porphyrin and chlorophyll metabolism"	33
95	"ko00130 Ubiquinone and other terpenoid-quinone biosynthesis"	20

96	"ko00900 Terpenoid backbone biosynthesis"	29
97	"ko00902 Monoterpenoid biosynthesis"	2
98	"ko00909 Sesquiterpenoid and triterpenoid biosynthesis"	5
99	"ko00904 Diterpenoid biosynthesis"	5
100	"ko00906 Carotenoid biosynthesis"	17
101	"ko00905 Brassinosteroid biosynthesis"	4
102	"ko00908 Zeatin biosynthesis"	3
103	"ko00903 Limonene and pinene degradation"	2
104	"ko00281 Geraniol degradation"	1
105	"ko01051 Biosynthesis of ansamycins"	1
106	"ko00253 Tetracycline biosynthesis"	4
107	"ko00523 Polyketide sugar unit biosynthesis"	2
108	"ko01053 Biosynthesis of siderophore group nonribosomal peptides"	1
109	"ko01055 Biosynthesis of vancomycin group antibiotics"	1
110	"ko00940 Phenylpropanoid biosynthesis"	18
111	"ko00945 Stilbenoid, diarylheptanoid and gingerol biosynthesis"	5
112	"ko00941 Flavonoid biosynthesis"	7
113	"ko00944 Flavone and flavonol biosynthesis"	1
114	"ko00942 Anthocyanin biosynthesis"	1
115	"ko00901 Indole alkaloid biosynthesis"	1
116	"ko00950 Isoquinoline alkaloid biosynthesis"	7
117	"ko00960 Tropane, piperidine and pyridine alkaloid biosynthesis"	9
118	"ko00232 Caffeine metabolism"	2
119	"ko00965 Betalain biosynthesis"	1
120	"ko00966 Glucosinolate biosynthesis"	1
121	"ko00332 Carbapenem biosynthesis"	1
122	"ko00521 Streptomycin biosynthesis"	5

123	"ko00524 Butirosin and neomycin biosynthesis"	1
124	"ko00401 Novobiocin biosynthesis"	2
125	"ko00254 Aflatoxin biosynthesis"	1
126	"ko00362 Benzoate degradation"	3
127	"ko00627 Aminobenzoate degradation"	4
128	"ko00364 Fluorobenzoate degradation"	1
129	"ko00625 Chloroalkane and chloroalkene degradation"	6
130	"ko00361 Chlorocyclohexane and chlorobenzene degradation"	2
131	"ko00623 Toluene degradation"	1
132	"ko00633 Nitrotoluene degradation"	1
133	"ko00643 Styrene degradation"	3
134	"ko00791 Atrazine degradation"	2
135	"ko00930 Caprolactam degradation"	1
136	"ko00351 DDT degradation"	1
137	"ko00363 Bisphenol degradation"	2
138	"ko00626 Naphthalene degradation"	3
139	"ko00624 Polycyclic aromatic hydrocarbon degradation"	2
140	"ko00980 Metabolism of xenobiotics by cytochrome P450"	6
141	"ko00982 Drug metabolism - cytochrome P450"	6
142	"ko00983 Drug metabolism - other enzymes"	13
143	"ko03020 RNA polymerase"	27
144	"ko03022 Basal transcription factors"	30
145	"ko03040 Spliceosome"	103
146	"ko03010 Ribosome"	127
147	"ko00970 Aminoacyl-tRNA biosynthesis"	27
148	"ko03013 RNA transport"	94
149	"ko03015 mRNA surveillance pathway"	48
150	"ko03008 Ribosome biogenesis in eukaryotes"	59
151	"ko03060 Protein export"	26

152	"ko04141 Protein processing in endoplasmic reticulum"	77
153	"ko04130 SNARE interactions in vesicular transport"	17
154	"ko04120 Ubiquitin mediated proteolysis"	59
155	"ko04122 Sulfur relay system"	9
156	"ko03050 Proteasome"	34
157	"ko03018 RNA degradation"	50
158	"ko03030 DNA replication"	23
159	"ko03410 Base excision repair"	22
160	"ko03420 Nucleotide excision repair"	36
161	"ko03430 Mismatch repair"	19
162	"ko03440 Homologous recombination"	19
163	"ko03450 Non-homologous end-joining"	8
164	"ko03460 Fanconi anemia pathway"	20
165	"ko02010 ABC transporters"	4
166	"ko03070 Bacterial secretion system"	6
167	"ko02020 Two-component system"	12
168	"ko04014 Ras signaling pathway"	10
169	"ko04015 Rap1 signaling pathway"	6
170	"ko04010 MAPK signaling pathway"	13
171	"ko04013 MAPK signaling pathway - fly"	2
172	"ko04011 MAPK signaling pathway - yeast"	4
173	"ko04012 ErbB signaling pathway"	5
174	"ko04310 Wnt signaling pathway"	15
175	"ko04330 Notch signaling pathway"	7
176	"ko04340 Hedgehog signaling pathway"	2
177	"ko04350 TGF-beta signaling pathway"	9
178	"ko04390 Hippo signaling pathway"	7
179	"ko04391 Hippo signaling pathway - fly"	7
180	"ko04370 VEGF signaling pathway"	6

