Genome-wide identification and mathematical model development for prediction of operon like gene clusters in Rice genome

A Thesis

Submitted in partial fulfillment of the requirements for the

award of the degree of

DOCTOR OF PHILOSOPHY

in

Biotechnology

By

Himanshu Singh

Registration No: 41400180

Supervised By

Dr. Vikas Kaushik

LOVELY PROFESSIONAL UNIVERSITY

PUNJAB

2021

DECLARATION

I hereby declare that the thesis entitled, "***Genome-wide identification and mathematical model development for prediction of operon like gene clusters in Rice genome***" submitted for Ph.D. Biotechnology degree to Department of Biotechnology, Lovely Professional University is entirely original work and all ideas and references have been duly acknowledged. The research work has not been formed the basis for the award of any other degree.

.

Himanshu Singh

Date: 16th August, 2021

Himanshu Singh

Registration No: 41400180

CERTIFICATE

This is to certify that **Mr. Himanshu Singh** has completed the Ph. D in Biotechnology titled "***Genome-wide identification and mathematical model development for prediction of operon like gene clusters in Rice genome***" under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of this thesis has ever been submitted for any other degree or diploma.

The thesis is fit for the submission for the partial fulfilment of the condition for the award of degree of Ph.D. in Biotechnology.

Advisor Name & Signature: Vikas Kaushik

Date: 16th August, 2021

Acknowledgement

My most sincere gratitude goes out to my advisor Dr. Vikas Kaushik, Associate professor, department of Bioinformatics, School of Bio-engineering and Biosciences, Lovely Professional University, Punjab, India. He has provided me with intellectual and editorial support before and during the writing of this thesis. His guidance throughout my PhD has been invaluable to me. Interaction with Dr. GPS Raghava, IMTECH Chandigarh has shaped both this thesis and my understanding of the machine learning area. In particular, I have been lucky to get guidance from Dr. Atul Kumar Upadhyay, Assistant professor as well from Thapar University, Punjab.

I am thankful to Lovely Professional University; Punjab for providing me all support which has made completing my research studies that much easier. I acknowledge scholarly work support by Prof. (DR.) Neeta Raj Sharma, Head of School of Bio-engineering and Biosciences, Lovely Professional University, Punjab, India. I also gratefully acknowledge help rendered by Mr. Amit Joshi, Ph.D. Scholar of LPU in collection of related literatures and in software-based management of references. I also gratefully acknowledge Mr. Bhupendra, PhD Scholar from Lovely Professional University for his immense support in SVM data analysis.

Thank you to all the members of the School of Bioengineering and Biosciences LPU, especially Dr. Anjuvan Singh for providing a lively and stimulating atmosphere. My wife Sweta Singh help has also been invaluable in many ways at times scientifically and psyclogically.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

SMO	Sequential Minimal Optimization
MIBiG	Minimum Information about a Biosynthetic Gene cluster
BGCs	Biosynthetic Gene Clusters
UV	Ultra Violet
antiSMASH	Antibiotics & Secondary Metabolite Analysis Shell
SMURF	Secondary Metabolite Unique Regions Finder
DIBOA	2,4-dihydroxy-1,4-benzoxazin-3-one
DIMBOA	2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one
plantiSMASH	Plant Secondary Metabolite Analysis Shell
CYP72	Cytochrome P450 enzymes
GAME7	Glycoalkaloid metabolism 7
Mu	Mutator
bx1	Benzoxazineless
BAC	Bacterial Artificial Chromosome
3D	Three dimensional
QMEANDisCo	Qualitative Model Energy ANalysisDIStanceCOnstraint
BLAST	Basic Local Alignment Search Tool
PSI-BLAST	Position-Specific Iterative BLAST
I-TASSER	Iterative Threading ASSEmbly Refinement
GUI	Graphical user interface
RMSD	Root Mean Square Deviation
DNA	Deoxyribonucleic acid
GBK	GenBank file format
EMBL	European Molecular Biology Laboratory
GFF	General feature format
FASTA	Fast-all
SOFT	Simple Omnibus Format in Text
CSV	Comma-separated values
NCBI	National Center for Biotechnology Information
UniprotKB	Universal Protein Resource Knowledgebase
GRAVY	Grand Average of Hydropathicity
CASP13	Critical Assessment of protein Structure Prediction 13
CAMEO	Continuous Automated Model Evaluation
LOMETS	Local Meta-Threading Server
BLASTP	BLAST-Protein
SVMLib	Support Vector Machines Library
ANN	Artificial Neural Network
IBk	Instance Based Learner
2OG-FeII_Oxy	2-oxoglutarate (2OG) and Fe (II)-dependent oxygenase

DIOX_N	Non-haem dioxygenase in morphine synthesis N-terminal
UDPGT	Uridine 5'-diphospho-glucuronosyltransferase
Chal_sti_synt_C	Chalcone and stilbene synthases, C-terminal domain
Chal_sti_synt_N	Chalcone and stilbene synthases, N-terminal domain
Aminotran	Aminotranaferase
Methyltransf	Methyltransferase
Glycos_transf	Glycosyltransferase
Terpene_synth	Terpene synthase
Cellulose_synt	Cellulose Synthase
ADH	Alcohol dehydrogenase
Cu_amine_oxid	Copper amine oxidase
AMP-binding	Adenosine monophosphate binding
Acetyltransf	Acetyltransferase
Str_synth	Strictosidine synthase
SQHop_cyclase_C	Squalene-hopene cyclase C-terminal domain
SQHop_cyclase_N	Squalene-hopene cyclase N-terminal domain
SQS_PSY	Squalene/phytoene synthase
TPMT	Thiopurine methyltransferase
PR-10	10 of plant pathogenesis-related protein
rbcS1	Ribulose bisphosphate carboxylase
GluA-3	Ionotropic glutamate receptors
CDKB1	Cyclin-dependent kinase
UGE	UDP-glucose/galactose-4-epimerase
CHS8	Chalcone synthase
XSP1	xylem serine peptidase 1
PC	Pyruvate carboxylase
GLN2	Glutamine synthetase
PDF1	Protodermal factor 1
C4H	Cinnamate-4-hydroxylase
CHS8	Chitin synthase 8
TM	Template Modelling score
C-Score	Confidence score
ACE	Atomic Contact Energy
MW	Molecular Weight
PI	Isoelectric Point
TMindex	Melting Temperature index
TP	True positive
FP	False positive
ROC	Receiver Operating Characteristic
MCC	Mathew’s correlation coefficient

LIST OF WEB RESOURCES

1. **BLAST:** (https://blast.ncbi.nlm.nih.gov/Blast.cgi)
2. **PSI_BLAST:**(https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROGRAM=blastp&BLAST_PROGRAMS=psiBlast)
3. **CASCADE BLAST:** (http://crick.mbu.iisc.ernet.in/~CASCADE/CascadeBlast.html.)
4. **PlantiSMASH:** (http://plantismash.secondarymetabolites.org/about.html)
5. **Genomic Viewer:** (https://www.ncbi.nlm.nih.gov/genome/gdv/)
6. **Weka 3_Tool:** (https://www.cs.waikato.ac.nz/ml/weka/)
7. **ProtParam by ExPASy:** (https://web.expasy.org/protparam/)
8. **Protein-Solubility Server:** (https://protein-sol.manchester.ac.uk/)
9. **PepCalc:** (https://pepcalc.com/)
10. **UniprotKB:** (https://www.uniprot.org/)
11. **Swiss-Model:** (https://swissmodel.expasy.org/)
12. **trRosetta web server:** (https://yanglab.nankai.edu.cn/trRosetta/)
13. **I-TASSER:** (https://zhanglab.ccmb.med.umich.edu/I-TASSER/)
14. **PatchDock server:** (https://bioinfo3d.cs.tau.ac.il/PatchDock/php.php)

ABSTRACT

Organized microbial gene expression under the concept of operons is a well-established concept. Operon-Esque gene arrangements in the plant kingdom have been brought to the limelight with the help of recent developments in genetics, biochemistry, and bioinformatics. Plant Gene clusters contains signature and tailoring genes. Signature genes are responsible for forming backbone of the structure of the molecule. We aim to explore the interactions between various gene products (enzymes) from the gene clusters of the 12 chromosomes in *Oryza sativa Japonica* and *Oryza sativa Indica*. Sequence information of all the reviewed 'signature' and 'tailoring' genes was retrieved. A simple search with the name of the gene product and the species name was performed on UniProtKB. The 3-dimensional structures of these proteins were predicted using various bioinformatics tools. The FASTA sequences of the selected gene products (enzymes/proteins) were extracted from UniProtKB which was in turn used as the input for Swiss-Model, TrRosetta and I-TASSER servers to predict the 3D structures. The interactions among various predicted structures were ascertained by molecular docking techniques. The obtained 3D structures were then loaded to the PatchDock server to conduct molecular docking and interaction analysis. The hypothesis that the various signature gene products and the corresponding tailoring gene products of the gene clusters of *Oryza sativa* undergo protein-protein interaction was substantiated by the reliable molecular docking results that reveal perfect interaction. Phylogenetic trees were developed by utilizing the BLAST TREE widget on pBLAST. The dendrogram generated for each signature gene product displays the evolutionary relationship with similar proteins from other species or the predicted/hypothetical proteins suggested by the algorithm. We want to create a universal cluster prediction algorithm based on the distribution of selected specific physicochemical properties, using terpene synthase gene clusters as a guide and classifying genes into terpene synthases and non-terpene synthases. In the WEKA tool, random forest and sequential minimal optimization (SMO) classifiers were used to construct the machine learning model. A set of ten physicochemical properties were selected and their values were predicted for each of the protein molecules (terpene synthases and non terpene synthases). Employing the random forest and SMO classifiers, we were able to obtain significantly promising accuracy of over 90 percent with 66 percent percentage split testing. Accurate prediction of BGCs in the plants, especially in the major food crops like rice, wheat and corn revolutionize farming and nutrition for the better future.

CHAPTER 1:
INTRODUCTION

INTRODUCTION

1.1 OPERONS

Operons are organized sets of genes found in microbes that work in a coordinated fashion **[Zheng et al., 2002; Rocha, 2008; Koonin, 2009]**. Each operon's genes are coordinated and regulated by the same promoter, which produces a single polycistronic message. *E. coli* Lac operon was the first to be discovered and it's necessary for lactose to be used as a carbon and energy source. **[Jacob, F., & Monod, J., 1961; Jacob et al., 2005].** Historically, bacteria were thought to be the only carriers of clustered genes. The fact that operons coordinate over half of the bacterial genes bolstered this belief. MIBiG is a specific chromosomal cluster archive (A biosynthetic gene cluster's minimal information), records 1,221 clusters of bacterial genes with a variety of chemical applications in medical, agricultural, and industrial applications **[Cimermancic et al., 2014; Medema., 2015]**. Recent Advancements in the fields of genetics, biochemistry, and bioinformatics have led to a paradigm-shifting discovery in the plant kingdom wherein they were observed to have operon-like arrangements of genes within their genome. This phenomenon albeit not as established as the microbial operons is called Biosynthetic Gene Clusters (BGCs). Genes of a BGC work together under the control of the same promoter for the production of the respective proteins **[Qin et al., 2010]**. BGCs are found to be comprised of two types of genes, namely, Tailoring genes and signature genes. Tailoring genes produce tailoring enzymes and signature genes produce the bigger, more complex signature enzymes. Tailoring gene products, which are produced first, trigger a cascade of catalytic reactions that culminate in the synthesis of the signature enzyme **[Ghosh, Ali, &Gantait, 2016].**

1.2 METABOLITES

The improvements and accelerated rates of discovery of new techniques to uncover gene clusters have made advancements in plant biology and natural product discovery easier and more convenient than they have ever been. Improved farming (allelopathic interactions), drug discovery, better nutrition, and synthetic biology are a few of the areas that would reap the benefits of BGC discovery **[Nützmann& Osbourn, 2014]**.

Metabolites are chemicals generated by organisms as a direct or indirect product of their system's various pathways. Plants create these unique metabolites to aid in their growth and development by interacting with their surroundings and allowing other mechanisms such as insect and UV defense **[Kautsar et al., 2017]**. The majority of plant metabolite reservoirs are

shrouded in mystery, making synthetic methods difficult to meet the demands for successful and inexpensive pharmaceutical, agricultural, and industrial chemicals. Some of the most common metabolites are terpenes, alkaloids, and cyanogenic glycosides **[Nützmann& Osbourn, 2014]**.

1.3 PLANT GENE CLUSTERS

Genetic and biochemistry research, as well as serendipity, have discovered metabolite gene clusters in plants. Cluster discovery relies heavily on the delineation of plant genomic sequences **[Field et al, 2011; Field, & Osbourn,2008; Osbourn et al., 2012; Castillo, Kolesnikova, &Matsuda, 2013]**, This would allow bioinformatics workflows similar to those used for microorganisms (antiSMASH, SMURF, and ClusterMine360) to develop and be produced more quickly. **[Blin et al., 2013; Khaldi et al., 2010; Conway, &Boddy, 2012]**. Plant genomes have always been difficult to study due to their sheer scale. However, with these cutting-edge techniques, massive gymnosperm genome projects are now feasible **[Mackay et al., 2012; Nystedt et al., 2013]**. Gene duplication and neofunctionalization have been identified as the key drivers of gene clustering **[Chu, Wegel, & Osbourn, 2011]**. Bx1-Bx5, the proximal Penta gene organization seen in *Zea mays*, was created to code for enzymes involved in the creation of 2,4-dihydroxy-1,4-benzoxazin-3-one, a cyclic hydroxamic acid (DIBOA). Bx6-Bx9 genes were found to be linked to Bx1-Bx5, suggesting that they help in the conversion of DIBOA to 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA) and subsequent glycosylation **[Von Rad et al., 2001; Sakamoto et al., 2004; Shimura t al., 2007; Jonczyk et al., 2008]**. The oat (Avna spp.) avenacin cluster and the rice (*Oryza sativa*) phytocassane cluster were identified seven years later, contributing to the increasing number of gene clusters **[Qi et al., 2004; Wildermanen et al., 2004]**. By 2012, nine metabolite plant gene clusters had been discovered **[Field et al.,2011; Field, & Osbourn, 2008; Shimura et al., 2007; Takos et al., 2011]**, and four more were discovered in the following four years **[Winzer et al., 2012; Matsuba et al., 2013; Krokida et al., 2013; Itkin et al., 2013]**. Because of tandem duplications, some BGCs are similar, such as repetitive leucine-rich genes in plants that aid disease resistance and self-incompatibility/heterostyly deciders in plants **[Field, & Osbourn 2008; Li et al., 2016]**. While horizontal transfer from microbes may seem to be a logical assumption for the formation of BGCs in plants, this is not the case; rather, the development of BGCs in plants is aided by the transfer of genes somewhere else in the genome via replication/duplication, neofunctionalization, and other yet undiscovered

mechanisms.**[Boutanaev et al., 2015; Dutartre, Hilliou, &Feyereisen,2012; Osbourn, 2010]**.The presence of frequency gene clusters in the plant kingdom has yet to be completely determined and, in order to do so, the entire plant kingdom must be sequenced, which is a difficult task. Focusing research on Corn, rice, and other major staple food plants, incorporating molecular biology, biochemistry, and reverse genetics approaches with high throughput genomic analysis, is, frankly, the sensible thing to do **[Von Rad et al., 2001; Sakamoto et al., 2004; Shimura t al., 2007; Jonczyk et al., 2008; Wildermanen et al., 2004; Takos et al., 2011; Swaminathan et al., 2009]**. Genes that are often elicitor inducible have been found to be useful in the fast study of BGCs. For the oat avenacin community, for example, screening for fluorescence loss at the roots was critical when using a forward genetic approach **[Papadopoulou et al., 1999; Qi et al., 2006]**. The presence of a cluster in *Arabidopsis thaliana* was determined by measuring the activity of the triterpene synthase (oxidosqualene cyclase) gene **[Field, & Osbourn,2008]**. A BGC's transcriptional response can be expressed in three ways: coexpression, coregulation, and coordinated expression. These processes are influenced by an individual's developmental or environmental circumstances **[Nützmann, Scazzocchio, &Osbourn, 2018].**

1.4 RICE *(ORYZA SATIVA)*

After wheat, rice is the world's most frequently eaten food crop on the planet. A member of the Poaceae and genus Oryza, rice (Oryza *sativa)* has over 20 wild varieties that are globally cultivated out of which *Oryza sativa* is the widely cultivated one and *Oryza glaberrima* of West African origin is one of the oldest subspecies that have been around for 3500 years. With n=12, rice is usually diploid or triploid. The Asian rice or Japanese rice (*Oryza* sativa *Japonica*) was the first food crop to have its whole genome sequenced. Rice is loaded with niacin, zinc, and rice proteins which have an astounding 88 percent which makes it the richest biological protein. All these desirable qualities have made rice the staple food of over 3 billion people around the world and a meaningful candidate for most of the BGC studies **[Yi, Sze, &Thon, 2007]**.

The Indica rice ecotype is generally cultivated in tropical and subtropical rice-growing regions with low latitudes and altitudes, whereas the japonica rice ecotype is primarily grown in temperate rice-growing areas of high latitudes and altitudes. Japanese cultivars are known for having lesser production potential than Indica types. Japonica rice may be grown in hilly locations with high altitudes in several low-latitude rice-growing regions, such as China's

Yunnan and Guizhou provinces, Laos, Myanmar, and Vietnam, as well as many other Southeast Asian nations. Because of their adaptation to distinct ecological circumstances, indica and japonica rice exhibit substantial differences in morphological (e.g., plant height and pubescence), agronomical characteristics (e.g., length/width ratio and grain persistence), and physiological–biochemical properties (e.g., winter hardiness and starch types in grains).Some rice experts split Indica and japonica rice into two subspecies, *Oryza sativa* subsp. *Indica* Kato and *Oryza sativa* subsp. *japonica* Kato, due to significant differences in many respects **[Vaughan, Lu, &Tomooka, 2008; Wang, &Li, 1997].** Indica and Japonica rice types have quite different genetic connections. There will be substantial genetic recombination and variety as a result of the inter-subspecies hybridization of Indica and japonica rice, offering rice breeders greater possibilities for selecting optimum variation types for rice development. The Indica–Japonica hybrid rice has a lot of potential in rice cultivation if the difficulties of inter-subspecies sterility can be solved **[Khush, 2001]**. Over the lengthy history of rice domestication, the Indica and Japonica rice varieties have diversified in morphological features, agronomic attributes, physiological and biochemical aspects, yield, quality, and stress tolerance. However, the proteins and genes that cause these variations, as well as their functions in these two rice varieties, are still unknown. Furthermore, because of the large geographical overlap in adaptability between the two types, distinguishing between Indica and Japonica rice is challenging **[Peng et al., 2004].**

1.5 BGC PREDICTION

Accurate detection of these BGCs would be a tremendous catalyst that would accelerate developments in the field of agriculture. We are aiming to do just that by developing an algorithm/tool for identifying and classifying BGCs with better performance statistics than already available tools. We studied the interactions between the enzymes (gene products) formed by the gene clusters within the 12 total chromosomes of *Oryza* sativa *Japonica* and *Oryza sativa Indica* in the first part of the study. The sequence information for the genes that were checked and annotated was retrieved. To make predictions, Bioinformatic web servers were utilized for three-dimensional models of these proteins, which were then molecularly docked together. This was done to test the hypothesis that the interaction between the proposed tailoring and signature gene products results in overall BGC expression. The dendrograms generated for selected gene products provided a better understanding of these enzymes' evolutionary pathways **[Chu, Wegel, &Osbourn, 2011]**.