181	"ko04630 Jak-STAT signaling pathway"	2
182	"ko04064 NF-kappa B signaling pathway"	6
183	"ko04668 TNF signaling pathway"	4
184	"ko04066 HIF-1 signaling pathway"	14
185	"ko04068 FoxO signaling pathway"	20
186	"ko04020 Calcium signaling pathway"	7
187	"ko04070 Phosphatidylinositol signaling system"	17
188	"ko04071 Sphingolipid signaling pathway"	19
189	"ko04024 cAMP signaling pathway"	8
190	"ko04022 cGMP-PKG signaling pathway"	8
191	"ko04151 PI3K-Akt signaling pathway"	21
192	"ko04152 AMPK signaling pathway"	24
193	"ko04150 mTOR signaling pathway"	12
194	"ko04075 Plant hormone signal transduction"	38
195	"ko04144 Endocytosis"	41
196	"ko04145 Phagosome"	28
197	"ko04142 Lysosome"	33
198	"ko04146 Peroxisome"	37
199	"ko04140 Regulation of autophagy"	14
200	"ko04810 Regulation of actin cytoskeleton"	17
201	"ko04110 Cell cycle"	41
202	"ko04111 Cell cycle - yeast"	38
203	"ko04112 Cell cycle - Caulobacter"	4
204	"ko04113 Meiosis - yeast"	27
205	"ko04114 Oocyte meiosis"	29
206	"ko04210 Apoptosis"	5
207	"ko04115 p53 signaling pathway"	11
208	"ko04510 Focal adhesion"	7
209	"ko04520 Adherens junction"	7

210	"ko04530 Tight junction"	12
211	"ko04540 Gap junction"	5
212	"ko04550 Signaling pathways regulating pluripotency of stem cells"	3
213	"ko04611 Platelet activation"	4
214	"ko04620 Toll-like receptor signaling pathway"	5
215	"ko04621 NOD-like receptor signaling pathway"	4
216	"ko04622 RIG-I-like receptor signaling pathway"	6
217	"ko04623 Cytosolic DNA-sensing pathway"	15
218	"ko04650 Natural killer cell mediated cytotoxicity"	4
219	"ko04612 Antigen processing and presentation"	9
220	"ko04660 T cell receptor signaling pathway"	4
221	"ko04662 B cell receptor signaling pathway"	4
222	"ko04664 Fc epsilon RI signaling pathway"	5
223	"ko04666 Fc gamma R-mediated phagocytosis"	13
224	"ko04670 Leukocyte transendothelial migration"	2
225	"ko04062 Chemokine signaling pathway"	4
226	"ko04910 Insulin signaling pathway"	18
227	"ko04922 Glucagon signaling pathway"	15
228	"ko04920 Adipocytokine signaling pathway"	7
229	"ko03320 PPAR signaling pathway"	7
230	"ko04912 GnRH signaling pathway"	6
231	"ko04913 Ovarian steroidogenesis"	4
232	"ko04915 Estrogen signaling pathway"	7
233	"ko04914 Progesterone-mediated oocyte maturation"	20
234	"ko04917 Prolactin signaling pathway"	3
235	"ko04921 Oxytocin signaling pathway"	13
236	"ko04918 Thyroid hormone synthesis"	5
237	"ko04919 Thyroid hormone signaling pathway"	17
238	"ko04916 Melanogenesis"	4

239	"ko04614 Renin-angiotensin system"	1
240	"ko04260 Cardiac muscle contraction"	8
241	"ko04261 Adrenergic signaling in cardiomyocytes"	8
242	"ko04270 Vascular smooth muscle contraction"	5
243	"ko04970 Salivary secretion"	1
244	"ko04971 Gastric acid secretion"	1
245	"ko04972 Pancreatic secretion"	3
246	"ko04976 Bile secretion"	2
247	"ko04973 Carbohydrate digestion and absorption"	4
248	"ko04974 Protein digestion and absorption"	2
249	"ko04975 Fat digestion and absorption"	4
250	"ko04977 Vitamin digestion and absorption"	1
251	"ko04978 Mineral absorption"	5
252	"ko04962 Vasopressin-regulated water reabsorption"	4
253	"ko04960 Aldosterone-regulated sodium reabsorption"	2
254	"ko04961 Endocrine and other factor-regulated calcium reabsorption"	6
255	"ko04964 Proximal tubule bicarbonate reclamation"	3
256	"ko04966 Collecting duct acid secretion"	12
257	"ko04724 Glutamatergic synapse"	8
258	"ko04727 GABAergic synapse"	9
259	"ko04725 Cholinergic synapse"	3
260	"ko04728 Dopaminergic synapse"	8
261	"ko04726 Serotonergic synapse"	6
262	"ko04720 Long-term potentiation"	7
263	"ko04730 Long-term depression"	5
264	"ko04723 Retrograde endocannabinoid signaling"	3
265	"ko04721 Synaptic vesicle cycle"	23
266	"ko04722 Neurotrophin signaling pathway"	14
267	"ko04744 Phototransduction"	2