The second part of the project entailed modeling the algorithm based on the relevant data sets that are unique and reliable for classifying input data. The popular machine learning tool, WEKA was employed to develop the models using the physicochemical properties of proteins as the data set. Terpene synthase was picked to be the candidate while non-terpene synthases were selected as the negative control. A total of 159 proteins, (terpene synthases and non-terpene synthases) were classified on WEKA using the Sequential minimal optimization (SMO) and random forest classifiers at a 66 percent split value. The outputs obtained showed significantly high accuracy of over 90 percent, which was a promising sign **[Ho, 1995; Ho, 1998; Platt, 1998].**

1.6 APPLICATION

When it comes to this field of study, the practical challenge is determining how the knowledge gleaned from studying various genetic variants will support agricultural and industrial use, pestand diseasetolerance, improved-nutritional values, higher levels of high-value items. Clustering genetic knowledge may be used in genetic engineering to help deal with genetic defects linked to undesirable traits (such as bitterness). Allelopathy is another field where BGC discovery and associated technologies may be useful. It's a process in which plants control the rhizosphere by releasing a group of chemicals known as allelochemicals, which can affect the growth of nearby plants **[Albuquerque et al., 2011]**. Allelopathic interactions are usually negative, and they're used to manage weeds naturally **[Khanh et al., 2005; Cheng, & Cheng, 2015; Guo et al., 2017; Boycheva et al., 2014; Xu et al., 2012]**.

By the use of high-throughput screening techniques. The area of BGC exploration has seen exponential development, uncovering new pathways, enzymes, and chemistries in the plant kingdom by combining systematic genome mining and functional analysis of candidate clusters with artificial intelligence and machine learning. With this initiative, we aim to contribute to the revolutionary movement.

CHAPTER 2:
REVIEW OF LITERATURE

REVIEW OF LITERATURE

2.1 ORGANIZATION OF GENE CLUSTERS

Within a gene cluster, one gene usually codes for a signature enzyme that defines a certain metabolite scaffold, while a variable number of additional genes code for tailoring enzymes [**Mugford et al., 2013; Osbourn, 2010]**. Gene duplication and neofunctionalization of primary metabolism genes tend to result in signature genes within a cluster, which may be direct or indirect **[Chu, Wegel, & Osbourn, 2011]**. The tailoring genes are also recruited by this signature gene **[Field et al; 2011; Dutartre, Hilliou, &Feyereisen, 2012]**. These pathways are being illuminated by comparative genomics **[Field, 2011; Takos et al., 2011; Matsuba et al., 2013; Dutartre, Hilliou, &Feyereisen, 2012]**. The secondary metabolites are made from simple building blocks that are readily available in plants, including amino acids, fatty acids and sugars. **[F. pengxiang et al., 2020]**

The gene coding for the first step may not appear to be closely linked to other gene genes in certain cases (The CYP72 gene GAME7, for example, is 8 Mb distant from other coupled genes on the same chromosome and may be involved in the initial phase of steroidal glycoalkaloids production in tomatoes) **[Takos et al., 2011; Gao et al., 2012].**

2.2 DISCOVERY

Following the discovery of a cluster, a cognate metabolic pathway must be identified. For product recognition, heterologous expression of the putative gene in bacteria or yeast may be used. The same findings may be obtained via transient expression in Nicotiana benthamiana. The expression mechanism is determined by the function of the substance, such as the signature enzyme. For example, yeast strain GIL77 accumulates 2,3-oxidosqualene, making it perfect for oxidosqualene cyclase functional testing **[Gao et al., 2012]**. The use of T-DNA insertion mutants in Arabidopsis to establish that the thalianol and marneral pathways need the THAS, THAH, and MRO genes **[Field et al., 2011]** is a recent example of a reverse genetic approach to cluster identification. Other clusters (such as the avenacin cluster in oats and the thalianol cluster in A. thaliana) are compact and don't have a lot of genes, but the cyanogenic glucoside cluster in Lotus japonicus has a lot of genes that don't have anything to do with secondary metabolism **[Field, & Osbourn, 2008]**. Chemical genetic methods can be used to reduce the number of potential genes for a gene cluster pathway. The CYP inhibitor uniconazole-P, for example, was utilized to look into the involvement of CYPs in the

momilactone pathway **[Shimura et al., 2007]**, Bx6 in the maize DIMBOA pathway was discovered using the 2-oxoglutarate-dependent dioxygenase inhibitor Prohexadion-Ca.

Genetic analysis was utilized to find a secondary metabolic component for the manufacture of defense compounds in corn and oats. The cyclic hydroxamic acid DIBOA and its methylated counterpart DIMBOA give pest and microbial resistance in maize and grasses. DIBOA synthesis is stopped when max bx1 (benzoxazineless) is mutated. Bx1 was genetically modified using the Mutator (Mu) transposon maize marking technique, and heterologous expression in *Escherichia coli* revealed that it incorporates tryptophan synthase, a homologous indole (DIBOA DIMBOA precursor) rather than tryptophan **[Frey et al., 1997]**. Another gene product, Bx3, may also be cloned using the Mu method. On maize chromosome 4, Bx3 is one of four closely related cytochrome P450 (CYP) genes (Bx2-5) located near Bx1**[Frey et al., 1995]**. According to Bx2-5 expression in yeast, these four CYPs support sequential stages in indole conversion in DIBOA. Bx1-5 was discovered and characterized using a mix of genomic and chemical research. DIBOA is produced by converting indole-3-glycerol phosphate to indole-3-glycerol phosphate, these five genes, which are part of the gene pool, are needed and necessary. Additional genes implicated in this process have been discovered in other research, such as those results in the transformation of DIBOA to DIMBOA as well as the glucosylation of these molecules **[Jonczyk et al., 2008].** The avenacin series of antimicrobial triterpene glycosides (saponins) derived from oat roots is a second example **[Qi et al., 2004]**. A forward screen approach identified diploid oats (*Avena strigosa*) mutant genes that were unable to generate avenacins, which led to the discovery of this set of genes (known as saponin deficient sad mutant). According to earlier genetic study, six of the Sad loci discovered by genetic mutation were connected to genes and mapped to a 3.6-cM regional map of the diploid oat genome's D relation community. A single gene (Sad4, which is necessary for glucosylation) is not linked or connected **[Chaudhury et al., 2011]**. Sad1 is the first of these genes, and it produces b-amyrin synthase, an enzyme that promotes the contribution to avenacin synthesis in the first step **[Haralampidis et al., 2001]**. Since Sad1 is generated and characterized by the same product with active activity in yeast, we created a diploid oat chromosomal (BAC) library (*A. strigosa*). Itallowed us to find Sad1-containing BAC clone and build a 400-kb BAC contig around this species' DNA. The BAC sequence revealed four other genes in the gene cluster near Sad1, all of which show to impart crucial role in avenacin biosynthesis **[Chu, Wegel, & Osbourn, 2011]**.

It is necessary to demonstrate that genetic products function together in order to demonstrate that a genetic community forms a cluster of genes. This can be accomplished in two ways: I using genetic modification techniques such as genetic knockouts and knockdowns on the plant from which it originated; and/or (ii) using genetic techniques such as genetic knockouts and knockdowns on the plant from which it originated. These and other related mechanisms are also being investigated for the identification of non-clustered metabolic pathways **[Dixon et al., 2006]**.

Genetic variations and gene clusters in plants have been studied using a variety of expression methods. Many distinct enzymes, including glycosyl transferases, oxoglutarate-dependent dioxygenases, and methyltransferases, have been successfully produced by expressing cluster products in *E. Coli*, followed by protein purification and substrate molecule testing **[Jonczyk et al., 2008]**. Reverse genetic methods, as detailed in the DIBOA/ DIMBOA maize metabolite production path, require knowledge on genes that may be targeted for mutation. Knocking out a gene causes an increase in the output of the enzyme in question, as well as an accumulation of the enzyme's substrate and the loss of the metabolic pathway's final product. To fully understand metabolism, new metabolites synthesized by recombinant enzymes in vitro assays or by reverse genetics must be identified. The dynamic existence of plant extracts makes metabolite detection difficult. In addition to control samples, chromatographic separation accompanied by mass spectrometry is commonly used for identification (e.g., wild plant species). Mass spectrometry methods may be used indefinitely to collect structural information. Tandem mass spectrometry techniques with precise weight, for example, are useful for detecting multiple changes associated with hydroxylation, acylation, and glycosylation **[McCallum et al., 2000; Prasad et al., 2011]**. Biosynthetic pathways of the numerous metabolites can be determined by characterization of genes, interactions and gene products **[Bharadwaj et al., 2021].**

2.3 TOOLS AND TECHNIQUES

The three-dimensional structure of proteins exposes important data about their function at the cellular level, as well as a wide range of medicinal applications. Many biological functions are built on the foundation of protein structures. A thorough grasp of biological processes, as well as how protein structures and networks work, is required, and how we can quantify them requires a thorough explanation of their interactions and the complete structure of their counterparts **[Nim et al., 2016]**.

When 3D structures of binding partners are available or can be drawn in a trustworthy manner, drawing methods may be utilized to generate a three-dimensional complex model based on the geometric and physicochemical cohesion of interacting molecules **[Kurkcuoglu et al., 2018]**.

Homology modeling is a prominent structural biology approach that has had a significant influence on narrowing the gap between known protein sequences and empirically confirmed structures. The modeling homology process is streamlined and guided by fully automated processes and servers, allowing users without advanced programming skills to create accurate protein models and access modeling, vision, and translation results with ease. The SWISS-MODEL server, which was developed 25 years ago and has been continuously updated, was the first to discover automatic modeling. Its utility has recently been expanded to modeling homomeric and heteromeric properties. From the amino acid sequence of interacting proteins, a homology model guides stoichiometry and its all-encompassing complex structure. Two additional important improvements are the adoption of a new modeling engine, ProMod3, and the installation of a new local quality measurement tool, QMEANDisCo. SWISS-MODEL may be downloaded for free at https://swissmodel.expasy.org **[Waterhouse et al., 2018].**

Threading is a bioinformatics technique for identifying template proteins using standardized knowledge from a structure that has the same structure or structural structure as the sequence protein in question. With BLAST, the query sequence is compared to the unnecessary sequence database (PSI-BLAST) **[Altschul et al., 1997]** in the first step of I-TASSER to distinguish evolving relatives. A duplicate assembly server (TASSER) is an integrated system that automates protein synthesis and output prediction using the sequence-to-function paradigm. Starting with the amino acid sequence, I-TASSER creates three-dimensional (3D) atomic models by aligning numerous threads and simulating a structural assembly. By accurately matching 3D models with other known proteins, the function of a protein may subsequently be identified. Normal server responses produce estimates of full secondary and tertiary structures, replete with active ligand binding sites, Enzyme Commission numbers, and Gene Ontology terms. The precision of the predictions is estimated using the model's confidence measurement. This protocol offers new ideas and recommendations for developing cutting-edge online protein systems and predictions. http://zhanglab.ccmb.med.umich.edu/TASSER **[Roy, Kucukural, & Zhang, ,2010]** is the address for the server.

One of the most difficult problems in structural biology is predicting protein-protein and small molecule-protein interactions automatically. More precise predictions will help many biological studies, both academic and industrial. Docking is a technique for determining the optimal combination of two interlocked molecules given a description of each molecule's structure.

When it comes to protein-protein docking, the correct prediction will disclose the bulk of the remaining contacts in the targeted interaction. Over the last thirty years, several docking algorithms **[Gray et al., 2003]** have been created. At the moment, however, only a few algorithms are offered as a free web service (ClusPro server). Algorithms vary significantly in how they look for and evaluate fixed structures in the six-dimensional transformation space they use. The majority of these algorithms were able to withstand extensive testing. PatchDock **[Connolly, 1983]** is a fast algorithm for docking small proteins with ligands and proteins with other proteins. Predicting the structure of protein-protein and protein-molecule complexes is easier with the PatchDock process. http://bioinfo3d.cs.tau.ac.il/) provides access to the facilities. A variety of localized features of static molecules with compatible features are guaranteed to be included in the robust GUI. The PatchDock method splits the surface representation of the Connolly dot of molecules into concave, convex, and flat patches. Candidates are then examined for geometrical fit and atomic energy once the relevant patches are matched. Finally, a variant of RMSD (root mean square deviation) is utilized to eliminate undesired candidate solutions **[Duhovny et al., 2005] [Frey., et al 1995]**. Phylogenetic trees represent the relationship between protein sequences. The tree's topology indicates how the sequence should be organized, and the length of the branch gives an indication of the true stages of evolution. The precision of a tree is naturally influenced by the sequence of events. The reference tree's phylogenetic species are used as a guide tree for orthologous protein clustering. Intramatrix junctions in orthologous protein matrices can be extracted using the phylogenetic tree as a reference, where these interactions are interpreted as intermediate distances between orthologous ancestral proteins **[Craig, & Liao,2007].**

In all related matters, the nodes reflect the collective ancestor. The diversity of evolutionary history indicates Phylogeny, a tree-like structure. The matrix's similarity or distance is normally calculated by comparing two DNA or protein sequences. The sequences that are the most similar will be close together, while those that are the most dissimilar will be far apart. Ancestor-related sequences have a lot of functional similarities **[Lewi, 1994]**.Analysis of Secondary Metabolite in Plants Shell (plantiSMASH) and PlantClusterFinder are gene cluster mining methods that allow users to utilize compound analysis to automatically discover BGC

plants, compare BGCs across all genomes, and predict active genetic interactions inside and across BGCs.It has the same easy-to-use, rich analysis and detection, and modular architecture as the widely used anti-SMASH predictive technique for microbial and fungal prediction. To utilize PlantiSMASH, users must have genomic data (with or without annotations) as well as transcriptome data. PlantiSMASH offers a number of configuration options as well as a number of aesthetic effects that need expert interpretation. PlantiSMASH supports both nonspecific (FASTA) and listed (GBK / EMBL / GFF + FASTA) genomic data **[Medema et al., 2011]**. By design, the method ignores scaffolds less than 1000 bp. PlantiSMASH includes a genetic analysis module that makes it easier to investigate compound patterns in BGCs predicted byan algorithm. SOFT files and CSV files are the two formats used by plantiSMASH. https://plantismash.secondary metabolites.org **[Kautsar, Duran, & Medema, 2017]**.

2.4 MACHINE LEARNING

Many issues in bioinformatics research can be recast as machine learning tasks. People are categorized in classification and regression because they share certain qualities; in clustering, individuals are grouped because they share certain properties; and in selecting features, the objective is to choose those features that are critical in predicting an individual's result. All three problem forms have algorithms in the Weka data mining suite. It has been used in bioinformatics for automated protein annotation **[Kretschmann, Fleischmann, & Apweiler,2001]**, Gene-expression array probe selection **[Tobler et al., 2002]**, automatic cancer diagnosis experiments **[Li, & Wong, 2002]**, creating a computational model for frame-shifting sites, plant genotype discrimination, and Gene expression profiles are being classified and rules are being extracted **[Witten, & Frank, 2002]** describes several of the algorithms in Weka. Real-world datasets differ, and no one algorithm is better than the others for all data mining problems. The algorithm must suit the structure of the issue in order to acquire relevant information or an accurate model. Weka was created with the goal of allowing the most versatility possible when experimenting with machine learning methods on new datasets. This comprises feature selection strategies (quick filtering and wrapper approaches), and pre-processing techniques for various types of models (e.g., decision trees, rule sets, and linear discriminants) (e.g., discretization, arbitrary mathematical transformations, and combinations of attributes).By providing a varied collection of techniques available through a single interface, Weka makes it simple to evaluate multiple solution strategies based on the same evaluation methodology and select the one that is most

suited for the situation at hand. It's written in Java and should work on nearly any machine. Random forests algorithm is a statistical and machine learning algorithm utilized for classification and regression analysis **[Schonlau, & Yuyan Zou, 2020]**. Random forest classifier is inbuilt in Weka library and utilized for the decision making and prediction. The random forest classification and regression method induces each constituent decision tree from a bootstrap sample of the training data. The support vector machines (SVMs) are one of the most robust and significant classification and regression analysis algorithm having application in various fields. The SVMs are playing important role in pattern recognition which is a most talked and popular research area. **[Cervantes et al., 2020].** A computerized Rice spikelet blast grade scale based on neural characteristics of tiny convolutional neural networks is proposed in a new analysis **[Sethy et al., 2021]**. In another study artificial neural network (ANN) plus gene-expression coding could accurately estimate rate of growth using cumulative weather conditions depending upon growth degrees days **[Liu et al., 2022]**. One Latest tool, GRAiN is a web-based interactive visualization application that enables to investigate the functional links among transcription factors and genomic components that underpin abiotic stress **[Gupta et al., 2021]**.

CHAPTER 3:
RATIONALE AND SCOPE OF STUDY

RATIONALE AND SCOPE OF STUDY

With the help of high throughput screening methods, combining systematic genome mining and functional analysis of candidate clusters along with artificial intelligence and machine learning the field of BGC discovery has seen exponential growth that has unraveled new pathways, enzymes, and chemistries in the plant kingdom. We hope to be part of the revolutionary movement with this project by exploring and studying the BGCs and their products in rice, the second most popular food crop in the world, and subsequently designing a machine-learning algorithm to predict potential gene clusters from an input plant genome. This could potentially be helpful in avenues such as plant crop weed interaction, nutrition, and synthetic biology.

Plants are well known to produce wide varieties of the specialized metabolites. This ability of the plants make them suitable candidate for modification at gene level to increase the production of these metabolites. These specialized metabolites are having many applications in food, cosmetics, pharmaceutical and chemical industry. Most of the gene cluster prediction tools available focus on the bacteria and fungi. There is need to extend the use of these tools to more complex genome to understand the diversity of specialized metabolites. Omics data and advances in genomics and bioinformatics along with sophisticated bioinformatics tools, have expanded our understanding of genomic and structural diversity, revealing practically limitless natural product discovery possibilities. Consisting of 4 objectives carried out in two stages, the first of which validates the fact that both signature and tailoring enzymes interact with one another for the complete expression of the BGC and are highly conserved and most functional, through molecular docking and evolutionary analysis with the help of dendrograms. The second stage entailed the development of the machine learning algorithm for predicting and classifying the genes of an input genome and identifying the BGCs where terpene synthases, which are involved in the production of different kinds of terpenes, one of the most common and important plant metabolites responsible for aroma, taste, and pigments of plants, were selected as the candidate used for training and testing against non-terpene synthases.

CHAPTER 4:
OBJECTIVES

OBJECTIVES

1. **To identify gene clusters in rice genome**
2. **To identify the regulatory mechanisms governing gene cluster expression.**
3. **Revealing the evolutionary forces behind the formation and maintenance of metabolic gene cluster**
4. **To device search engines for metabolic gene clusters in the sequenced plant genomes.**

CHAPTER 5:
METHODOLOGY

METHODOLOGY

- Objective 1:

To identify novel or known putative gene clusters common to plant homology search tools such as BLAST (1), PSI-BLAST and CASCADE-BLAST will be used. First, we were download complete genome sequences of desired plant. The homologous sequence for "signature genes" was searched against these genomes. We also perform these searches online against nr database to confirm the results. The specific gene locus of these genes was extracted from GenBank NCBI. Manual curation and GENOME VIEWER was used for identification of tailoring genes on the chromosomes having signature genes. Annotation of the signature and tailoring genes was performed.

The most commonly used and most reliable technique for characterizing newly discovered sequences is sequence similarity searches, which is generally done with BLAST. By finding excess statistically significant similarity that implies common ancestry, sequence similarity searches can uncover "homologous" proteins or genes. The Dynamic Programming approach produces an alignment in a time proportional to the product of the two sequence lengths being compared. As a result, while searching a large database, the calculation time scales linearly with the database size. Calculating a full Dynamic Programming alignment for each database sequence is too slow with existing databases (unless implemented in specialized parallel hardware).

PlantiSMASH: It enables the discovery, annotation, and analysis of secondary metabolite biosynthesis gene clusters across the plant kingdom on a genome-wide scale. It's a plant-specific version of the widely used antiSMASH tool (http://plantismash.secondarymetabolites.org/about.html).In this software we were required to give the email to which we want our results to be sent. In the second input box we upload our file of genomic or nucleotide sequence in GenBank or EMBL format. The file as a whole is quite big in size. To avoid and cut off time consumption we can use the third input box. In this we upload the NCBI accession no. of our desired file. Another option would be to segment whole genome into their chromosomes and upload each individually. This helps in reducing time consumption significantly. Once the results were obtained, we analyzed them for finding gene clusters, their size, and location and core domains.

- **Objective 2:**

1. FASTA sequences of amino acids were used as input for the 3D structure prediction web servers.
2. The obtained 3D structures (.pdb) were docked for exploring the protein-protein interactions. Two protein products from a BGC were loaded at the same time into PatchDock web server for molecular docking, one as the receptor and the other as the ligand.