268	"ko04745 Phototransduction - fly"	2
269	"ko04740 Olfactory transduction"	1
270	"ko04742 Taste transduction"	2
271	"ko04750 Inflammatory mediator regulation of TRP channels"	4
272	"ko04320 Dorso-ventral axis formation"	2
273	"ko04360 Axon guidance"	5
274	"ko04380 Osteoclast differentiation"	5
275	"ko04710 Circadian rhythm"	7
276	"ko04713 Circadian entrainment"	3
277	"ko04712 Circadian rhythm - plant"	17
278	"ko04626 Plant-pathogen interaction"	25
279	"ko05200 Pathways in cancer"	24
280	"ko05230 Central carbon metabolism in cancer"	12
281	"ko05231 Choline metabolism in cancer"	12
282	"ko05202 Transcriptional misregulation in cancer"	11
283	"ko05206 MicroRNAs in cancer"	17
284	"ko05205 Proteoglycans in cancer"	12
285	"ko05204 Chemical carcinogenesis"	4
286	"ko05203 Viral carcinogenesis"	38
287	"ko05210 Colorectal cancer"	7
288	"ko05212 Pancreatic cancer"	4
289	"ko05214 Glioma"	6
290	"ko05216 Thyroid cancer"	3
291	"ko05221 Acute myeloid leukemia"	4
292	"ko05220 Chronic myeloid leukemia"	5
293	"ko05217 Basal cell carcinoma"	1
294	"ko05218 Melanoma"	4
295	"ko05211 Renal cell carcinoma"	7
296	"ko05219 Bladder cancer"	3

297	"ko05215 Prostate cancer"	10
298	"ko05213 Endometrial cancer"	5
299	"ko05222 Small cell lung cancer"	7
300	"ko05223 Non-small cell lung cancer"	5
301	"ko05322 Systemic lupus erythematosus"	8
302	"ko05323 Rheumatoid arthritis"	14
303	"ko05340 Primary immunodeficiency"	1
304	"ko05010 Alzheimer's disease"	47
305	"ko05012 Parkinson's disease"	46
306	"ko05014 Amyotrophic lateral sclerosis (ALS)"	7
307	"ko05016 Huntington's disease"	63
308	"ko05020 Prion diseases"	5
309	"ko05031 Amphetamine addiction"	3
310	"ko05032 Morphine addiction"	2
311	"ko05033 Nicotine addiction"	1
312	"ko05034 Alcoholism"	14
313	"ko05410 Hypertrophic cardiomyopathy (HCM)"	4
314	"ko05412 Arrhythmogenic right ventricular cardiomyopathy (ARVC)"	1
315	"ko05414 Dilated cardiomyopathy"	1
316	"ko05416 Viral myocarditis"	5
317	"ko04940 Type I diabetes mellitus"	2
318	"ko04930 Type II diabetes mellitus"	4
319	"ko04932 Non-alcoholic fatty liver disease (NAFLD)"	38
320	"ko05110 Vibrio cholerae infection"	19
321	"ko05120 Epithelial cell signaling in Helicobacter pylori infection"	15
322	"ko05130 Pathogenic Escherichia coli infection"	10
323	"ko05132 Salmonella infection"	10
324	"ko05131 Shigellosis"	13

325	"ko05133 Pertussis"	5
326	"ko05134 Legionellosis"	13
327	"ko05152 Tuberculosis"	21
328	"ko05100 Bacterial invasion of epithelial cells"	10
329	"ko05166 HTLV-I infection"	39
330	"ko05162 Measles"	11
331	"ko05164 Influenza A"	18
332	"ko05161 Hepatitis B"	11
333	"ko05160 Hepatitis C"	9
334	"ko05168 Herpes simplex infection"	31
335	"ko05169 Epstein-Barr virus infection"	60
336	"ko05146 Amoebiasis"	3
337	"ko05145 Toxoplasmosis"	8
338	"ko05140 Leishmaniasis"	4
339	"ko05142 Chagas disease (American trypanosomiasis)"	8
340	"ko05143 African trypanosomiasis"	2
341	"ko01502 Vancomycin resistance"	2