- **Objective 3:**

1. FASTA sequences of BGC products were subjected to BLAST search.
2. Evolutionary analysis was performed for each of the BGC products by utilizing the phylogenetic tree/dendrogram produced using BLAST results.

Signature gene selection: Signature genes were selected from the gene clusters.

Information gathering on Signature and tailoring genes.

3D structure prediction.

Docking and interaction analysis.

Phylogenetic tree analysis.

- **Objective 4:**

1. From across the plant kingdom, Terpene synthase genes and non-terpene synthases and their products was selected as the candidates.
2. Ten physicochemical properties were selected and predicted using web servers. They were used as classification attributes to train the algorithm to distinguish between terpene synthases and non-terpene synthases.

3. A data was prepared in the required. arff format for WEKA tool and loaded to develop the machine learning algorithm.
4. Random forest and SMO classifiers available in WEKA was used to classify the input by a 66 percent split where 66 percent of the data was used for training and 34 percent were used for testing.

Collected FASTA sequences of Terpene synthases and Non terpene synthases

A total of 10 physicochemical properties were selected and predicted

The physicochemical properties were converted into. arff format required for WEKA

Random forest and SMO model building

The following were the web servers used for predicting the physicochemical properties.

5.1 SOFTWARES AND SERVERS

Softwares and servers used for objectives 1, 2, and 3 are; UniProtKB, Swiss-Model, Rosetta web server, Pymol, PatchDock server, BLAST, BLAST tree view

The following are the servers used for objective 4.

- ProtParam by ExPASy: Chain length, molecular weight, PI, instability index, aliphatic index, GRAVY(Hydropathicity) https://web.expasy.org/protparam/ **[Gasteiger et al., 2005]**
- PROTEIN CALCULATOR v3.4: Charge at pH7 http://protcalc.sourceforge.net/
- TM predictor: TMindex http://tm.life.nthu.edu.tw/
- Protein-sol: Solubility https://protein-sol.manchester.ac.uk/ **[Hebditch et al., 2017]**
- PepCalc.com-Peptide property calculator by INNOVAGEN: Extinction coefficient https://pepcalc.com/ **[Lear et al., 2016]**

5.1.1 UniProtKB:

The UniProtKB Knowledgebase (UniProtKB) is a database that collects detailed data about proteins in a uniform, accurate, and high-quality manner. Each UniProtKB item (which consists mostly of the amino acid sequence, protein name or description, taxonomy data, and citation information) must assemble the core data and, where feasible, add annotation data.This comprises generally established biological ontologies, classifications, and cross-references, as well as unambiguous indicators of annotation quality in the form of experimental and computational data evidence attribution.The UniProt Knowledgebase is divided into two sections: manual-annotation records with data from the literature and curator-evaluated computational analysis, and computationally analyzed records that are still waiting for preprocessing.The two parts are referred to as "UniProtKB / Swiss-Prot" (reviewed, manually annotated) and "UniProtKB / Trembl" (reviewed, manually annotated) for consistency and name familiarity (not reviewed, automatically quoted) (https://www.uniprot.org/).

5.1.2 Swiss-Model:

SWISS-MODEL, protein homology modelling server which is fully automated and accessed via the ExPASsy Web server or using the Deep View (Swiss PDB-Viewer) program, was used (https://swissmodel.expasy.org/) **[Waterhouse et al., 2018; Bienert et al., 2017]**

5.1.3 trRosetta web server:

The trRosetta method predicts protein structures from scratch rapidly and reliably. Direct energy minimization was used to construct protein structures. The deep residual neural network's predictions of inter-residual distance and orientation distributions point to the technique's shortcomings. Benchmark tests such as CASP13 and CAMEO-derived sets have shown how well trRosetta outperforms all of the previously described methods.

The query protein's amino acid sequence is fed into trRosetta. The inter-residual distance and orientation distributions of the provided query sequence are estimated using a deep residual neural network. The anticipated distance and orientation distributions are then translated to smooth constraints, which instruct the Rosetta to construct 3D structural models via direct energy minimization. (https://yanglab.nankai.edu.cn/trRosetta/) **[Yang et al., 2020]**.

5.1.4 ITASSER:

I-TASSER (Iterative Threading ASSEmbly Refinement) predicts protein structures and annotates them based on their function using a hierarchical method. For finding structural templates from PDB, LOMETS, a multiple threading technique, was utilized. To create full-length atomic models, iterative template-based fragment assembly simulations are used. BioLiP, a protein function library was used for obtaining function insights by re-threading the protein templates. (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) **[Roy, Kucukural, &Zhang, 2010, Yang et al.,2015].**

5.1.5 Pymol:

Warren Lifford Delano created PyMOL, an open-source molecular visualization toolkit. PyMOL was first commercialized by Delano Scientific LLC, a private software business devoted to create helpful tools for the scientific and academic community that are widely available, and is now sold by Schrödinger, Inc.Pymol can create high-resolution 3D pictures of biological macromolecules like proteins. According to the original author, PyMOL was used to create roughly a quarter of all published pictures of 3D protein structures in the scientific literature as of 2009.

5.1.6 PatchDock server:

PatchDock algorithm is used for molecular docking. Proteins, DNA, peptides, and medicines can all be used as input molecules. The docking findings were a list of possible complexes ordered by shape complementarity criteria. The patchdock approach makes use of computer vision techniques such as object identification and picture segmentation. Docking is similar to putting together a jigsaw puzzle. We seek for patterns that fit in with the rest of the problem while focusing on patterns that are particular to the puzzle piece. According to the form of the surface, the two molecules' surfaces are split into patches. These patches correlate to the patterns that distinguish puzzle pieces visually. Shape matching algorithms can be used to superimpose identified patches. (https://bioinfo3d.cs.tau.ac.il/PatchDock/php.php) **[Duhovny, Nussinov, &Wolfson,2002, Duhovny et al.,2005].**

5.1.7 BLAST:

As the name suggests, BLAST (Basic Local Alignment Search Tool) performs local alignments. Most proteins in nature have at least one useful area within them. These areas can be common among the proteins from various species. The BLAST calculation is tuned to find these gaps or low amplitudes of group intimacy. The near-alignment approach similarly implies that an mRNA can be encapsulated with a little bit of genomic DNA, as it requires more time in gene collection and testing. BLASTP performs the protein-protein alignment correlation, and its computation is the base for many different BLAST findings. BLAST identifies similar sequences between the two sequences using short sequences, which is a process called seeding, which is followed by fitting. A heuristic method where BLAST finds similar words that are short sequences in the input sequence and sequence in the database is employed. After which they are assembled and compiled, and then matched with database sequences. The protein sequence is uploaded in a FASTA format or the sequence file is uploaded directly. Necessary parameters (if any) are put in, such as explosion against particular organism database information, excluding organism information. The appropriate algorithm is selected for our search, such as quick BLAST, phi BLAST among others. The results can be further analyzed and presented by superfamilies, similar sequences, conserved regions, and similarity trees. (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

5.1.8 BLAST tree view:

NCBI Blast automatically generates a Blast Tree, which is a basic pairwise alignment tree. The phylogenetic tree is a real tree that depicts numerous gene/protein alignments. This (BLAST tree) tries to show the relationship between BLAST hits in a hazy way. A genuine phylogenetic tree demonstrates and interprets the proper evolution of a gene or protein. The NCBI Web Blast service now has a Tree View option that displays a dendrogram or tree that clusters sequences based on their distances from the query sequence. The anomaly of sequences or natural groupings of similar sequences in the BLAST output, such as members of gene families or homologs of other species, is identified in this presentation.

For all DNA-DNA or protein-protein comparisons, the BLAST output includes a link to the Tree View display known as the 'results distance tree.' By choosing the corresponding tab from the tree output, the trees can be shown as rectangular, slanted, radial, or force displays.

The distance between scenes is scaled in both rectangular and radial outputs. By mouse-overing the 'Show subtree' or 'Show alignment' links pop-up menu that displays at the tree's inner nodes; you may see subsets or any arrangement of the tree.

5.1.9 WEKA:

Loaded with several algorithms such as Baysian Network, SVMLib, Artificial Neural Network (ANN), Nearest Neighbor (IBk), Random Forest, etc. Ans user-friendly graphical front end, Weka is a convenient machine learning tool. For this study, we developed Random Forest and SMO models.

Chapter 6
RESULT & DISCUSSION

RESULT AND DISCUSSION

6.1 Identifications of Gene Clusters

- Results obtained from *Oryza sativa japonica* have the most occurrence of saccharide and putative gene clusters in chromosomes (1,2,3,4,5,6,8,9,10,11,12) and (8,9,6) respectively. (Table 1)
- Results obtained from *Oryza sativa indca* have the most occurrence of saccharide and putative gene clusters in chromosomes (1 to 6 and 9 to 12) and (1 to 3, 6 and 9 to 10) respectively. (Table 2)

Table 1. Gene clusters of *Oryza sativa japonica* (chr-1 to chr-12)

	Sr. No.	Gene cluster	Size (kb)	Core Domains
Chromosome 1	1.	Saccharide	71.44	2OG-FeII_Oxy, DIOX_N, Peptidase_S10, UDPGT_2
	2.	Lignan-Polyketide	70.90	Chal_sti_synt_C, Chal_sti_synt_N, Dirigent, p450
	3.	Saccharide	82.51	Aminotran_1_2, UDPGT_2
	4.	Saccharide	72.22	UDPGT_2, p450
	5.	Alkaloid	33.28	Bet_v_1, Epimerase, Methyltransf_11
Chromosome 2	1.	Saccharide	139.97	Glycos_transf_1, p450
	2.	Saccharide-Polyketide	211.17	Chal_sti_synt_C, UDPGT_2, p450
	3.	Terpene	369.98	COesterase, Terpene_synth, Terpene_synth_C, p450
Chromosome 3	1.	Lignan-Saccharide	97.55	Cellulose_synt, Dirigent, Methyltransf_11, UDPGT_2
	2.	Saccharide	64.12	Amino_oxidase, UDPGT_2, adh_short

Chromosome 4	1.	Terpene	212.71	Terpene_synth, Terpene_synth_C, adh_short_C2, p450
	2.	Saccharide-Alkaloid	360.51	Cu_amine_oxid, UDPGT_2, adh_short
	3.	Saccharide	169.20	Peptidase_S10, UDPGT_2
	4.	Terpene	334.35	2OG-FeII_Oxy, Terpene_synth, Terpene_synth_C
	5.	Saccharide	42.28	Peptidase_S10, UDPGT_2
	6.	Terpene	61.50	Terpene_synth, Terpene_synth_C, Transferase
	7.	Lignan	82.15	2OG-FeII_Oxy, DIOX_N, Dirigent, Methyltransf_7
Chromosome 5	1.	Saccharide	207.12	2OG-FeII_Oxy, DIOX_N, Transferase, UDPGT_2
Chromosome 6	1.	Putative	71.58	2OG-FeII_Oxy, DIOX_N
	2.	Putative	105.71	Peptidase_S10, Transferase, adh_short_C2
	3.	Saccharide	165.15	Transferase, UDPGT_2
	4.	Polyketide	133.31	Chal_sti_synt_C, p450
Chromosome 7	1.	Lignan	86.46	Aminotran_1_2, Dirigent
	2.	Lignan-Saccharide	88.18	Aminotran_1_2, Dirigent, Glycos_transf_1
	3.	Lignan	86.37	COesterase, Dirigent, p450

Chromosome 8	1.	Saccharide-Terpene	127.02	Methyltransf_2, Terpene_synth, Terpene_synth_C, UDPGT_2
	2.	Lignan-Alkaloid	132.28	Bet_v_1, Dirigent, Epimerase
	3.	Putative	83.82	COesterase, adh_short
Chromosome 9	1.	Saccharide	99.62	AMP-binding, UDPGT_2, p450
	2.	Putative	150.15	COesterase, Peptidase_S10, adh_short
Chromosome 10	1.	Saccharide	141.74	Transferase, UDPGT_2, p450
	2.	Lignan-Saccharide	432.20	Dirigent, UDPGT_2, p450
	3.	Polyketide	141.94	Acetyltransf_1, COesterase, Chal_sti_synt_C, Epimerase
	4.	Polyketide	139.12	Amino_oxidase, Chal_sti_synt_C, GMC_oxred_C, GMC_oxred_N
Chromosome 11	1.	Alkaloid	41.98	HMGL-like, Str_synth, p450
	2.	Lignan	130.12	Dirigent, Peptidase_S10
	3.	Saccharide	468.44	2OG-FeII_Oxy, UDPGT_2, adh_short, adh_short_C2
Chromosome 12	1.	Lignan	323.86	Dirigent, Methyltransf_2, p450
	2.	Saccharide	67.79	Glycos_transf_1, p450

Table 2. Gene clusters of *Oryza sativa indica* (chr-1 to chr-12)

Chromosome	Sr. No.	Gene cluster	Size (kb)	Core Domains
Chromosome 1	1.	Putative	90.75	adh_short, p450
	2.	Saccharide	106.76	2OG-FeII_Oxy, DIOX_N, Peptidase_S10, UDPGT_2
	3.	Polyketide	47.74	Chal_sti_synt_C, Chal_sti_synt_N,p450
	4.	Putative	143.09	COesterase, Peptidase_S10, p450
	5.	Saccharide	55.04	Methyltransf_11, UDPGT_2
	6.	Saccharide	63.62	UDPGT_2, p450
	7.	Alkaloid	33.55	Bet_v_1, Epimerase, Methyltransf_11
	8.	Lignan	12.69	Dirigent, Epimerase
Chromosome 2	1.	Terpene	88.21	SQHop_cyclase_C, SQHop_cyclase_N
	2.	Saccharide	221.17	Glycos_transf_1, p450
	3.	Saccharide-Polyketide	97.73	Chal_sti_synt_C, UDPGT_2
	4.	Terpene	334.45	COesterase, Terpene_synth,Terpene_synth_C, p450
	5.	Saccharide	41.63	Glycos_transf_1, Glycos_transf_2, Methyltransf_11
Chromosome 3	1.	Putative	107.70	Acetyltransf_1, Amino_oxidase, Aminotran_1_2
	2.	Alkaloid	89.76	Aminotran_1_2, Bet_v_1,adh_short, adh_short_C2
	3.	Saccharide	53.65	AMP-binding, SQS_PSY, UDPGT_2
	4.	Lignan-Saccharide	82.00	Cellulose_synt, Dirigent, Methyltransf_11, UDPGT_2
	5.	Saccharide	79.53	Amino_oxidase, TPMT, UDPGT_2, adh_short
Chromosome 4	1.	Terpene	269.58	2OG-FeII_Oxy, DIOX_N,Terpene_synth, Terpene_synth_C, p450

	2.	Putative	60.73	Methyltransf_2, Transferase
	3.	Saccharide-Alkaloid	259.95	Cu_amine_oxid, UDPGT_2,adh_short
	4.	Terpene	159.38	Terpene_synth, Terpene_synth_C,adh_short
	5.	Saccharide-Polyketide	101.87	Chal_sti_synt_C, Chal_sti_synt_N, UDPGT_2
	6.	Saccharide	36.88	Peptidase_S10, UDPGT_2
	7.	Terpene	73.91	Terpene_synth, Terpene_synth_C, Transferase
	8.	Lignan	96.10	2OG-FeII_Oxy, DIOX_N, Dirigent, Methyltransf_7
Chromosome 5	1.	Saccharide	46.81	2OG-FeII_Oxy, DIOX_N, UDPGT_2
	2.	Putative	124.24	2OG-FeII_Oxy, Acetyltransf_1, DIOX_N, Methyltransf_2, p450
Chromosome 6	1.	Lignan-Saccharide	156.97	Aminotran_1_2, Dirigent, Transferase, UDPGT_2
	2.	Putative	186.18	Methyltransf_2, Methyltransf_7, Peptidase_S10
	3.	Saccharide	221.10	Methyltransf_2, UDPGT_2
	4.	Saccharide	160.38	UDPGT, UDPGT_2
	5.	Alkaloid	152.33	Cellulose_synt, Cu_amine_oxid
Chromosome 7	1.	Saccharide-Polyketide	145.89	Chal_sti_synt_C, Chal_sti_synt_N, Glycos_transf_1, Transferase
	2.	Lignan-Saccharide	75.21	Aminotran_1_2, Dirigent, Glycos_transf_1
	3.	Lignan	67.44	COesterase, Dirigent, p450
Chromosome 8	1.	Saccharide-Terpene	96.89	Methyltransf_2, Terpene_synth, Terpene_synth_C, UDPGT_2
	2.	Alkaloid	92.81	Str_synth, Transferase
	3.	Alkaloid	103.26	Epimerase, FA_desaturase,Str_synth
Chromoso	1.	Saccharide	104.2	AMP-binding, UDPGT_2, p450

me 9			8	
	2.	Putative	123.70	COesterase, Peptidase_S10, adh_short
Chromosome 10	1.	Polyketide	169.74	Chal_sti_synt_C, Chal_sti_synt_N, p450
	2.	Saccharide-Polyketide	138.86	Chal_sti_synt_C, Chal_sti_synt_N, UDPGT_2, p450
	3.	Saccharide	119.93	Transferase, UDPGT_2, p450
	4.	Lignan-Saccharide	169.86	Dirigent, UDPGT_2
	5.	Polyketide	41.54	Acetyltransf_1, COesterase,Chal_sti_synt_C
	6.	Putative	64.83	2OG-FeII_Oxy, DIOX_N
Chromosome 11	1.	Alkaloid	34.30	HMGL-like, Str_synth, p450
	2.	Lignan	62.02	Dirigent, Peptidase_S10
	3.	Saccharide	125.71	Peptidase_S10, UDPGT_2
	4.	Polyketide	130.39	Chal_sti_synt_C, Chal_sti_synt_N
Chromosome 12	1.	Alkaloid	57.99	Epimerase, HMGL-like, Str_synth, p450
	2.	Lignan	175.27	COesterase, Dirigent, Peptidase_S10
	3.	Terpene	55.16	2OG-FeII_Oxy, DIOX_N,Terpene_synth_C
	4.	Saccharide	100.29	Glycos_transf_1, Peptidase_S10, p450

6.1.2. Analysis of gene clusters:

We analysed results obtained for putative clusters to find similarities in order to identify and predict their functions via Blast, as also described in Table 1 and Table2. Table 3 provides comparative analysis between clusters of *Oryza sativa Indica and japonica verities.*

Japonica group Putatives:

- Chromosome 6, Cluster 1: Most similar sequence that was obtained was Proteosome subunit alpha-type. (Query coverage: 9%. Similarity: 100%)
- Chromosome 6, Cluster 2: Most similar sequence that was obtained was HoubaCopia like Retrotransposon. (Query coverage: 21%. Similarity: 100%)

- Chromosome 8, Cluster 3: Most similar sequence that was obtained was Carboxylesterase. (Query coverage: 11%. Similarity: 97%.)
- Chromosome 9, Cluster 2: Most similar sequence that was obtained was Glycosyltransferase. (Query coverage: 17%. Similarity: 99%.)

Indica group Putatives:

- Chromosome 1, Cluster 1: Most similar sequence that was obtained was Proteosome subunit alpha-type. (Query coverage: 7%. Similarity: 89%.)
- Chromosome 1, Cluster 4: Most similar sequence that was obtained was Peroxidase precursor. (Query coverage: 8%. Similarity: 96%.)
- Chromosome 3, Cluster 1: Most similar sequence that was obtained was Isoflavanone. (Query coverage: 9%. Similarity: 87%.)
- Chromosome 4, Cluster 2: Most similar sequence that was obtained was Carboxylesterase. (Query coverage: 17%. Similarity: 100%.)
- Chromosome 5, Cluster 2: Most similar sequence that was obtained was Proteosome subunit alpha-type. (Query coverage: 8%. Similarity: 88%.)
- Chromosome 6, Cluster 2: Most similar sequence that was obtained was Amyloplastic. (Query coverage: 13%. Similarity: 94%.)
- Chromosome 9, Cluster 2: Most similar sequence that was obtained was prx 89 gene. (Query coverage: 21%. Similarity: 83%.)
- Chromosome 10, Cluster 6: Most similar sequence that was obtained was Hydroxyl Isoflavanone. (Query coverage: 7%. Similarity: 88%.)

Table 3. Comparison between Japonica and Indica

Japonica group	**Indica Group**
There were **thirteen saccharide gene clusters** in the obtained result. **Inference**: Sugar content lower than rice of Indica group	1)There were **fifteen saccharide gene clusters** in the obtained result. **Inference**: Sugar content higher than rice of Japonica group.
2)**Two Alkaloid gene clusters** were obtained. **Inference**: Nitrogenous organic compound content is much lower than rice of Indica group.	2)**Seven Alkaloid gene clusters** were found. **Inference**: Nitrogenous organic compound content is much higher than rice of Japonica group.
3)There were **four terpene gene clusters** in the obtained result.	3)There were **six terpene gene clusters** in the obtained result.

Inference: Less fragrant than rice of Indica group. 4)There were **five lignan gene clusters** in the obtained result. **Inference**: Phytoestrogen levels are similar.	**Inference**: More fragrant than rice of Japonica group. 4)There were **five lignan gene clusters** in the obtained result. **Inference**: Phytoestrogen levels are similar.

1.) Saccharides:

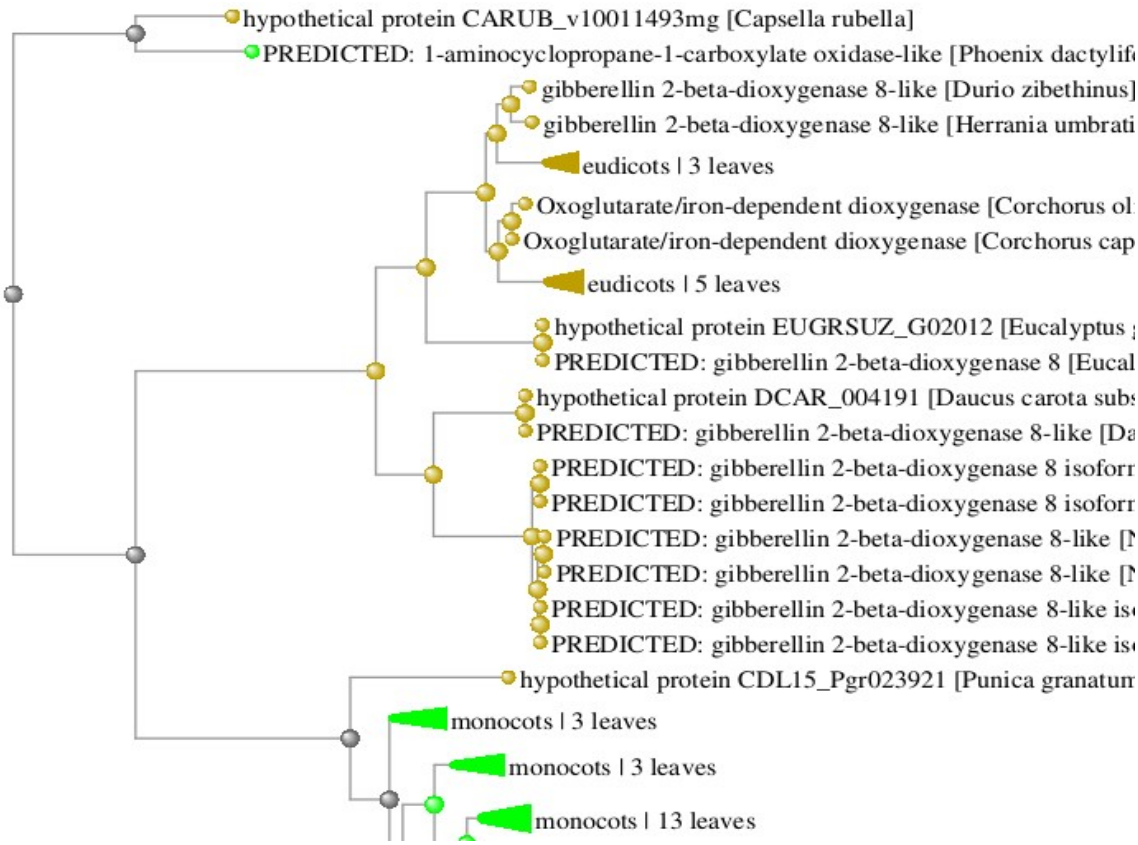
Results obtained have the most occurrence of saccharide and putative gene clusters in chromosomes (1,2,3,4,5,6,8,9,10,11,12) and (8,9,6) respectively.

Analyzed results of putative:

2OG_fell_oxy:

Out of all the sequences, sequence 1 was used for obtaining results.

Figure 1. Phylogenetic tree obtained in cluster one of putative:



Out of 26 similar sequences that were obtained in the tree, predicted sequences obtained were 9. The most similarity was exhibited by the hypothetical protein CARUB. Least was exhibited by the predicted proteins. Super families for cluster one are defined in Table 4, and phylogenetic tree for the same is provided in Figure1.

Table 4. Super families of cluster one

Name	Accession	Description	Interval	e-value
PLN00417	PLN00417	oxidoreductase, 2OG-Fe (II) oxygenase family protein	51-352	5.15e-89
PcbC	COG3491	Isopenicillin N synthase and related dioxygenases [Secondary metabolites biosynthesis]	50-347	3.29e-33
DIOX_N	pfam14226	non-haem dioxygenase in morphine synthesis N-terminal; This is the highly conserved N-terminal.	51-180	3.00e-26

Name	Accession	Description	Interval	e-value
PLN00417	PLN00417	oxidoreductase, 2OG-Fe (II) oxygenase family protein	16-311	1.37e-74
2OG-FeII_Oxy	pfam03171	2OG-Fe (II) oxygenase superfamily; This family contains members of the 2-oxoglutarate (2OG)	170-263	1.86e-28
PcbC	COG3491	Isopenicillin N synthase and related dioxygenases [Secondary metabolites biosynthesis]	44-306	6.53e-26

2.) Alkaloids:

In the Rice genome we got alkaloid clusters in 3 different chromosomes; chromosome number 1, chromosome 8 and chromosome 11. In chromosome number 1 the alkaloid gene cluster consists of 3 genes that are Bet_V_1, Epimerase and Methyltransf_11. Similarly in the chromosome number 8 also has 3 different genes which are Bet_V_1, Dirigent and Epimerase. Finally, in the chromosome number 11 the Alkaloid gene cluster had 3 different genes which were HMGL-like (Pyruvate Carboxylase), Str_synth and p450 (Cytochrome Superfamily). To begin with this research, we first used the tool String in order to predict interactions between various proteins. These interactions might be direct (physical) association or indirect(functional). These data are collected from various preexisting databases as well as other sources like automated text mining, genomic context prediction,

etc. Through this tool we got an idea of how the gene product in our gene cluster interacted with various other proteins. However, conclusive data wasn't actually received from that application alone thus we further went forward to get our results more conclusive. Then we used a tool that predicted the motifs of regulatory elements of the gene to see if the genes were actually regulated by the same regulatory elements or not. This would help us get a more conclusive result and help us further understand our gene cluster and its function. For the regulatory element inspection, we used a tool called NSitePL developed by Softberry. This program looks for functional patterns in plant promoter/regulatory sequences that are statistically significant. Plant functional motifs were chosen from the RegSite Database based on published research on plant gene transcription regulation. The data collected are kept in order of their chromosome.

Chromosome 1: Bet_V_1:

Bet v 1 refers to a protein allergen family. Allergens are molecules that, when they come into touch with the human body, cause an immune reaction. The main birch pollen allergen is named after this family. The protein belongs to the plant pathogenesis-related protein family 10. (PR-10). Cytoplasmic proteins of 15-17kd are abundant in dicotyledonous plants. Many plant proteins with low sequence similarity to Bet v 1 (Figure2) have recently been discovered, and when categorized by sequence similarity, many subfamilies related to PR-10 have emerged. Some of them were shown to be the most common tree pollen allergens in birch and similar species such as hazel and alder. These allergens were found in significant concentrations in fruits, vegetables, and seeds, as well as as a pathogenesis-related protein that is activated by abiotic stressors, injury, or pathogen infection. Most of the proteins that belonged to this family were reported in dicotyledonous plants however, related sequences of these proteins were also identified in monocots.

Figure 2: Birch Pollen Allergen Bet V 1 PDB 1bv1

This sequence was analyzed using the tool soft berry Nsite using the following parameters:

- Expected_Mean_Number :0.0100000
- Statistical_Significance_Level:0.9500000
- Level of homology between known_RE and motif:80%
- Variation of Distance between RE Blocks:20%

After running the sequence in the tool for finding the motifs of regulatory elements we got a result where 11 motifs of 9 different Regulatory Element were found. The search was done by putting our organism as *Oryza sativa* of the *Japonica* cultivar but we got results in which there were many such genes and motifs of homology from various other organisms. Thus, the entire results are presented in tabular form below for the gene Bet_V_1(Table5) which has length of 477 bp.

Table 5. Regulatory elements for Bet_V_1

S. N	Strand	Sequence of RE	Location	Organism Species	Gene
1	Negative	ATATTTAAA ATATTTAAA	243-235 198-190	Soybean (Glycine max)	beta-conglicinin (7S globulin) alpha' and beta subunits
2	Positive	gAAATTAAA aTTgTG	301-315	tomato (*Lycopersicon esculentum)*	rbcS1(Ribulose bisphosphate carboxylase small chain)
3	Negative	CAGAATAA	72-65	tomato (*Lycopersicon esculentum*)	rbcS1(Ribulose bisphosphate carboxylase small chain)
4	Positive	GAAATTAAA AtTGt	301-314	carrot (*Daucus carota*)	DC 59
5	Positive	GATGAAGTG G	214-223	maize (*Zea mays*)	b-32
6	Negative	TTGTtgCTAA TTTTCaGAAt G	356-336	soybean (Glycine max)	Ferritin
7	Positive	ATTaAAAAA aTTA	157-169	soybean (Glycine max)	BiP9
8	Negative	TTTAAAGT	240-233	rice (*Oryza sativa*)	GluA-3(Ionotropic glutamate receptors)
9	Positive Negative	GGCGCGCC GGCGCGCC	45-52 52-45	Arabidopsis (*Arabidopsis thaliana*)	CDKB1(Cyclin-dependent kinase)

Epimerase:

Epimerase are enzymes that help in catalyzing the sterochemical inversion of the configuration of a molecule about an asymmetric carbon atom in a substrate having more than one center of symmetry or in other words it works by interconverting epimers. In plants epimers are very important mainly because plants have to utilize light energy to convert inorganic carbon into organic forms which means that simple sugars derived from photosynthesis serve as an energy source and also work as a building block for cellular components such as cell walls, membranes and glycoproteins. In rice UDP-glucose/galactose-4-epimerase (UGE) is the epimerase we have chosen for the experiment as it is one of the epimerases that is present with proper characterization as well as annotation. The sequence of this epimerase of rice was analyzed as well in order to look for its motifs of regulatory elements. This sequence was analyzed using the tool Softberry Nsite using the following parameters:

- Expected Mean Number – 0.0100000
- Statistical Significance Level- 0.9500000
- Level of homology between known RE and motif- 80%
- Variation of Distance between RE Blocks- 20%

After running the sequence in the tool for finding the motifs of regulatory elements we got a result where 7 motifs of 7 different Regulatory Element were found. The search was done by putting our organism as *Oryza sativa* of the *Japonica* cultivar but we got results in which there were many such genes and motifs of homology from various other organisms. Thus, the entire results are presented in tabular form below for the gene Epimerase (Table6) which has length of 1062 bp.

Table 6. Regulatory elements for Epimerase:

S. N	Strand	Sequence of RE	Location	Organism Species	Gene
1	Negative	CAGAATAA	546-539	tomato (*Lycopersicon esculentum)*	rbcS1(Ribulose bisphosphate carboxylase small chain)
2	Negative	AAAAATCT	234-227	Arabidopsis (*Arabidopsis thaliana)*	Lhcb1*3(Chlorophyll a-b binding protein 1, chloroplast)
3	Positive	AACAACCTGGATT	175-187	rice (*Oryza sativa*)	GluA-3(Ionotropic glutamate receptors)
4	Negative	CTTCAAAGCGC	253-243	Arabidopsis (*Arabidopsis*	At3g62160(acyl-transferase family protein)

				thaliana)	
5	Positive	GAAAATTATTCT	534-547	Arabidopsis (*Arabidopsis thaliana*)	AtSUC2(sucrose-proton symporter 2)
6	Negative	aAACACcAGCGGA	897-885	soybean (Glycine max)	CHS8 (Chalcone synthase)
7	Positive	AAAAATTCCgC	879-889	barley (*Hordeum vulgare*)	HvHsp17.5CI (heat shock proteins)

Methyltransf_11:

This sequence was analyzed using the tool Softberry Nsite using the following parameters:

- Expected Mean Number – 0.0100000
- Statistical Significance Level- 0.9500000
- Level of homology between known RE and motif- 80%
- Variation of Distance between RE Blocks- 20%

After running the sequence in the tool for finding the motifs of regulatory elements we got a result where 8 motifs of 8 different Regulatory Element were found. The search was done by putting our organism as *Oryza sativa* of the Japonica cultivar but we got results in which there were many such genes and motifs of homology from various other organisms. Thus, the entire results are presented in tabular form below for the gene Methyl Transferase (Table 7) which has length of 1104 bp

Table 7. Regulatory elements for Methyltransf_11:

S. N	Strand	Sequence of RE	Location	Organism Species	Gene
1	Positive	AGCCGCCAT	670-678	Arabidopsis (*Arabidopsis thaliana*)	rd29A(Low-temperature-induced 78 kDa protein)
2	Negative	tAAATTCAAT	1099-1090	arabidopsis (*Arabidopsis thaliana*)	AtpC (ATP synthase epsilon chain)
3	Positive	CTTTAAAGCGA	1053-1063	Arabidopsis (*Arabidopsis thaliana*)	At3g62160(acyl-transferase family protein)
4	Positive	CTTTAAAGCGA	1053-1063	Arabidopsis (*Arabidopsis*	XSP1 (xylem serine peptidase 1)

				thaliana)	
5	Positive	aACGCGATTgAAC	103-115	Arabidopsis (*Arabidopsis thaliana*)	2CPA
6	Positive	GCATATTC	681-688	eggplant (*Solanum melongena*)	SmPT4(
7	Positive	CGCATATTCg	680-689	Arabidopsis (*Arabidopsis thaliana*)	AtIPS1(Inositol-3-phosphate synthase isozyme 1)
8	Negative	GAATATGC	688-681	Arabidopsis (*Arabidopsis thaliana*)	AtIPS3(Inositol-3-phosphate synthase isozyme 3)

Chromosome 8:

The gene consists of 2 genes that are already explained in above:

Dirigent: Dirigent proteins are members of a class of proteins which dictate the stereochemistry of a compound synthesized by other enzymes. Figure 3 shows structure of DRR206.

Figure 3: Structure of the dirigent protein DRR206

Dirigent proteins impart stereoselectivity on the phenoxy radical-coupling response, yielding optically dynamic lignans from two particles of coniferyl alcohol in the biosynthesis of lignans, flavonolignans, and alkaloids and in this way assumes a focal part in plant secondary metabolism. In lignan biosynthesis, oxidative proteins perform proton coupled electron exchange to expel a hydrogen particle from monolignols, framing a radical middle of the road.

These intermediates at that point couple in a radical end response to frame one of an assortment of dimers, known as lignans. The sequence was analyzed using the tool Softberry Nsite using the following parameters:

- Expected Mean Number – 0.0100000
- Statistical Significance Level- 0.9500000
- Level of homology between known RE and motif- 80%
- Variation of Distance between RE Blocks- 20%

After running the sequence in the tool for finding the motifs of regulatory elements we got a result where 6 motifs of 6 different Regulatory Element were found. The search was done by putting our organism as *Oryza sativa* of the *Japonica* cultivar but we got results in which there were many such genes and motifs of homology from various other organisms. Thus, the entire results are presented in tabular form below for the gene Dirigent (Table8) which has length of 492 bp.

Table 8. Regulatory elements for Dirigent

S. N	Strand	Sequence of RE	Location	Organism Species	Gene
1	Positive	CGCGGATC	336-343	Arabidopsis (*Arabidopsis thaliana*)	H4A748(Arabidopsis histone promoter)
2	Positive	CGCGGATC	336-343	Maize (*Zea mays*)	H3C4
3	Negative	TATAcAGAATaTC	70-58	Soybean (*Glycine max*)	Gmhspl7.5-E (Heat Shock Promoter)
4	Positive	ATTTTTTTAT	392-401	Soybean (*Glycine max*)	CHS8
5	Positive	ACATTATTG	164-172	Arabidopsis (*Arabidopsis thaliana*)	ANAC019(Arabidopsis thaliana NAC transcription factors)
6	Negative	ACATTATTG	115-106	Soybean (*Glycine max*)	synthetic oligonucleotides

Chromosome11: HMGL-Like:

Pyruvate carboxylase (PC) (Figure 4) is an enzyme which catalyzes the physiologically irreversible carboxylation of pyruvate to form oxaloacetate. It is a critical anaplerotic response that makes oxaloacetate from pyruvate.

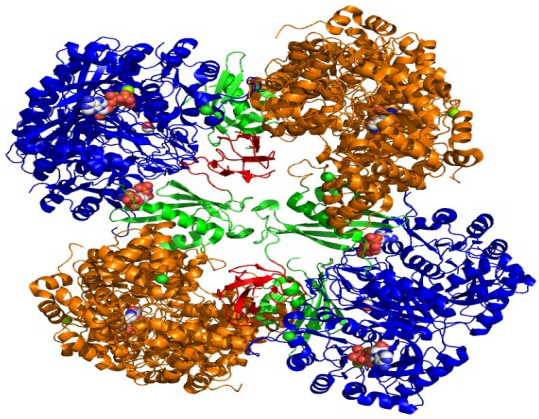


Figure 4: Crystallographic structure of pyruvate carboxylase from *Rhizobium etli*: biotin carboxylase domain (blue); allosteric linking domain (green); biotin binding domain (red); and carboxyl transferase domain (orange*)*

The chemical is a mitochondrial protein containing a biotin prosthetic gathering, requiring magnesium or manganese and acetyl CoA. It is found in a wide assortment of prokaryotes and eukaryotes including growths, microscopic organisms, plants, and creatures. In warm blooded creatures, PC assumes a urgent part in gluconeogenesis and lipogenesis, in the biosynthesis of neurotransmitters, and in glucose-initiated insulin discharge by pancreatic islets. Oxaloacetate delivered by PC is a vital middle of the road, which is utilized as a part of these biosynthetic pathways. In warm blooded creatures, PC is communicated in a tissue-particular way, with its movement observed to be most elevated in the liver and kidney (gluconeogenic tissues), in fat tissue and lactating mammary organ (lipogenic tissues), and in pancreatic islets. Movement is direct in cerebrum, heart and adrenal organ, and minimum in white platelets and skin fibroblasts.

This sequence was analyzed using the tool Softberry Nsite using the following parameters:

- Expected Mean Number – 0.0100000
- Statistical Significance Level- 0.9500000
- Level of homology between known RE and motif- 80%
- Variation of Distance between RE Blocks- 20%

After running the sequence in the tool for finding the motifs of regulatory elements we got a result where 20 motifs of 16 different Regulatory Element was found. The search was done by putting our organism as Oryza sativa of the Japonica cultivar but we got results in which there were many such genes and motifs of homology from various other organisms. Thus, the entire results are presented in tabular form below for the gene Pyruvate Carboxylase (Table 9) which has length of 3495 bp.