Table A6 Functional classification by NCBI Biosystems (AHSR)

1. Metabolism	1638
1.1 Carbohydrate Metabolism	414
Butanoate metabolism	12
Fructose and mannose metabolism	29
Glycolysis / Gluconeogenesis	43
Starch and sucrose metabolism	77
Pyruvate metabolism	31
Amino sugar and nucleotide sugar metabolism	63
Citrate cycle (TCA cycle)	47
Propanoate metabolism	17
Galactose metabolism	20
Ascorbate and aldarate metabolism	11
Inositol phosphate metabolism	28
Inositol phosphate metabolism, Ins(1,3,4)P3 => phytate	1
C5-Branched dibasic acid metabolism	1

Pentose and glucuronate interconversions	17
Glyoxylate and dicarboxylate metabolism	17
1.2 Energy Metabolism	239
Oxidative phosphorylation	137
Carbon fixation in photosynthetic organisms	16
Nitrogen metabolism	27
Methane metabolism	9
Photosynthesis - antenna proteins	11
Photosynthesis - antenna proteins	11
Photosynthesis	10
Sulfur metabolism	18
1.3 Lipid Metabolism	216
Glycerophospholipid metabolism	47
Fatty acid biosynthesis, elongation	5
Fatty acid biosynthesis	13
alpha-Linolenic acid metabolism	16
Glycerolipid metabolism	29
Fatty acid metabolism	22
Synthesis and degradation of ketone bodies	2
Biosynthesis of unsaturated fatty acids	19
Ether lipid metabolism	12
Steroid biosynthesis	17
Linoleic acid metabolism	6
Sphingolipid metabolism	23
Arachidonic acid metabolism	5
1.4 Nucleotide Metabolism	143
Pyrimidine metabolism	48
Purine metabolism	95
1.5 Amino Acid Metabolism	265
Histidine metabolism	4
Lysine biosynthesis	8
Lysine biosynthesis, 2-oxoglutarate => 2-aminoadipate => lysine	1
Lysine biosynthesis, 2-oxoglutarate => 2-oxoadipate	2
Alanine, aspartate and glutamate metabolism	34
Cysteine and methionine metabolism	42
Tryptophan metabolism	11
Valine, leucine and isoleucine degradation	21
Arginine and proline metabolism	42
Glycine, serine and threonine metabolism	24
Valine, leucine and isoleucine biosynthesis	15
Phenylalanine, tyrosine and tryptophan biosynthesis	21
Phenylalanine metabolism	21
Lysine degradation	7
Lysine degradation, lysine => saccharopine =>	1

acetoacetyl-CoA	
Tyrosine metabolism	11
1.6 Metabolism of Other Amino Acids	93
Taurine and hypotaurine metabolism	11
Glutathione metabolism	56
beta-Alanine metabolism	16
Cyanoamino acid metabolism	10
1.7 Glycan Biosynthesis and Metabolism	74
N-Glycan biosynthesis	35
Other glycan degradation	10
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	17
Glycosphingolipid biosynthesis - globo series	3
Glycosaminoglycan degradation	8
Glycosphingolipid biosynthesis - ganglio series	1
1.8 Metabolism of Cofactors and Vitamins	112
Pantothenate and CoA biosynthesis	24
Thiamine metabolism	6
Porphyrin and chlorophyll metabolism	28
Lipoic acid metabolism	4
Ubiquinone and other terpenoid-quinone biosynthesis	9
One carbon pool by folate	11
Nicotinate and nicotinamide metabolism	7
Vitamin B6 metabolism	7
Riboflavin metabolism	5
Folate biosynthesis	6
Biotin metabolism	5
1.9 Metabolism of Terpenoids and Polyketides	46
Terpenoid backbone biosynthesis	15
Carotenoid biosynthesis	13
Zeatin biosynthesis	6
Limonene and pinene degradation	1
Diterpenoid biosynthesis	4
Brassinosteroid biosynthesis	7
1.10 Biosynthesis of Other Secondary Metabolites	33
Phenylpropanoid biosynthesis	21
Flavonoid biosynthesis	2
Stilbenoid, diarylheptanoid and gingerol biosynthesis	1
Tropane, piperidine and pyridine alkaloid biosynthesis	5
Anthocyanin biosynthesis	1
Isoquinoline alkaloid biosynthesis	3
1.11 Xenobiotics Biodegradation and Metabolism	3