Table 9. Regulatory elements for HMGL-Like

S. N	Strand	Sequence of RE	Location	Organism Species	Gene
1	Positive	AtTATTTTTATT	355-366	Tobacco (*Nicotiana plumbaginifolia*)	cab-E(Polyketide oxygenase)
2	Positive	AtTATTTTTATT	355-366	Tomato (*Lycopersicon esculentum*)	rbcS-3A(Ribulose bisphosphate carboxylase small chain)
3	Positive	AtTATTTTTATt	355-366	Tobacco (*Nicotiana plumbaginifolia*)	cab-E(Polyketide oxygenase)
4	Positive	cATTATTTTTATT	354-366	Pea (*Pisum sativum*)	rbcS-3A(Ribulose bisphosphate carboxylase small chain)
5	Negative	ATgATGATGATgA ATgATGATGATgA ATgATGATGATgA ATgATGATGATgA ATgATGATGATgA	30-18 27-15 24-12 21-9 18-6	Tomato (*Lycopersicon esculentum*)	rbcS3B(Ribulose bisphosphate carboxylase small chain)
6	Positive	AGCAGGTgAAAC	1427-1438	Tobacco (*Nicotiana plumbaginifolia*)	Cab-E(Polyketide oxygenase)
7	Positive	AtTATTTTTATT	355-366	Rice (*Oryza sativa*)	osRACD
8	Positive	GATGATATGG	472-481	Maize (*Zea mays*)	b-32
9	Negative	CtTTCACCAgCACT	3448-3435	Arabidopsis (*Arabidopsis thaliana*)	C4H(Cinnamate-4-hydroxylase)
10	Positive	TATTTTAT	913-920	soybean (*Glycine max*)	GS15
11	Negative	CTTTAAAGCGA	169-159	Arabidopsis (*Arabidopsis thaliana*)	At3g62160 (Chloramphenicol acetyltransferase-like domain)
12	Negative	CTTTAAAGCGA	169-159	Arabidopsis (*Arabidopsis thaliana*)	XSP1(xylem serine peptidase 1)
13	Negative	TCAGATATT	889-881	Arabidopsis (*Arabidopsis thaliana*)	AGAMOUS
14	Positive	ATGTGACCGT	1478-1487	*Boea hygrometrica*	BhGolS1(galactinol synthase)
15	Negative	TTCAATCACTT	2940-2930	Arabidopsis (*Arabidopsis thaliana*)	BhGolS1(galactinol synthase)
16	Negative	tGCCGCgGCCGCCGCCGCC	559-541	Arabidopsis (*Arabidopsis thaliana*)	Synthetic oligonucleotides

Cytochrome P450

Cytochrome P450 (CYPs) are hemoproteins, which are proteins in the P450 superfamily that include heme as a cofactor. In enzymatic reactions, CYPs use a variety of small and large atoms as substrates. They are commonly found in P450-containing frameworks and are the terminal oxidase proteins in electron exchange chains. When the catalyst is in the reduced state and complexed with carbon monoxide, the spectrophotometric crest at the wavelength of the assimilation maximum of the catalyst (450 nm) is obtained. Chemicals derived from CYP (Figure5) have been found in all kingdoms of life, including animals, plants, parasites, protists, tiny organisms, archaea, and even viruses.

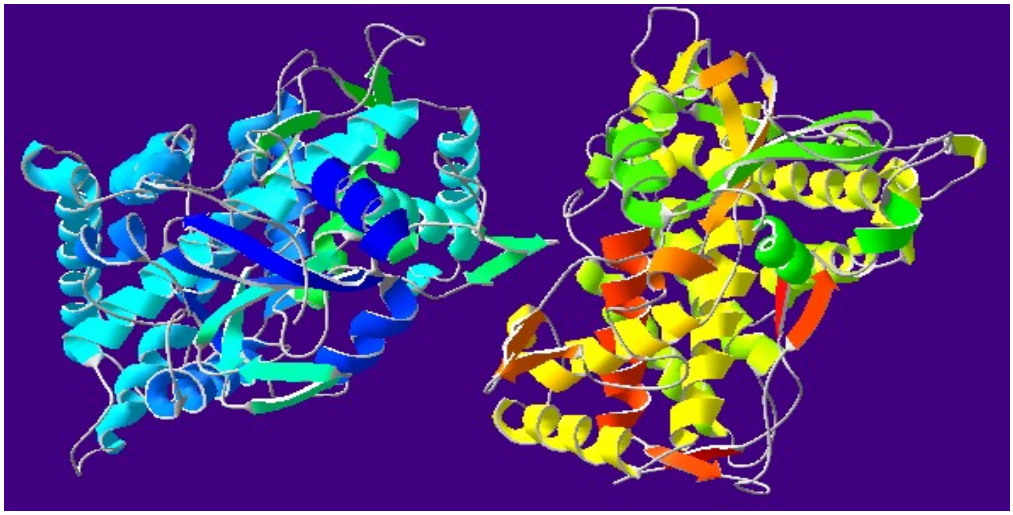


Figure 5: Cytochrome P450 Oxidase (CYP2C9)

More than 50,000 unique CYPproteins. To reduce the iron, most CYPs require a protein partner to transport at least one electron (and in the long run sub-atomic oxygen). Regulatory elements for P450 given in Table 10. In light of the electron transport proteins' concept. This sequence was analyzed using the tool Softberry Nsite using the following parameters:

- Expected Mean Number – 0.0100000
- Statistical Significance Level- 0.9500000
- Level of homology between known RE and motif- 80%
- Variation of Distance between RE Blocks- 20%

Table 10. Regulatory elements for P450

S.N	Strand	Sequence of RE	Location	Organism Species	Gene

1	Positive	AAAGCGGTGCaG	697-708	alfalfa (*Medicago sativa*)	MSPRP2
2	Positive	TGgAATTTCAG	1388-1398	tomato (*Lycopersicon esculentum*)	rbcS3B(Ribulose bisphosphate carboxylase small chain)
3	Positive	TATAAAGAATaTC	1105-1117	soybean (*Glycine max*)	Gmhspl7.5-E Heat Shock Promoter
4	Negative	TCATAGCCGCC	872-862	tobacco (*Nicotiana tabacum*)	GLN2(Glutamine synthetase)
5	Negative	TgCCATGCAGCaGC	1186-1173	barley (*Hordeum vulgare*)	ITR1(Myo-inositol transporter 1)
6	Negative	TCAGATATT	1120-1112	Arabidopsis (*Arabidopsis thaliana*)	AGAMOUS
7	Positive	AAAAAGAAA	971-979	rice (*Oryza sativa*)	CPD3(Class II CPD photolyase)
8	Negative	ATTTTTTTAT	1016-1007	soybean (*Glycine max*)	CHS8(Chitin synthase 8)
9	Positive	GCAGCCGCT	203-211	Arabidopsis (*Arabidopsis thaliana*)	PDF1(Protodermal factor 1)
10	Positive	CTAtAAAGAAT	1104-1114	Arabidopsis (*Arabidopsis thaliana*)	AmidP

After running the sequence in the tool for finding the motifs of regulatory elements we got a result where 10 motifs of 10 different Regulatory Element were found. The search was done by putting our organism as *Oryza sativa* of the *Japonica* cultivar but we got results in which there were many such genes and motifs of homology from various other organisms. Thus, the entire results are presented in tabular form below for the gene Cytochrome P450 which has length of 1440 bp.

Strictosidine synthase

The gene for strictosidine synthase, the enzyme that catalyzes the stereospecific condensation of tryptamine and secologanin to create the main indole alkaloid 3 alpha(S)-strictosidine, has been identified from *Rauvolfia serpentina* (India) and *Rauvolfia mannii* genomic libraries (West Africa). Str1 has no introns and exhibited 100 percent nucleotide sequence similarity between the two species spanning 1180 bp, including the complete reading frame. The R. serpentina gene's transcription begins 81 nucleotides upstream of the AUG (26 nucleotides downstream from the TATA box). Transient expression experiments in *Nicotiana plumbaginifolia* protoplasts of the R. serpentina str1 5'-noncoding region fused to the

beta-glucuronidase reporter gene indicated promoter activity of 4 +/- 2% of 35 S CaMV promoter control. The existence of three regions of minor, but repeatable, negative control was shown by a series of shortened portions of the str1 promoter region. Several areas of the 5'-flanking sequences selectively bound nuclear protein from R. serpentina, but at least one region did not bind nuclear protein from *R. mannii*, according to gel retardation tests. A study of str1 expression in *R. serpentina* plants found that strictosidine synthase poly(A)+ RNA was mostly found in the root, although not exclusively. This sequence was analyzed using the tool Softberry Nsite using the following parameters:

- Expected Mean Number – 0.0100000
- Statistical Significance Level- 0.9500000
- Level of homology between known RE and motif- 80%
- Variation of Distance between RE Blocks- 20%

After running the sequence in the tool for finding the motifs of regulatory elements we got a result where 6 motifs of 5 different regulatory Element was found. The search was done by putting our organism as *Oryza sativa* of the *Japonica* cultivar but we got results in which there were many such genes and motifs of homology from various other organisms. Thus, the entire results are presented in tabular form below for the gene Strictosidine synthase (Table11) which has length of 906 bp.

Table 11. Regulatory elements for Strictosidine synthase

S. N	Strand	Sequence of RE	Location	Organism Species	Gene
1	Negative	ATATTTAAT ATATTTAAT	483-475 108-100	soybean (*Glycine max*)	beta-conglicinin (7S globulin) alpha' and beta subunits genes/
2	Negative	CGgCCATGCAT	743-733	maize (*Zea mays*)	C1
3	Positive	CCATTTTTGG	693-702	tomato (*Solanum lycopersicum*)	Genomic fragment 133R
4	Negative	CGGAAtTTTCAc	660-649	tomato (*Solanum lycopersicum*)	SlHsp21.5-ER
5	Negative	CAAAcgGCGGCGGCAGCg	817-800	maize (*Zea mays*)	ZmArf34

Thus, from here we have observed and gained information about the motifs of the regulatory elements and thus now we can further analyze these data into refining our studies and coming to a bigger conclusion about our gene clusters.

- The gene clusters obtained via plantismash of rice chromosomes 1 to 12 of Japonica group and Indica group were varied in each with quite a few similar clusters present in each respectively.
- The predominant cluster that was obtained was saccharide in both groups while Putative has second-most occurrence. Core domains that have the most occurrence are 2OG-FeII_Oxy, transferases, and peptidases.
- The sequences that were obtained were annotated using Prosite and Blast. 2OG-FeII_Oxy: Proteins with Fe and 2-oxoglutarate (2OG) dioxygenase space typically catalyze the oxidation of a characteristic substrate using a dioxygen molecule, generally by using ferrous iron as the dynamic site cofactor and 2OG as a co-substrate which is decarboxylated to succinate and CO2. In rice, FeII-2OG dioxygenase space impetuses catalyze the advancement of plant hormones like gibberellin, ethylene, hues, and flavones.
- Transferases: The most common transferases found in the collected results were methyltransferases and UDPGT transferases. Uridine 5'-diphospho-glucuronosyltransferase (UDP glucuronosyltransferase, UGT) is a cytosolic glycosyltransferase that transfers the glucuronic acid component of Uridine 5'- diphospho-glucuronic acid to a hydrophobic atom. This is a response to glucuronidation. Methyltransferases as the name suggest transfers Methyl group from one functional group to another.
- Peptidases**:** Peptidase_S10 most observed peptidase in obtained results, this cluster of serine peptidases (SPep) belongs to MEROPS peptidase family S10 (group SC). All known carboxypeptidases are serine carboxypeptidases (SCPep.) or metallo carboxypeptidases. The reactant action of the SCPep. is similar to that of the trypsin family SP, which is given by a transfer framework including aspartic acid buildup hydrogen-attached to a histidine, which is itself hydrogen-linked to a serine.

Table 12. Signature genes details

Signature genes	PDB	Family	Superfamily	Uniprot	Function
Acetyl transferase	no	PF00450	SSF53474	P37890	Type of transferase enzyme that transfers an acetyl group.
Alcohol dehydrogenase (ADH)	no	PF08240 PF00107	SSF50129 SSF51735	Q2R8Z5	Reduce nicotinamide adenine di nucleotide (NAD+) to NADH to facilitate interconversion

					between alcohols and aldehydes or ketones.
Cellulose synthase	no	PF03552	NA	Q8W3F9	cellulose synthase is the main enzyme that produces cellulose
Chalcone synthase	no	PF02797	SSF53901	Q2R3A1	Chalcones, a family of chemical molecules found mostly in plants as natural defensive mechanisms and as synthetic intermediates, are linked to the synthesis of chalcones.
Carboxyl esterase (COesterase)	no	PF02230	SSF53474	Q0J968	catalyzes a chemical reaction to form an alcohol + a carboxylate from a carboxylic ester.
DIOX N	no	PF14226	NA	Q5ZA21	Morphine synthesis.
Epimerase	no	PF16363	SSF51735	Q6K2E1	Epimerases and racemases are isomerase enzymes that catalyze the inversion of stereochemistry in biological molecules.
Methyl transferase	no	PF00891	SSF53335 SSF46785	Q7XXI9	The transfer of a methyl group to the oxygen atom of an acceptor molecule is catalyzed.
Cytochrome P450	no	PF00067	SSF48264	Q7XU38	These proteins are required in the production of defense chemicals, fatty acids, and hormones in plants.
Peptidase S10	no	PF00450	SSF53474	P37890	A catalytic mechanism is used to catalyze the hydrolysis of a peptide bond that is less than three residues from the C-terminus of a polypeptide chain.
UDPGT(Glucuronosyltransferase)	no	PF03016	NA	Q8S1X7	The transfer of the glucuronic acid component of UDP-glucuronic acid to a tiny hydrophobic molecule is catalyzed by this enzyme.

6.2 Identification of the regulatory mechanisms governing gene cluster expression.

- **3-D structure prediction**

Oryza stativa japonica

None of the signature gene products had 3D structures available in PDB. Signature genes details given in Table 12. Swiss Model server was used for the prediction of all possible 3D conformations of these proteins by virtue of homology **[Waterhouse et al., 2018; Yang et al., 2020].** Due to some unforeseen errors, the models predicted using Swiss-Model were fragmented. Ab-Initio protein modeling using trRosetta and I-TASSER webserver were used to circumvent this issue **[Roy, Kucukural,& Zhang, 2010, Yang, 2015]**.TM Scores for selected enzymes was successfully interpreted: Acetyltransferase : Estimated TM-score: 0.753, Alcohol dehydrogenase : Estimated TM-score: 0.776, Cellulose synthase : Estimated TM-score = 0.70±0.12; C-score=-0.14,Chalcone synthase : Estimated TM-score: 0.815, COesterase : Estimated TM-score: 0.803, DIOXN : Estimated TM-score: 0.729, Epimerase : Estimated TM-score: 0.788, Methyl transferase : Estimated TM-score: 0.759, P450 : Estimated TM-score: 0.694, Peptidase S10 : Estimated TM-score: 0.706, UDPGT : Estimated TM-score: 0.476, trRosetta server predicts a total of 5 possible 3D structures for a given sequence (Figure 6 and 7). The quality of the predicted structures can be deduced from the Template Modelling scores (TM score). TM score signifies similarity which is akin to Root Mean Square Deviation (RMSD). A 100 percent identical match between two given structures is indicated by a TM score of 1. The structure of Cellulose synthase predicted through I-Tasser web server is validated by another parameter called the C-score. The C-score, which is a confidence score for quality, may be used to assess the quality of models projected by the I-Tasser server. For C-score estimation, the importance of threading template alignments and the convergence parameters of structure assembly simulations are taken into account.C-score values lie between (-5, 2). A higher C-score signifies that the model is of high confidence and a lower score indicates otherwise.

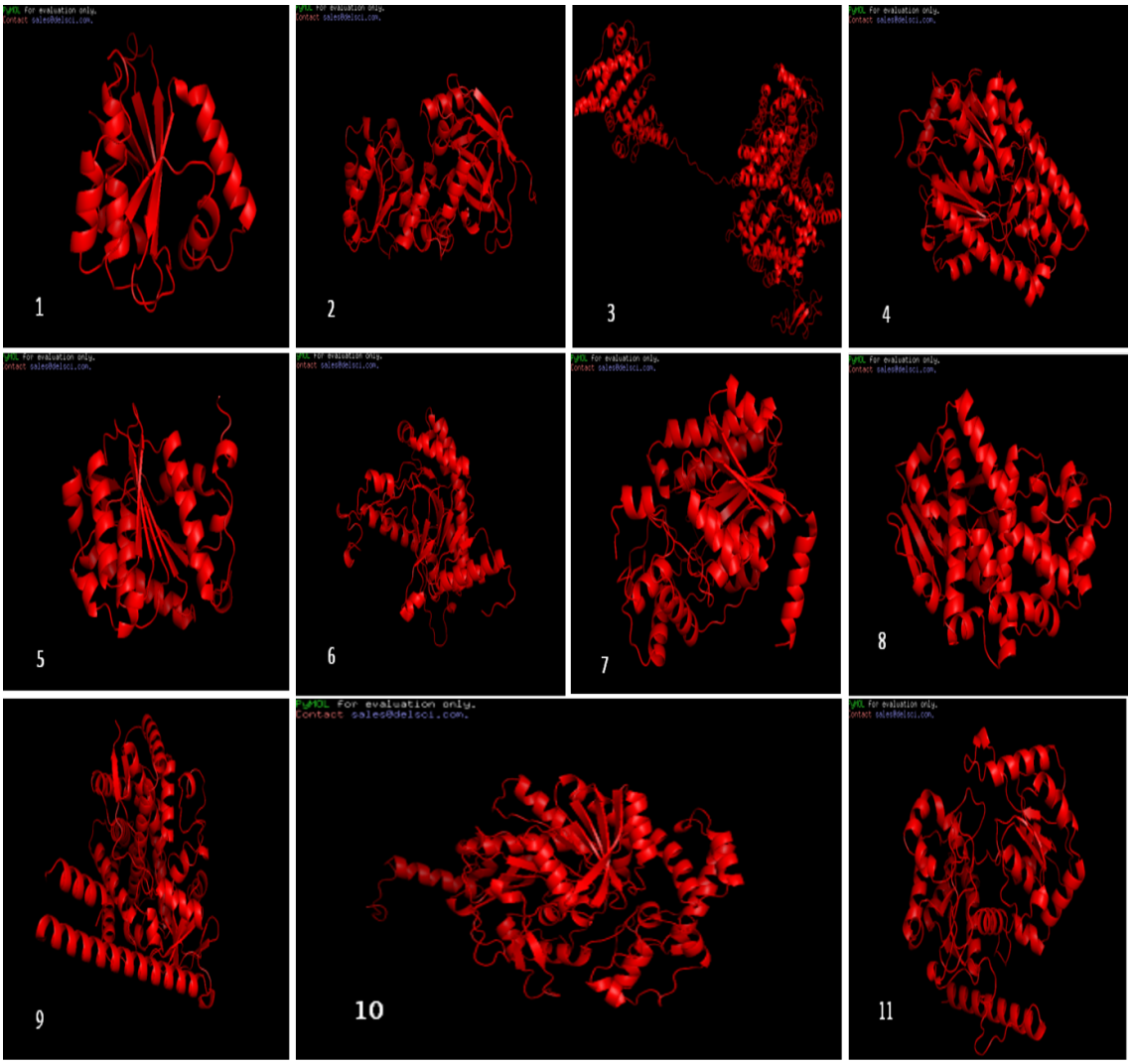


Figure 6: 3Dstructures of gene products (Japonica)

1. Acetyltransferase : Estimated TM-score: 0.753
2. Alcohol dehydrogenase : Estimated TM-score: 0.776
3. Cellulose synthase : Estimated TM-score = 0.70±0.12; C-score=-0.14
4. Chalcone synthase : Estimated TM-score: 0.815
5. COesterase : Estimated TM-score: 0.803
6. DIOXN : Estimated TM-score: 0.729
7. Epimerase : Estimated TM-score: 0.788
8. Methyl transferase : Estimated TM-score: 0.759
9. P450 : Estimated TM-score: 0.694
10. Peptidase S10 : Estimated TM-score: 0.706
11. UDPGT : Estimated TM-score: 0.476

Oryza sativa indica

10 out of the 11 proteins had their 3-D structures available in the Swiss Model Repository (SMR) database. The structure of Cellulase synthase was determined using I-Tasser web server which had the following parameters: C score=0.82, TM score=0.82 +- 0.08, RMSD=7.4 +- 4.2 Angstroms.

1. Alcohol dehydrogenase
2. Chalcone synthase
3. Methyl transferase
4. Epimerase
5. Glycosyl transferase
6. Terpene synthase
7. Acetyl transferase
8. Amino oxidase
9. Amino transferase
10. Sqs_psy
11. Cellulose synthase

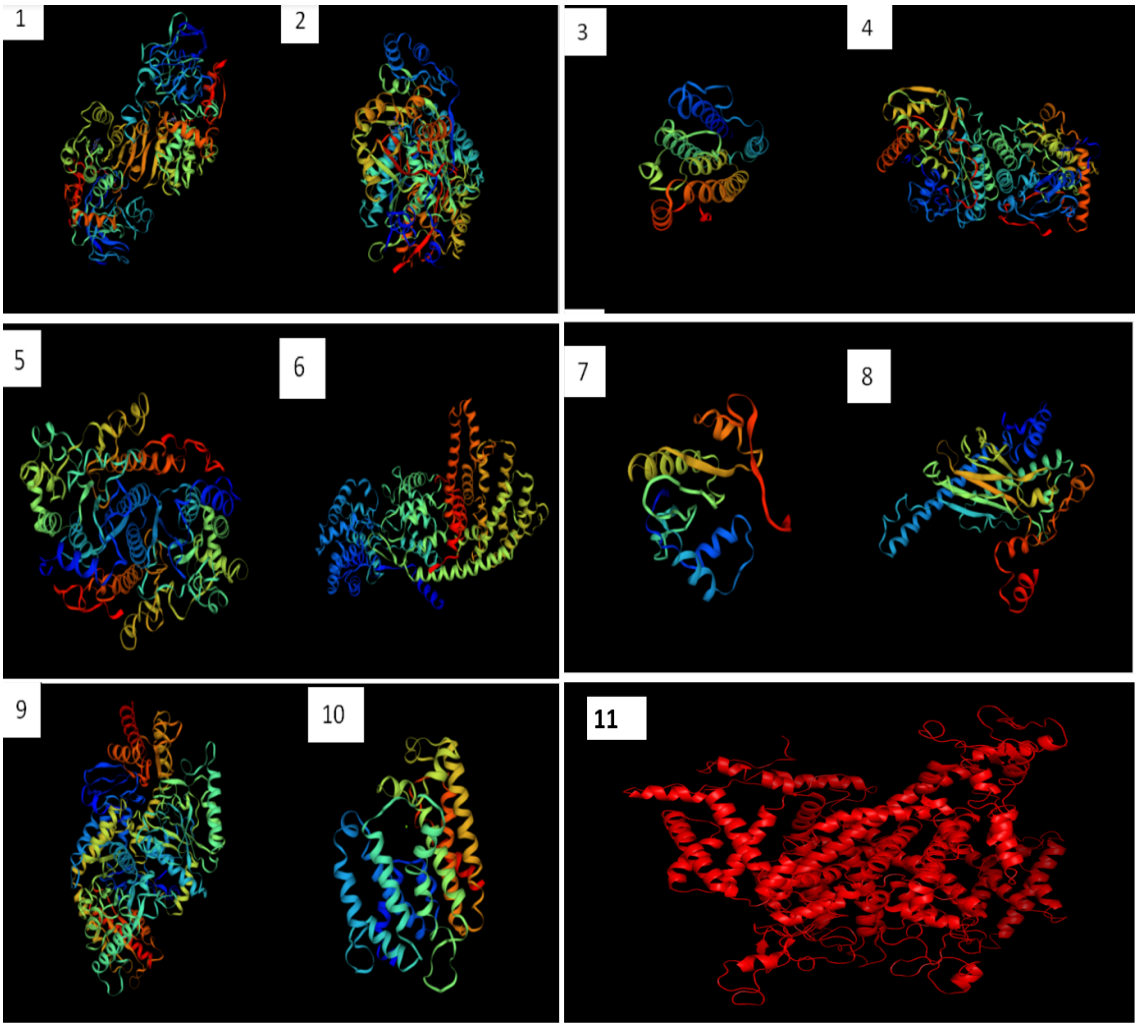


Figure 7:3D structures of gene products (Indica)

- **Interaction study**

The proteins produced by the signature genes were assumed to interact with the protein products of their respective tailoring genes of the cluster. To study the said interactions, the signature gene products were docked with the tailoring gene products using PatchDock server. Signature gene products in the docked complex are represented in red color and the tailoring gene products in green. Negative Atomic Contact Energies (ACE) between the docked proteins indicate the formation of a stable complex **[Duhovny, Nussinov, & Wolfson,2002; Duhovny et al., 2005]**. Docked complexes with the minimum ACE and considerably high docking score are selected since there is no one-to-one correlation between merely the docking score and the corresponding binding affinity.

Docked complexes:

Oryza sativa japonica

Various docked complexes were considered for molecular docking studies. Docked complexes are given in figure8 and enlisted below. Also, ACE values are provided in Table 13.

1.Methyltransferase-Cytochrome P450

2.Alcohol dehydrogenase (ADH)-Glucuronosyltransferase (UDPGT)

3.Glucuronosyltransferase (UDPGT) -Cytochrome P450

4.P450-UDPGT

5.Chalcone synthase-P450

6.UDPGT-Peptidase S10

7.Peptidase S10-ADH

8.Peptidase S10-UDPGT

9.UDPGT-DIOXN

10.Methyltransferase-UDPGT

11.Acetyl transferase-COesterase

12.Chalcone synthase-UDPGT

13.Acetyl Transferase-Epimerase

14.DIOX-Methyl transferase

15.P450-Chalcone synthase

16.COesterase-ADH

17.COesterase-P450

18.Peptidase S10-COesterase

19.Cellulose synthase-UDPGT

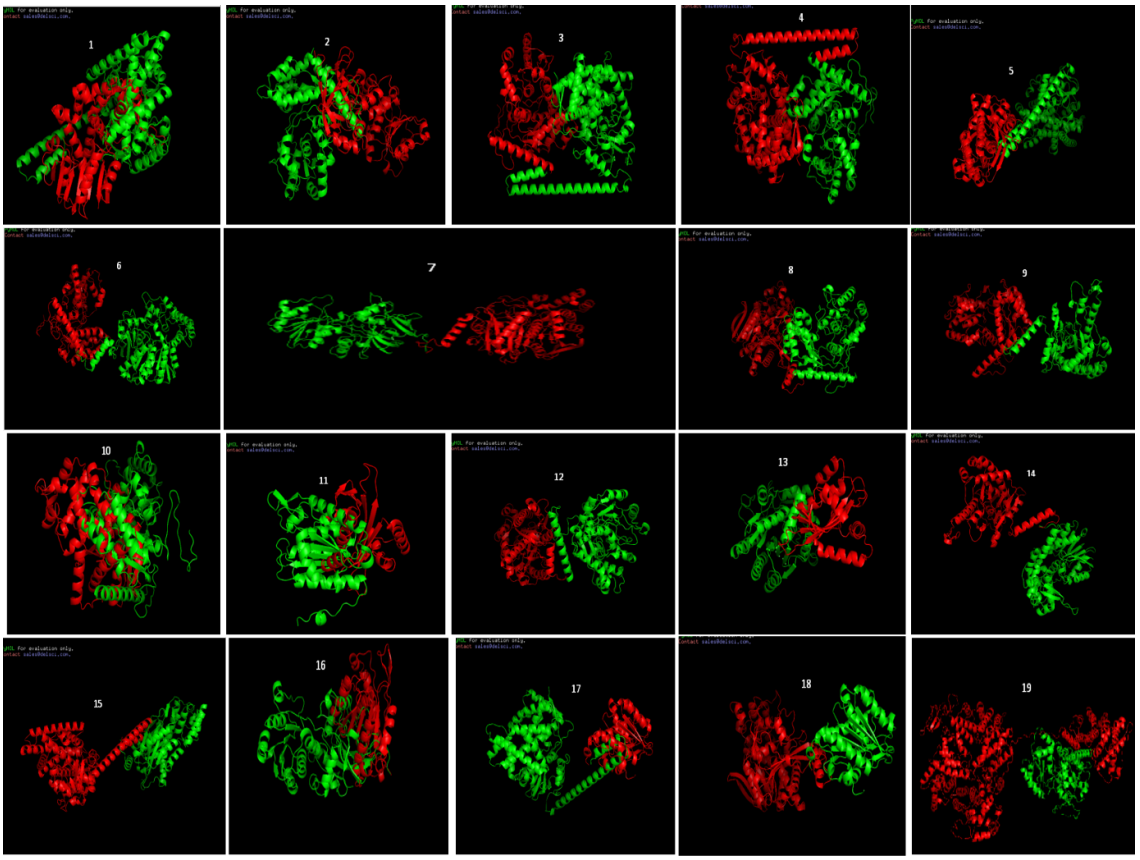


Figure 8: Docked complexes (Japonica).

Table 13. Docking results for Japonica

Complex	Score	ACE
1	13102	-776.1
2	11308	-195.48
3	15164	-201.53
4	12608	-738.46
5	12422	-764.76
6	11156	-154.14
7	11710	-145.8
8	12458	-767.48
9	11684	-420.05
10	11224	-225.08
11	13796	-506
12	11470	-314.33
13	12722	-348.01
14	15054	-298.56
15	14652	-483.03
16	12142	-406.08
17	16210	-530.46
18	15224	-369.26
19	14454	-715.05

Oryza sativa indica

The following 24 possible interactions were ascertained from the gene products. Docked complexes are given in figure 9 and enlisted below. Also, ACE values are provided in Table 14.

1. Chalcone synthase-Glycose transferase
2. Chalcone synthase-acetyl transferase
3. Methyl transferase-epimerase
4. Methyl transferase-glycosyl transferase
5. Methyl transferase-cellulase synthase
6. Methyl transferase-acetyl transferase
7. Methyl transferase-terpene synthase
8. Epimerase-methyl transferase
9. Glycosyl transferase-methyl transferase
10. Glycosyl transferase-chalcone synthase
11. Terpene synthase-methyl transferase
12. Terpene synthase-adh
13. Acetyl transferase-amino oxidase
14. Acetyl transferase-amino transferase
15. Acetyl transferase- methyl transferase
16. Acetyl transferase-chalcone synthase
17. Amino oxidase-acetyl transferase
18. Amino oxidase- amino transferase
19. Amino oxidase-adh
20. Amino transferase-amino oxidase
21. Amino transferase-acetyl transferase,
22. Amino transferase-ADH,
23. Amino tranferase-glycosyl transferase
24. Cellulose synthase-methyl transferase

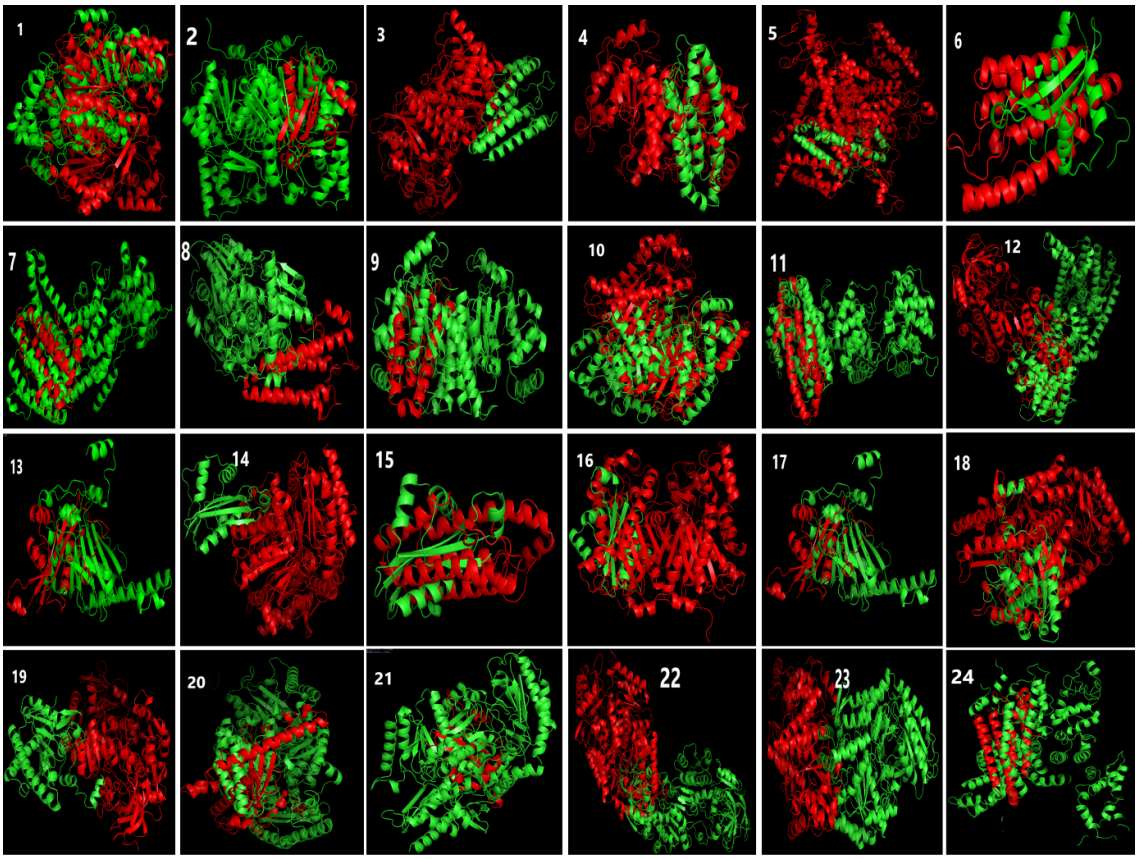


Figure 9: Docked complexes (Indica).

Table 14. Docking results for Indica

Complex	SCORE	ACE
1	11790	-159.44
2	12628	-453.91
3	13128	-389.17
4	13274	-650.6
5	18656	-677.43
6	11570	-705.53
7	16070	-541.07
8	12538	-475.57
9	14632	-540.81
10	14562	-123.81
11	14062	-600.66
12	9556	-283.86
13	12888	-658.49
14	11686	-228.65
15	11628	-598.1
16	12235	-322.94
17	13600	-442.66
18	11102	-152.6
20	11254	-126.55
21	12988	-142.99

22	6690	-197.77
23	10490	-133.43
24	20020	-1200.43

Sequences of the signature gene products were fed to pBLAST to obtain their homologs. This assisted in phylogenetic comparisons for selected enzymes and both were found highly conserved and most functional.

6.3 Revealing the evolutionary forces behind the formation and maintenance of metabolic gene clusters

6.3.1 Phylogenetic tree analysis:

Sequences of the signature gene products were fed to pBLAST to obtain their homologs. Using the obtained homology data, phylogenetic trees were created for each signature genes to study their evolution.

Oryza sativa Japonica

1. Acetyl transferase similarity tree

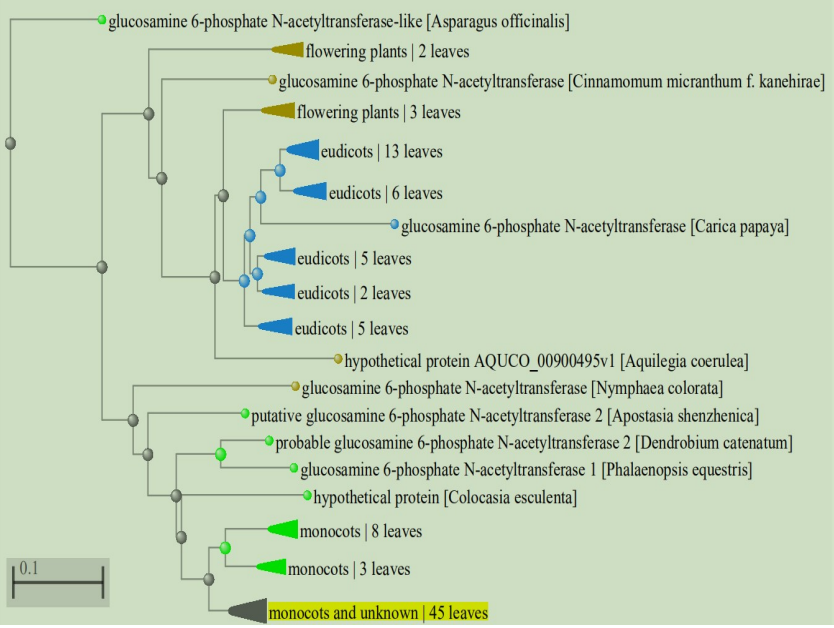


Figure 10: Acetyl Transferase similarity tree

Predicted sequences obtained were none. The most similarity was exhibited by Glucoseamine 6-phosphate N-acetyltransferase-like [*Asparagus officinalis*] proteins (Figure10).

2. ADH similarity tree

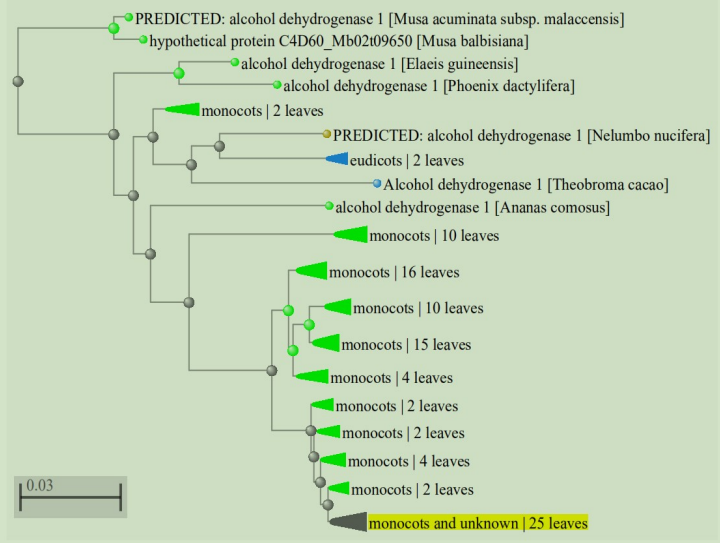


Figure 11: ADH similarity tree

Predicted sequences obtained were 2. The most similarity was exhibited by Alcohol dehydrogenase 1[*Musa acuminata* subsp. *malaccenis*] proteins (Figure11).

3. Cellulose synthase similarity tree

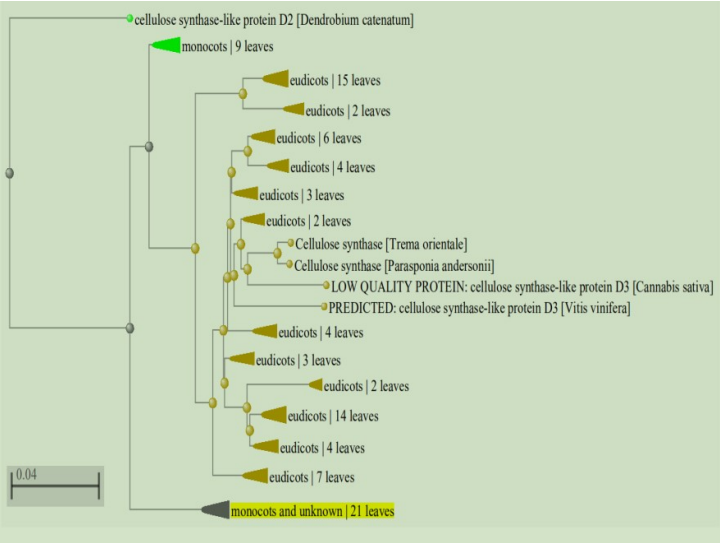


Figure 12: Cellulose synthase similarity tree

Predicted sequences obtained were 2. The most similarity was exhibited by Cellulose synthase-like protein D2 [*Dendrobium catenatum*] (Figure12).

4. Chalcone synthase similarity tree

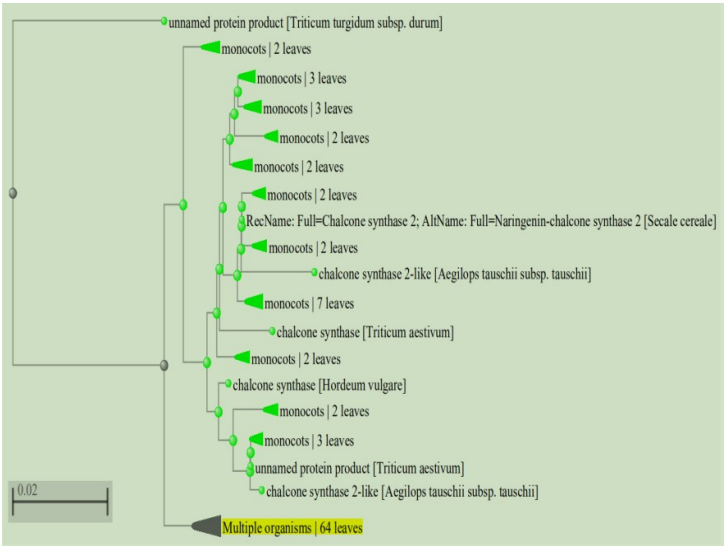


Figure 13: Chalcone synthase similarity tree

Predicted sequences obtained were none. The most similarity was exhibited by an unnamed protein product [*Triticum turgidum* subsp. *durum*] (Figure13).