Metabolism of xenobiotics by cytochrome P450	3
2. Genetic Information Processing	1821
2.1 Transcription	249
Basal transcription factors	21
RNA polymerase II, eukaryotes	3
RNA polymerase	15
RNA polymerase I, eukaryotes	2
RNA polymerase III, eukaryotes	1
RNA polymerase activity	1
Spliceosome, 35S U5-snRNP	22
Spliceosome	152
Spliceosome, U2-snRNP	10
Spliceosome, U4/U6.U5 tri-snRNP	15
Spliceosome, U1-snRNP	2
Spliceosome, Prp19/CDC5L complex	5
2.2 Translation	1107
Aminoacyl-tRNA biosynthesis, eukaryotes	28
Aminoacyl-tRNA biosynthesis	46
Ribosome	661
Ribosome, eukaryotes	272
Ribosome biogenesis in eukaryotes	97
Ribosome, bacteria	3
2.3 Folding, Sorting and Degradation	343
Proteasome, 19S regulatory particle (PA700)	17
Proteasome, 20S core particle	6
Proteasome	58
Protein export	47
RNA degradation	81
SNARE interactions in vesicular transport	31
Ubiquitin mediated proteolysis	103
2.4 Replication and Repair	122
Mismatch repair	17
Non-homologous end-joining	5
Nucleotide excision repair	35
Base excision repair	20
DNA replication	25
RNA-dependent DNA replication	1
Homologous recombination	19
3. Environmental Information Processing	46
3.1 Membrane Transport	21
ABC transporters	21
3.2 Signal Transduction	25
Phosphatidylinositol signaling system	25
4. Cellular Processes	170

4.1 Transport and Catabolism	170
Endocytosis	80
Peroxisome	70
Regulation of autophagy	20
5. Organismal Systems	113
5.1 Immune System	19
Natural killer cell mediated cytotoxicity	19
5.2 Environmental Adaptation	94
Circadian rhythm - plant	13
Plant-pathogen interaction	81

Table A7 Functional classification by NCBI Biosystems (AHSS)

1. Metabolism	415
1.1 Carbohydrate Metabolism	92
Butanoate metabolism	2
Fructose and mannose metabolism	6
Glycolysis / Gluconeogenesis	8
Starch and sucrose metabolism	21
Pyruvate metabolism	6
Amino sugar and nucleotide sugar metabolism	13
Citrate cycle (TCA cycle)	9
Propanoate metabolism	7
Galactose metabolism	5
Ascorbate and aldarate metabolism	2
Inositol phosphate metabolism	8
Pentose and glucuronate interconversions	4
Glyoxylate and dicarboxylate metabolism	1
1.2 Energy Metabolism	70
Oxidative phosphorylation	28
Carbon fixation in photosynthetic organisms	2
Nitrogen metabolism	9
Photosynthesis - antenna proteins	9
Photosynthesis - antenna proteins	9
Photosynthesis	8
Sulfur metabolism	5
1.3 Lipid Metabolism	44
Glycerophospholipid metabolism	13
Fatty acid biosynthesis, elongation	1
Fatty acid biosynthesis, elongation, endoplasmic reticulum	1

alpha-Linolenic acid metabolism	2
Glycerolipid metabolism	7
Fatty acid metabolism	4
Biosynthesis of unsaturated fatty acids	3
Ether lipid metabolism	1
Steroid biosynthesis	4
Linoleic acid metabolism	4
Sphingolipid metabolism	4
1.4 Nucleotide Metabolism	45
Pyrimidine metabolism	13
Purine metabolism	32
1.5 Amino Acid Metabolism	67
Lysine biosynthesis	5
Alanine, aspartate and glutamate metabolism	6
Cysteine and methionine metabolism	13
Tryptophan metabolism	5
Valine, leucine and isoleucine degradation	11
Arginine and proline metabolism	12
Glycine, serine and threonine metabolism	2
Valine, leucine and isoleucine biosynthesis	5
Phenylalanine, tyrosine and tryptophan biosynthesis	4
Lysine degradation	1
Tyrosine metabolism	3
1.6 Metabolism of Other Amino Acids	18
Taurine and hypotaurine metabolism	2
Glutathione metabolism	9
beta-Alanine metabolism	1
Cyanoamino acid metabolism	6
1.7 Glycan Biosynthesis and Metabolism	20
N-Glycan biosynthesis	6
Other glycan degradation	3
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	8
Glycosaminoglycan degradation	2
Glycosphingolipid biosynthesis - ganglio series	1
1.8 Metabolism of Cofactors and Vitamins	33
Pantothenate and CoA biosynthesis	6
Porphyrin and chlorophyll metabolism	9