5. COesterase similarity tree

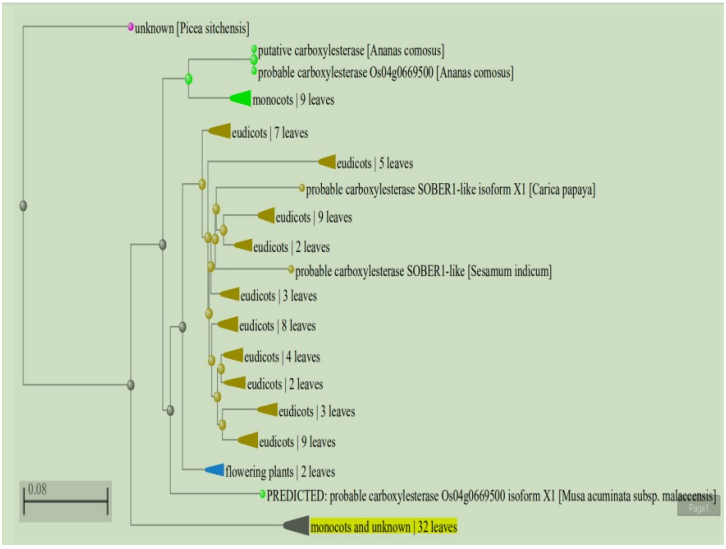


Figure 14: COesterase similarity tree

The predicted sequence obtained was1. The most similarity was exhibited by an unknown protein [*Picea sitchensis*] (Figure14).

6. DIOXN similarity tree

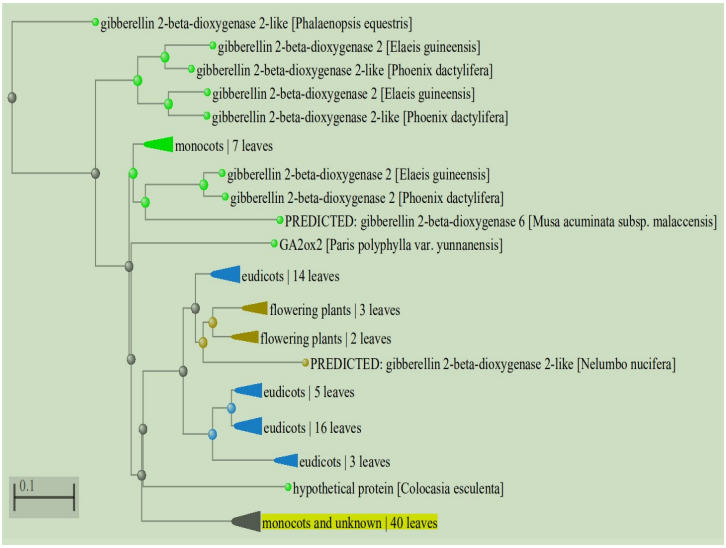


Figure 15: DIOXN similarity tree

Predicted sequences obtained were 2. The most similarity was exhibited by Gibberellin 2-beta-dioxygenase-2-like [*Phalaenopsis equestris*] proteins (Figure 15).

7. Epimerase similarity tree

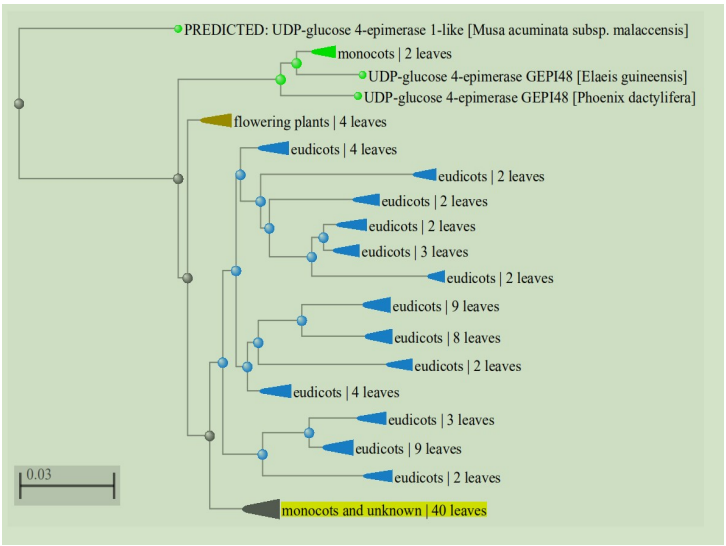


Figure 16: Epimerase similarity tree

The predicted sequence obtained was 1. The most similarity was exhibited by the predicted UDP-glucose-4-epimerase 1-like [*Musa acuminata* subsp. *malaccenis*] proteins (Figure16).

8. Methyl transferase similarity tree

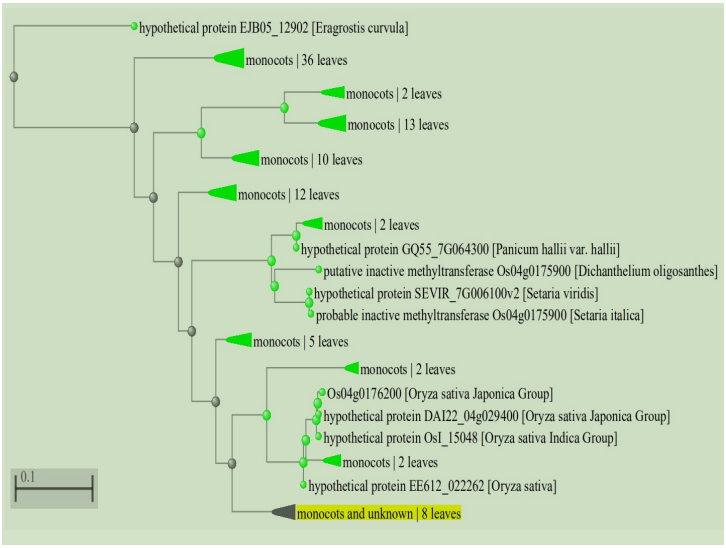


Figure 17: Methyl transferase similarity tree

Predicted sequences obtained were none. The most similarity was exhibited by hypothetical protein EJB05_12902 [*Eragrostis curvula*] proteins (Figure17).

9. P450 similarity tree

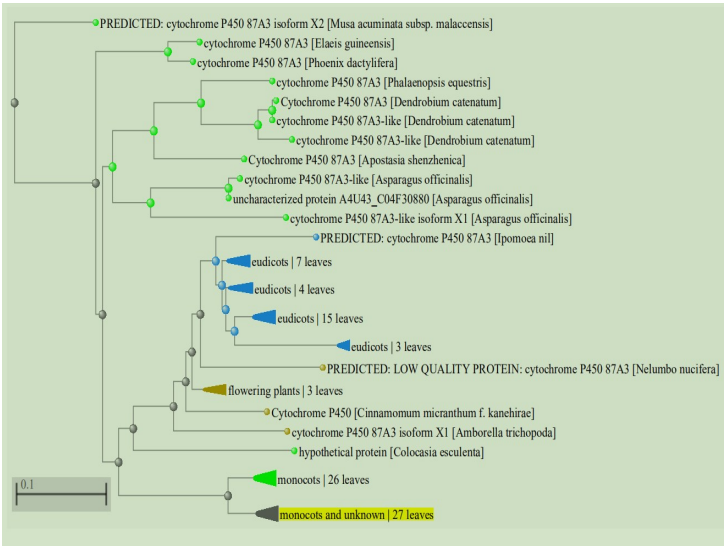


Figure 18: P450 similarity tree

Predicted sequences obtained were 3. The most similarity was exhibited by predicted Cytochrome p450 87A3 isoform X2[*Musa acuminata* subsp. *malaccenis*] proteins (Figure18).

10. Peptidase S10 similarity tree

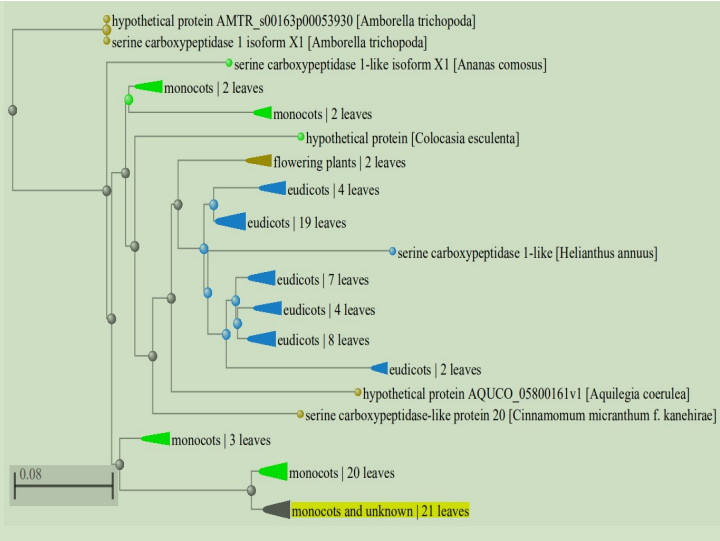


Figure 19: Peptidase S10 similarity tree

Predicted sequences obtained were 0. The most similarity was exhibited by hypothetical protein AMTR_s00163p00053930 [*Amborella trichopoda*] and serine carboxypeptidase 1 isoform X1 [*Amborella trichopoda*] proteins (Figure19).

11. UDPGT similarity tree

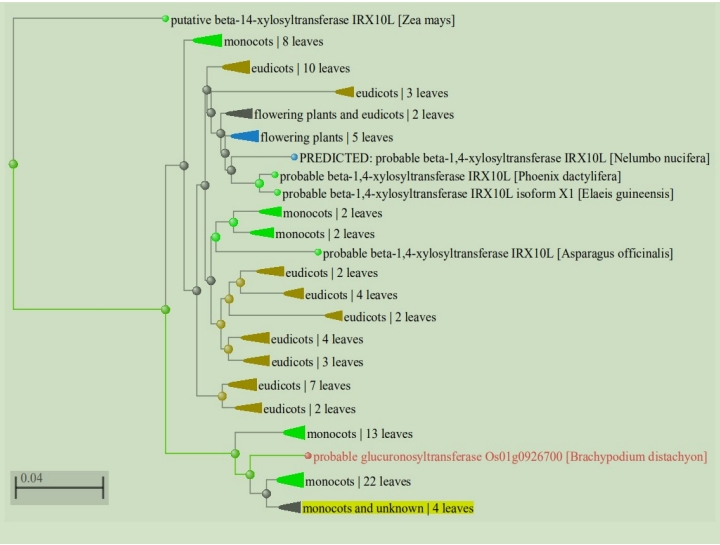


Figure 20: UDPGT similarity tree

The predicted sequence obtained was 1. The most similarity was exhibited by Putative beta 1-4-xylosyltransferase IRX10L [*Zea mays*] proteins (Figure 20).

Oryza sativa Indica

1. ADH

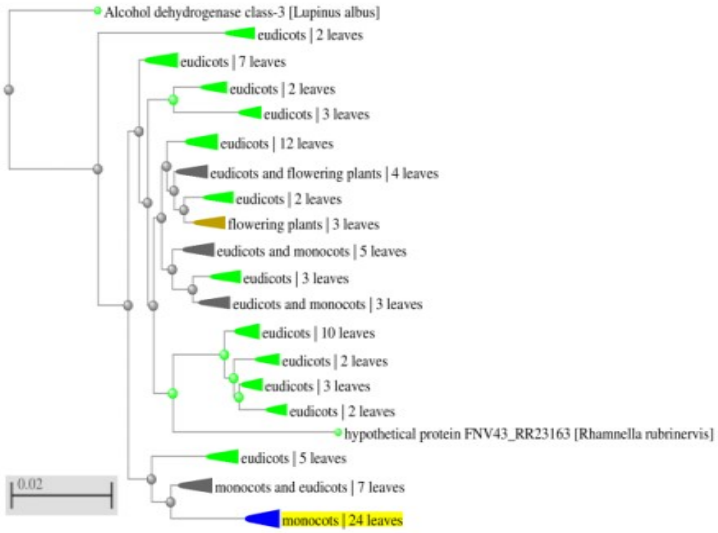


Figure 21: ADH similarity tree

The predicted sequences obtained were 0. The most similarity was exhibited by alcohol dehydrogenase class-3 [*Lupinus albus*] (Figure21).

2. Chalcone synthase

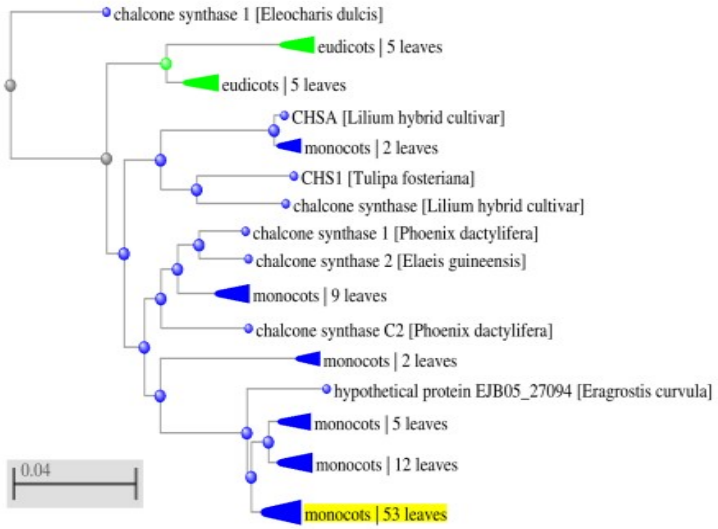


Figure 22: Chalcone synthase similarity tree

Predicted sequences obtained were 0. The most similarity was exhibited by chalcone synthase 1 [*Eleocharis dulcis*] (Figure22).

3. Methyltransferase

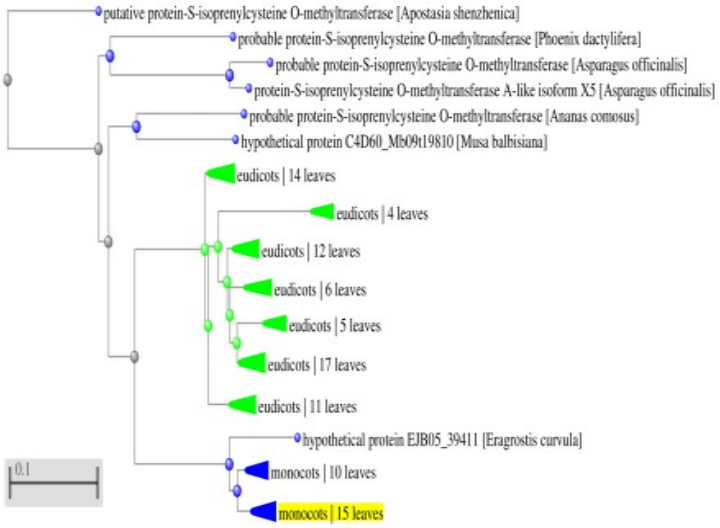


Figure 23: Methyltransferase similarity tree

Predicted sequences obtained were 0. The most similarity was exhibited by putative protein-S-isoprenylcysteine O-methyltransferase [*Apostasia shenzhenica*] (Figure23).

4. Epimrase

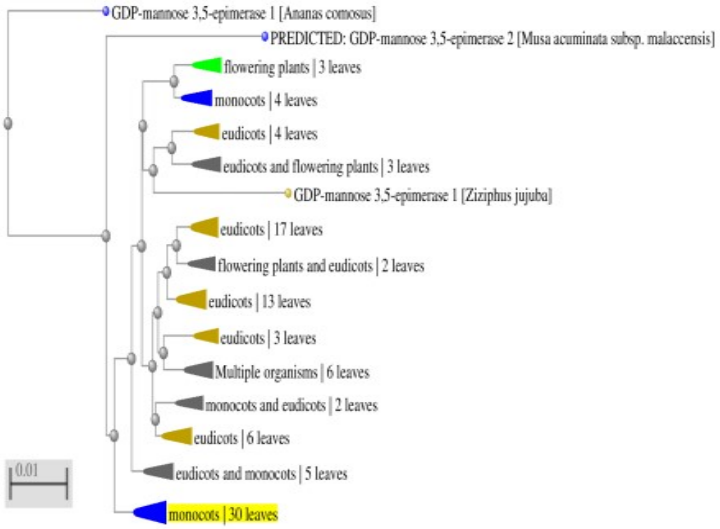


Figure 24: Epimerase similarity tree

The predicted sequence obtained was 1. The most similarity was exhibited by GDP-mannose 3,5-epimerase 1 [*Ananas comosus*] (Figure24).

5. Glycosyl transferase

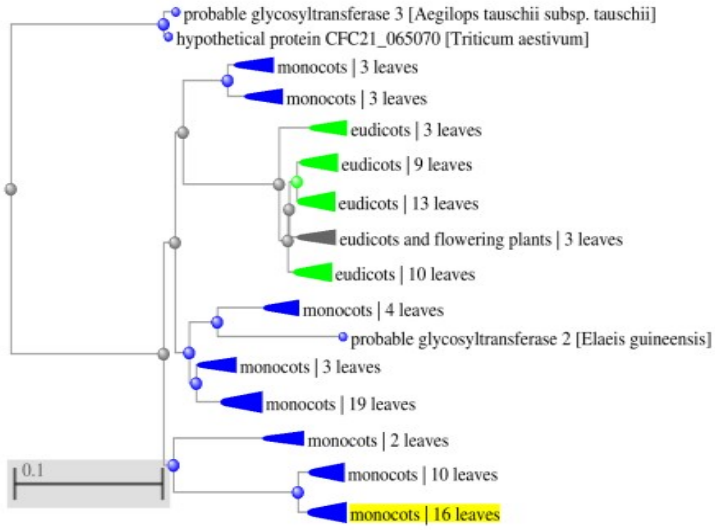


Figure 25: Glycosyl transferase similarity tree

Predicted sequences obtained were 0. The most similarity was exhibited by hypothetical protein CFC21_065070 [*Triticum aestivum*] (Figure25).

6. Terpene synthase

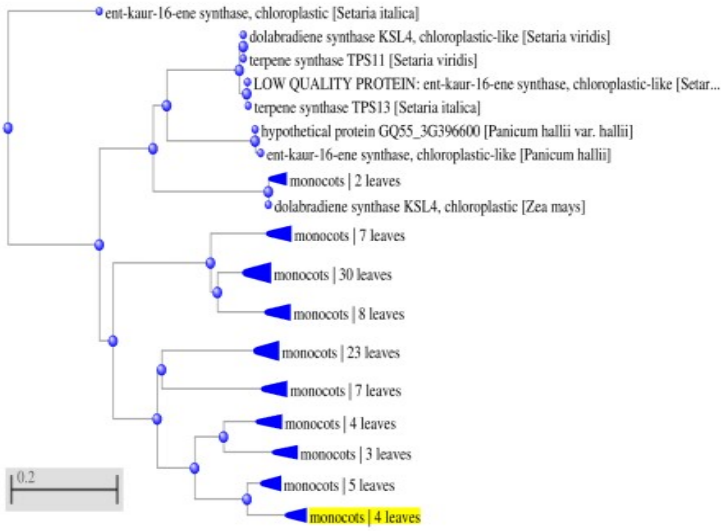


Figure 26: Terpene synthase similarity tree

The predicted sequence obtained was 1. The most similarity was exhibited by ent-kaur-16-ene synthase, chloroplastic [*Setaria italica*] (Figure26).