Ubiquinone and other terpenoid-quinone biosynthesis	5
One carbon pool by folate	2
Nicotinate and nicotinamide metabolism	3
Vitamin B6 metabolism	1
Folate biosynthesis	3
Biotin metabolism	4
1.9 Metabolism of Terpenoids and Polyketides	16
Terpenoid backbone biosynthesis	8
Carotenoid biosynthesis	2
Zeatin biosynthesis	2
Limonene and pinene degradation	1
Diterpenoid biosynthesis	1
Brassinosteroid biosynthesis	2
1.10 Biosynthesis of Other Secondary Metabolites	10
Phenylpropanoid biosynthesis	2
Flavonoid biosynthesis	2
Stilbenoid, diarylheptanoid and gingerol biosynthesis	1
Flavone and flavonol biosynthesis	2
Tropane, piperidine and pyridine alkaloid biosynthesis	2
Isoquinoline alkaloid biosynthesis	1
2. Genetic Information Processing	359
2.1 Transcription	74
Basal transcription factors	7
RNA polymerase II, eukaryotes	2
RNA polymerase	8
Spliceosome	43
Spliceosome, U4/U6.U5 tri-snRNP	5
Spliceosome, U2-snRNP	3
Spliceosome, 35S U5-snRNP	4
Spliceosome, Prp19/CDC5L complex	2
2.2 Translation	189
Aminoacyl-tRNA biosynthesis	10
Aminoacyl-tRNA biosynthesis, eukaryotes	10
Ribosome	97
Ribosome biogenesis in eukaryotes	25
Ribosome, eukaryotes	47
2.3 Folding, Sorting and Degradation	70
Proteasome	8

Proteasome, 19S regulatory particle (PA700)	3
Protein export	8
RNA degradation	23
SNARE interactions in vesicular transport	7
Ubiquitin mediated proteolysis	21
2.4 Replication and Repair	26
Mismatch repair	3
Non-homologous end-joining	2
Nucleotide excision repair	5
Base excision repair	6
DNA replication	4
RNA-dependent DNA replication	1
Homologous recombination	5
3. Environmental Information Processing	16
3.1 Membrane Transport	9
ABC transporters	9
3.2 Signal Transduction	7
Phosphatidylinositol signaling system	7
4. Cellular Processes	38
4.1 Transport and Catabolism	38
Endocytosis	14
Peroxisome	21
Regulation of autophagy	3
5. Organismal Systems	32
5.1 Immune System	5
Natural killer cell mediated cytotoxicity	5
5.2 Environmental Adaptation	27
Circadian rhythm - plant	2
Plant-pathogen interaction	25

Table A8 Functional classification of SCFG transcriptome by NCBI Biosystems

1. Metabolism	356
1.1 Carbohydrate Metabolism	99
Butanoate metabolism	2
Fructose and mannose metabolism	11
Glycolysis / Gluconeogenesis	11

Starch and sucrose metabolism	24
Pyruvate metabolism	4
Amino sugar and nucleotide sugar metabolism	15
Citrate cycle (TCA cycle)	2
Propanoate metabolism	6
Galactose metabolism	2
Ascorbate and aldarate metabolism	1
Inositol phosphate metabolism	12
Pentose and glucuronate interconversions	5
Glyoxylate and dicarboxylate metabolism	4
1.2 Energy Metabolism	64
Oxidative phosphorylation	28
Carbon fixation in photosynthetic organisms	7
Nitrogen metabolism	6
Photosynthesis - antenna proteins	5
Photosynthesis - antenna proteins	5
Photosynthesis	10
Sulfur metabolism	3
1.3 Lipid Metabolism	42
Glycerophospholipid metabolism	14
18:1 Fatty acid biosynthesis	5
Fatty acid biosynthesis, elongation	1
Glycerolipid metabolism	4
Fatty acid metabolism	5
Biosynthesis of unsaturated fatty acids	4
Ether lipid metabolism	2
Steroid biosynthesis	3
Linoleic acid metabolism	1
Sphingolipid metabolism	3

1.4 Nucleotide Metabolism	29
Pyrimidine metabolism	17
Purine metabolism	10
Histidine metabolism	2
1.5 Amino Acid Metabolism	48
Lysine biosynthesis	4
Alanine, aspartate and glutamate metabolism	6
Cysteine and methionine metabolism	9
Tryptophan metabolism	3
Valine, leucine and isoleucine degradation	7
Arginine and proline metabolism	8
Glycine, serine and threonine metabolism	4
Phenylalanine, tyrosine and tryptophan biosynthesis	3
Lysine degradation	2
Tyrosine metabolism	2
1.6 Metabolism of Other Amino Acids	25
Taurine and hypotaurine metabolism	1
Glutathione metabolism	12
beta-Alanine metabolism	8
Cyanoamino acid metabolism	4
1.7 Glycan Biosynthesis and Metabolism	10
N-Glycan biosynthesis	4
Other glycan degradation	1
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	3
Glycosaminoglycan degradation	1
Glycosphingolipid biosynthesis - ganglio series	1
1.8 Metabolism of Cofactors and Vitamins	22
Pantothenate and CoA biosynthesis	3
Porphyrin and chlorophyll metabolism	9

Ubiquinone and other terpenoid-quinone biosynthesis	5
One carbon pool by folate	1
Folate biosynthesis	1
Biotin metabolism	3
1.9 Metabolism of Terpenoids and Polyketides	10
Terpenoid backbone biosynthesis	1
Carotenoid biosynthesis	6
Zeatin biosynthesis	2
Monoterpenoid biosynthesis	1
1.10 Biosynthesis of Other Secondary Metabolites	7
Phenylpropanoid biosynthesis	6
Tropane, piperidine and pyridine alkaloid biosynthesis	1
2. Genetic Information Processing	369
2.1 Transcription	71
Basal transcription factors	4
RNA polymerase II, eukaryotes	1
RNA polymerase III, eukaryotes	4
RNA polymerase	7
Spliceosome	46
Spliceosome, U4/U6.U5 tri-snRNP	1
Spliceosome, U2-snRNP	2
Spliceosome, 35S U5-snRNP	6
2.2 Translation	190
Aminoacyl-tRNA biosynthesis	16
Aminoacyl-tRNA biosynthesis, eukaryotes	10
Ribosome	111
Ribosome biogenesis in eukaryotes	16
Ribosome, eukaryotes	37
2.3 Folding, Sorting and Degradation	80

Proteasome	20
Proteasome, 20S core particle	9
Protein export	6
RNA degradation	12
SNARE interactions in vesicular transport	3
Ubiquitin mediated proteolysis	30
2.4 Replication and Repair	28
Mismatch repair	7
Non-homologous end-joining	1
Nucleotide excision repair	7
Base excision repair	4
DNA replication	3
negative regulation of DNA-dependent DNA replication	1
regulation of DNA-dependent DNA replication	1
Homologous recombination	4
3. Environmental Information Processing	10
3.1 Membrane Transport	4
ABC transporters	4
3.2 Signal Transduction	6
Phosphatidylinositol signaling system	6
4. Cellular Processes	30
4.1 Transport and Catabolism	30
Endocytosis	20
Peroxisome	13
5. Organismal Systems	15
5.1 Immune System	4
Natural killer cell mediated cytotoxicity	4
5.2 Environmental Adaptation	11
Circadian rhythm - plant	1

Plant-pathogen interaction	10
----------------------------	----

Table A9 Functional classification of SCTC transcriptome by NCBI Biosystems

1. Metabolism	268
1.1 Carbohydrate Metabolism	61
Butanoate metabolism	1
Fructose and mannose metabolism	9
Glycolysis / Gluconeogenesis	4
Starch and sucrose metabolism	11
Pyruvate metabolism	4
Amino sugar and nucleotide sugar metabolism	11
Citrate cycle (TCA cycle)	3
Propanoate metabolism	4
Galactose metabolism	2
Ascorbate and aldarate metabolism	4
Inositol phosphate metabolism	6
Pentose and glucuronate interconversions	1
Glyoxylate and dicarboxylate metabolism	1
1.2 Energy Metabolism	58
Oxidative phosphorylation	25
Carbon fixation in photosynthetic organisms	5
Nitrogen metabolism	6
Photosynthesis - antenna proteins	4
Photosynthesis - antenna proteins	4
Photosynthesis	11
Sulfur metabolism	3
1.3 Lipid Metabolism	27
Glycerophospholipid metabolism	5
Fatty acid biosynthesis	3

Glycerolipid metabolism	5
Fatty acid metabolism	5
Ether lipid metabolism	1
Steroid biosynthesis	2
Sphingolipid metabolism	5
Arachidonic acid metabolism	1
1.4 Nucleotide Metabolism	26
Pyrimidine metabolism	11
Purine metabolism	13
Histidine metabolism	2
1.5 Amino Acid Metabolism	36
Lysine biosynthesis	2
Alanine, aspartate and glutamate metabolism	2
Cysteine and methionine metabolism	7
Tryptophan metabolism	3
Valine, leucine and isoleucine degradation	7
Arginine and proline metabolism	5
Glycine, serine and threonine metabolism	4
Valine, leucine and isoleucine biosynthesis	1
Phenylalanine, tyrosine and tryptophan biosynthesis	3
Lysine degradation	1
Tyrosine metabolism	1
1.6 Metabolism of Other Amino Acids	10
Taurine and hypotaurine metabolism	2
Glutathione metabolism	5
beta-Alanine metabolism	2
Cyanoamino acid metabolism	1
1.7 Glycan Biosynthesis and Metabolism	8
N-Glycan biosynthesis	3

Other glycan degradation	1
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	2
Glycosaminoglycan degradation	1
Glycosphingolipid biosynthesis - ganglio series	1
1.8 Metabolism of Cofactors and Vitamins	26
Pantothenate and CoA biosynthesis	3
Porphyrin and chlorophyll metabolism	7
Lipoic acid metabolism	1
Ubiquinone and other terpenoid-quinone biosynthesis	4
One carbon pool by folate	3
Nicotinate and nicotinamide metabolism	3
Riboflavin metabolism	3
Folate biosynthesis	2
1.9 Metabolism of Terpenoids and Polyketides	7
Terpenoid backbone biosynthesis	1
Carotenoid biosynthesis	2
Zeatin biosynthesis	4
1.10 Biosynthesis of Other Secondary Metabolites	9
Phenylpropanoid biosynthesis	6
Flavonoid biosynthesis	1
Tropane, piperidine and pyridine alkaloid biosynthesis	1
Isoquinoline alkaloid biosynthesis	1
2. Genetic Information Processing	280
2.1 Transcription	61
Basal transcription factors	4
RNA polymerase III, eukaryotes	3
RNA polymerase	6
RNA polymerase II, eukaryotes	1
RNA polymerase I, eukaryotes	1

Spliceosome, U2-snRNP	2
Spliceosome, U4/U6.U5 tri-snRNP	2
Spliceosome	39
Spliceosome, 35S U5-snRNP	3
2.2 Translation	137
Aminoacyl-tRNA biosynthesis	8
Aminoacyl-tRNA biosynthesis, eukaryotes	8
Ribosome	70
Ribosome biogenesis in eukaryotes	21
Ribosome, eukaryotes	30
2.3 Folding, Sorting and Degradation	57
13 Proteasome	14
76 Proteasome, 20S core particle	5
298 Proteasome, 19S regulatory particle (PA700)	1
36 Protein export	7
148 RNA degradation	4
122 SNARE interactions in vesicular transport	7
35 Ubiquitin mediated proteolysis	19
2.4 Replication and Repair	25
Mismatch repair	6
Non-homologous end-joining	1
Nucleotide excision repair	7
Base excision repair	2
DNA replication	5
Homologous recombination	4
3. Environmental Information Processing	7
3.1 Membrane Transport	4
ABC transporters	4
3.2 Signal Transduction	3

Phosphatidylinositol signaling system	3
4. Cellular Processes	20
4.1 Transport and Catabolism	20
Endocytosis	8
Peroxisome	11
Regulation of autophagy	1
5. Organismal Systems	13
5.1 Immune System	3
Natural killer cell mediated cytotoxicity	3
5.2 Environmental Adaptation	10
Plant-pathogen interaction	10

LIST OF PUBLICATIONS

Research Publications

- [1] **Tarun Pal**, Varun Jaiswal, Rajinder S. Chauhan (2016). DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. *Computers in Biology and Medicine*, 78, 42-48. **(Impact factor: 1.521) Scopus Indexed**
- [2] **Tarun Pal**, Nikhil Malhotra, Sree Krishna Chanumolu, Rajinder S. Chauhan (2015). Next-generation sequencing (NGS) transcriptomes reveal association of multiple genes and pathways contributing to secondary metabolites accumulation in tuberous roots of *Aconitum heterophyllum* Wall. *Planta*, 242(1), 239-258. **(Impact factor: 3.376) Scopus Indexed**
- [3] **Tarun Pal**, Jibesh Kumar Padhan, Pawan Kumar, Hemant Sood, Rajinder S. Chauhan (2016). Comparative transcriptomics uncovers differences in photoautotrophic versus photoheterotrophic modes of nutrition in relation to secondary metabolites biosynthesis in *Swertia chirayita*. *Molecular Biology Reports* **(Impact factor: 1.698) (Revision Submitted) Scopus Indexed**
- [4] Varun Gupta, Nikhil Malhotra, **Tarun Pal**, Rajinder S. Chauhan (2016). Molecular dissection of pathway components unravels atisine biosynthesis in a non-toxic *Aconitum* species, *A. heterophyllum* Wall. *3Biotech* 6(1), 1-10. **(Impact factor: 0.992) Scopus Indexed**

Conference Publications

- [1] **Tarun Pal**, Sree Krishna Chanumolu, Varun Jaiswal, Rajinder S. Chauhan (2013). Computational pipeline for analysis of ngs transcriptome data sets in medicinal herbs in *7th Annual Convention of ABAP & International Conference on Plant Biotechnology, Molecular Medicine & Human Health*, University of Delhi, South Campus, October 18-20, 2013. pp 193
- [2] **Tarun Pal**, Varun Jaiswal, Sree Krishna Chanumolu, Rajinder S. Chauhan (2014). Computational mining of *Aconitum heterophyllum* transcriptomes for transcription

factors controlling atisine biosynthesis in *National Symposium on Advances in Biotechnology for Crop Improvement*, Eternal University, Baru Sahib (H.P.), July 12, 2014. pp 37-38