7. Acetyl transferase

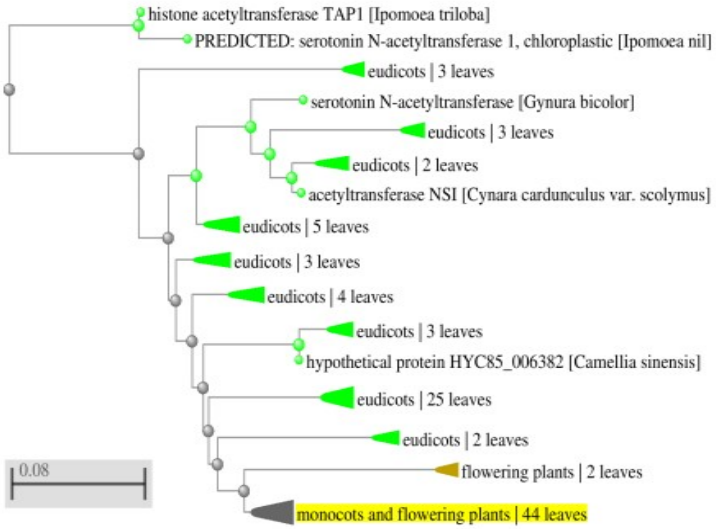


Figure 27: Acetyl transferase similarity tree

The predicted sequence obtained was 1. The most similarity was exhibited by histone acetyltransferase TAP1 [*Ipomoea triloba*] (Figure27).

8. Amino oxidase

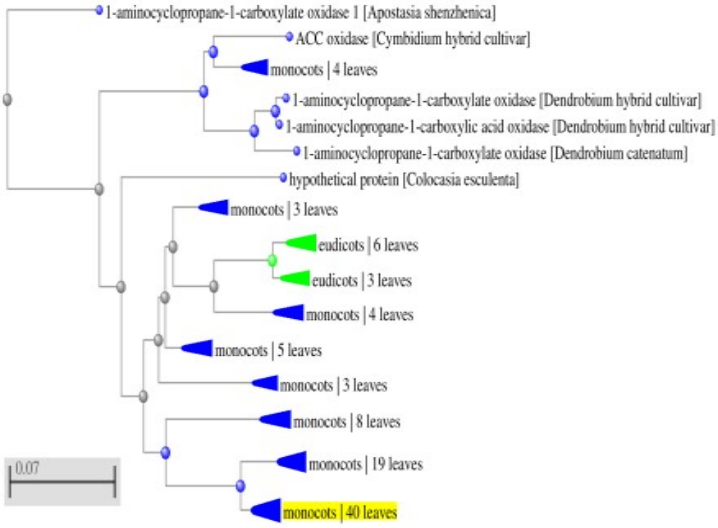


Figure 28: Amino oxidase similarity tree

Predicted sequences obtained were 0. The most similarity was exhibited by 1-aminocyclopropane-1-carboxylate oxidase 1 [*Apostasia shenzhenica*] (Figure28).

9. Amino transferase

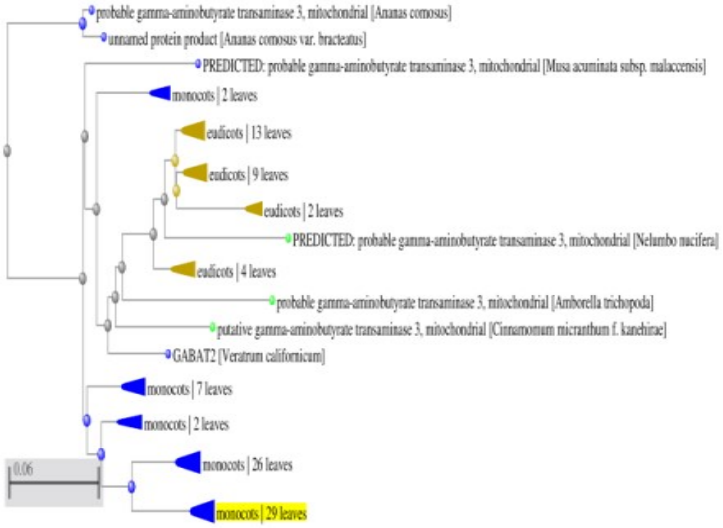


Figure 29: Amino transferase similarity tree

Predicted sequences obtained were 2. The most similarity was exhibited by probable gamma-aminobutyrate transaminase 3, mitochondrial [*Ananas comosus*] (Figure 29).

10.Sqs_psy

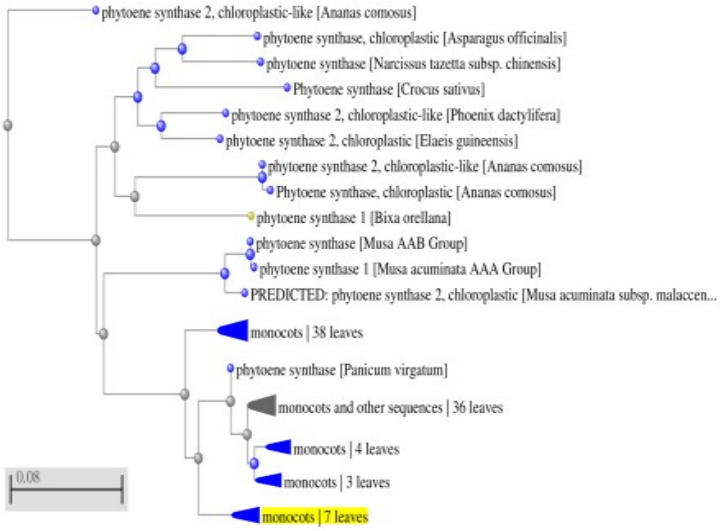


Figure 30: Sqs_psy similarity tree

The predicted sequence obtained was 1. The most similarity was exhibited by phytoene synthase 2, chloroplastic-like [*Ananas comosus*] (Figure30).

11.Cellulose synthase

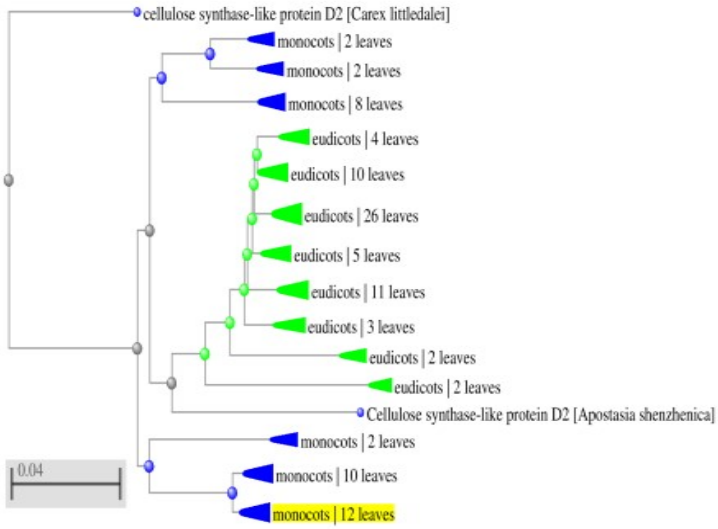


Figure 31: Cellulose synthase similarity tree

The predicted sequences obtained were 0. The most similarity was exhibited by cellulose synthase-like protein D2 [*Carex littledalei*] (Figure31).

6.4 To device search engines for metabolic gene clusters in the sequenced plant genomes.

WEKA tool utilized to prepare the Random Forest and SMO models. Given below are the values of the physicochemical properties that were obtained from the various web servers. These values were formatted as per the WEKA tool input requirements. The same ten physiochemical properties were considered for terpene synthases (Table 15) and non-terpene synthases (Table 16).

Table 15- Properties of terpenes synthases

Length	MW	PI	Instability index	Aliphatic index	hydropathicity	CHARGE AT PH7	TMindex	SOLUBILITY	EXTINCTION COEFFICIENT (M-1cm-1)
565	65753.22	7.3	42.15	90.21	-0.403	4.3	0.822113178	0.203	92740
591	69278.8	5.73	47.46	86.06	-0.397	-9.4	0.896357502	0.378	94020
600	70454.56	6.23	45.7	88.05	-0.435	-3.6	1.349081557	0.274	105400
600	70454.56	6.23	45.7	88.05	-0.435	-3.6	1.349081557	0.274	105400

877	101895.1	6.12	46.61	84.63	-0.349	-7.5	1.180332794	0.256	124460
547	63618.51	5.15	41.44	83.8	-0.305	-22.8	0.838713535	0.462	78800
66	7772.82	9.02	48.04	62.12	-0.403	4.5	2.910438886	0.53	15220
598	69619.93	5.66	53.79	80.55	-0.287	-9.3	1.109032653	0.348	81930
569	65400.88	6.47	54.06	89.77	-0.393	-0.6	0.520436714	0.284	86910
554	63918.89	5.29	48.65	84.53	-0.345	-15.3	2.046234991	0.351	76240
694	81629.53	5.63	47.92	75.61	-0.504	-10.9	0.670737685	0.341	128870
133	15477.56	6.19	43.63	96.62	-0.409	-0.9	0.584716619	0.464	12090
217	25462.37	9.1	53.4	91.24	-0.438	8	-0.203909394	0.46	34280
867	99914.52	5.53	41.44	89.99	-0.264	-17.5	0.858409326	0.293	142810
590	67327.76	5.76	49.53	89.86	-0.27	-10	1.365758382	0.227	74250
554	63840.77	5.17	49.56	84.19	-0.342	-17.5	2.100197873	0.386	81930
440	52389.23	6.39	43.22	92.82	-0.249	-0.8	1.264092932	0.208	87760
404	46890.12	5.36	56.84	96.31	-0.17	-12.8	1.280239308	0.392	66710
555	64014.61	5.14	48.03	88.58	-0.273	-19.6	0.825553048	0.382	78800
834	95150.3	6.16	51.3	95.05	-0.177	-6	1.768996677	0.142	112370
580	67132.52	5.71	46.96	92.66	-0.319	-8.9	1.161873104	0.275	90180
627	70784.54	5.75	46.42	85.57	-0.259	-9.6	0.36491398	0.253	93310
766	87634.18	5.93	45.55	82.11	-0.297	-8.6	0.255586887	0.159	195870
618	72123.77	5.77	52.02	93.51	-0.304	-11.1	0.335285879	0.255	89040
548	63930.16	5.7	45.49	85.42	-0.361	-9.6	0.486540152	0.342	66140
565	65424.64	5.53	50.51	89.4	-0.387	-11.3	0.762807887	0.31	87620
778	90819.41	6.62	42.91	90.23	-0.356	0.8	0.596019338	0.234	129440
286	32942.91	5.09	54.65	85.59	-0.111	-8.3	-0.247002893	0.335	50780
281	32667.26	5.27	43.98	84.98	-0.303	-6.3	0.657790795	0.407	53340
548	63938.25	5.14	41.36	90.49	-0.287	-19.3	1.324170781	0.464	83920
500	58396.71	5.55	47.77	86	-0.342	-13.6	0.927010442	0.449	70550
525	61232.43	5.5	54.58	93.58	-0.203	-11.8	1.594553807	0.376	73680
565	65733.37	5.58	55.73	92.99	-0.248	-12.8	1.260818837	0.344	83210
541	62888.97	5.62	47.74	89.56	-0.256	-10.7	0.60324774	0.289	83210
604	69262.16	5.99	48.13	86.26	-0.292	-5.9	0.706598795	0.237	66850
594	70604.9	5.88	39.77	87.63	-0.29	-7.5	1.161147039	0.288	114930
553	64186.44	5.33	52.44	92.53	-0.255	-14.1	0.960386037	0.378	88900
572	65857.87	5.89	41.8	92.05	-0.309	-9.4	1.146833258	0.252	99570
294	34011.31	4.64	32.97	88.2	-0.218	-19.2	0.373645295	0.49	53340
481	55012.66	5.46	45.42	89.6	-0.38	-11.3	0.740992701	0.395	58460
542	62242.07	5.65	41.4	90.33	-0.342	-9.9	0.867598471	0.311	83780
594	70510.91	5.99	43.24	87.17	-0.317	-5.7	1.308566163	0.296	120620
795	91379.57	5.63	39.59	84.67	-0.368	-18.3	0.687943275	0.339	146510
778	90685.24	6.35	45.39	91.98	-0.346	-2.7	0.770300519	0.241	125600
140	16107.01	4.74	39.17	108.5	0.444	-5.2	-1.683291437	0.717	23470
437	50454.89	5.19	33.31	92.88	-0.201	-16.1	1.205410883	0.337	58460
595	69414	5.75	48.29	88.22	-0.456	-10.6	1.513864776	0.315	115500

568	65844.55	6.09	40.89	97.48	-0.263	-4.9	1.576601596	0.267	105970
548	63731.34	5.31	42.16	90.86	-0.261	-14.5	1.337037043	0.408	80080
555	64262.6	5.5	43.75	92.76	-0.272	-11.2	1.123613699	0.35	82640
600	70439.88	5.28	57.32	90	-0.314	-20.8	0.886051557	0.41	81360
452	52622.85	5.47	52.97	89.73	-0.259	-12.6	0.996519076	0.267	60450
404	47513.22	5.17	43.91	86.16	-0.361	-14.5	0.624652969	0.405	76810
551	64630.08	5.29	39.64	92.74	-0.32	-16.5	0.785785431	0.364	109810
750	86619.28	6.01	41.54	87.93	-0.273	-6.4	0.932689807	0.233	133280
739	84559.67	5.81	44.13	86.6	-0.267	-7.9	1.344021108	0.19	139540
250	28930.1	5.56	41.02	87.84	-0.216	-4	1.23472925	0.321	70410
763	87908.7	5.69	46.98	78.1	-0.33	-13.5	1.250762272	0.286	199710
781	88760.08	5.36	48.77	88.04	-0.238	-19.1	0.469146959	0.278	121760
568	66316.95	5.33	44.25	96.64	-0.234	-15	1.376681164	0.416	99850
314	36128.69	8.12	37.48	92.64	-0.108	3.9	0.376720858	0.205	60880
93	10525.7	5.26	44.43	73.33	-0.391	-4.7	0.808595503	0.689	10810
548	63367.11	5.14	41.11	93.27	-0.319	-19.7	1.60034997	0.417	82640
607	70652.01	6.12	48.18	84.7	-0.326	-4.7	0.356290161	0.209	100990
622	72384.57	6.46	43.75	83.55	-0.42	-0.6	-0.390114797	0.255	97150
547	63657.75	5.32	47.82	88.79	-0.329	-18.8	1.381387174	0.33	74960
498	58329.63	5.2	45.11	92.43	-0.289	-16.2	0.736122591	0.305	89470
507	59375.16	6.1	45.19	91.34	-0.314	-3.9	1.082549472	0.19	75530
516	60486.14	5.32	46.07	93.18	-0.282	-15	0.706442552	0.273	89470
598	69852.75	5.66	42.67	90.69	-0.36	-11.1	1.258971302	0.313	123180
531	62269.33	5.7	47.08	92.35	-0.285	-8.8	1.016783632	0.214	95870
614	71947.62	5.74	35.78	93.34	-0.267	-9.5	1.56216211	0.242	103550
597	69633.59	5.78	41.62	89.98	-0.37	-9.1	1.228192776	0.301	119340
468	54778.76	5.43	39.89	91.71	-0.29	-10.6	0.835534114	0.327	86340
584	67733.64	5.82	43.42	94.83	-0.226	-9	0.666495649	0.233	97150
768	87865.15	5.6	41.38	88.89	-0.282	-12.5	0.848965855	0.26	142100
740	84758.9	5.86	44.22	85.96	-0.291	-7.7	1.101878399	0.209	136980
836	96737.71	5.53	49.45	77.97	-0.346	-19.4	1.631039497	0.352	144660
814	93079.21	6.71	50.88	82.35	-0.287	1.7	0.383663285	0.065	121620

Table 16- Properties of non-terpene synthases

LENGTH	MW	PI	Instability index	Aliphatic index	hydropathicity	CHARGE AT PH7	TMindex	SOLUBILITY	EXTINCTION COEFFICIENT (M-1cm-1)
1084	122502.11	7.24	38.87	87.03	-0.199	4.5	0.900667394	0.292	206680
1026	115798.12	6.25	41.31	82.07	-0.205	-4.7	1.263810521	0.273	181360
1065	119682.54	7.66	35.91	86.11	-0.181	4.9	0.982290643	0.274	192740
1081	122236.88	6.53	39.44	85.39	-0.236	-0.6	1.011707122	0.205	190320

985	111521.44	6.77	40.79	89.54	-0.104	1.1	0.912036192	0.271	168700
1140	128008.85	7.91	40.14	82.28	-0.212	8.2	1.180241658	0.171	190180
1069	120861.99	7.06	40.45	86.26	-0.209	3.8	0.831928881	0.3	209240
1049	119599.39	8.22	37.6	80.74	-0.284	10	0.86027541	0.291	209240
1145	128360.18	7.5	42.26	81.33	-0.22	6	0.784544687	0.197	197150
1084	122069.43	7.42	41.3	84.88	-0.205	5.2	0.659178264	0.267	210520
380	41408.62	6.47	30.06	86.16	-0.064	-0.5	0.670661042	0.381	26030
309	33860.18	6.22	36.3	90.78	0.04	-1.9	0.493442821	0.454	10810
380	41198.61	6.13	33.79	91.79	0.064	-3.3	0.198235623	0.368	24750
380	40924.91	5.87	29.6	84.11	-0.007	-5.5	0.045189479	0.366	27310
379	41110.24	5.98	30.3	82.53	-0.098	-5	0.595666777	0.4	23470
380	41408.62	6.47	30.06	86.16	-0.064	-0.5	0.670661042	0.381	26030
379	41076.04	5.97	29.68	81.24	-0.104	-5	0.296336564	0.378	23470
156	16655.84	8.41	27.26	78.72	0.047	1.8	1.170307154	0.576	33000
712	80138.08	5.98	35.69	81.28	-0.3	-6.6	0.872852738	0.243	102840
650	73783.05	8.99	37.16	80	-0.327	13.9	1.223433185	0.254	99710
776	86690.84	6.68	44.56	77.47	-0.357	1.4	0.589897032	0.218	103410
741	83233.15	5.8	37.22	78.87	-0.378	-11.4	0.364391106	0.339	85060
677	76673.67	6.98	34.3	77.52	-0.299	2.7	0.216668659	0.285	94590
681	76710.63	6.36	32.79	83.51	-0.243	-1.6	0.625505754	0.256	95870
687	77532.68	5.7	42.06	82.39	-0.275	-9.6	0.947478462	0.335	97150
300	33965.29	5.57	34.18	75.6	-0.541	-5.4	0.827052749	0.575	41960
460	51593.46	5.81	44.25	75.87	-0.38	-5.6	0.257888375	0.261	50780
92	10320.78	6.25	60.71	72.07	-0.293	-0.4	1.380046533	0.372	12660
493	54626.32	9.73	36.09	99.45	-0.006	27.6	0.863516331	0.326	55330
351	39157.71	6.14	27.47	82.79	-0.344	-1.7	0.144856232	0.448	46370
301	33597.33	5.73	31.57	85.48	-0.234	-3.2	-0.193724295	0.394	31150
351	38910.38	6.27	24.77	83.59	-0.27	-1.3	0.636286818	0.358	43810
281	30008.7	8.24	41.71	100.93	0.163	3.2	1.407278266	0.552	16500
377	42758.49	5.84	40.17	69.1	-0.493	-5.6	0.637837701	0.412	56470
667	75225.54	6.04	29.61	79.48	-0.399	-4.9	0.546089228	0.404	83210
350	38382.44	6.27	32.45	84.37	-0.273	-1.3	0.858121222	0.351	50210
437	48535.75	9.95	47.8	83.94	-0.242	29	-0.309460795	0.397	44520
351	38617.69	5.52	31.18	89.12	-0.222	-7.8	0.508032385	0.405	41960
395	43115.72	6.08	32.16	92.08	-0.074	-3.3	0.819543613	0.371	31720
388	42516.3	6.21	38.86	91.96	-0.055	-1.5	0.881346422	0.355	36840
392	42982.41	5.98	37.17	93.78	-0.167	-2.8	1.148594812	0.386	32430
395	43617.14	6.01	42.91	91.11	-0.228	-2.2	1.070041855	0.421	31150
385	42414.81	5.9	36.67	93.71	-0.168	-3	1.518460632	0.417	32430
389	42552.3	6.72	39.38	94.24	-0.06	0.7	1.083524965	0.36	34280
389	42874.38	5.96	35.53	89	-0.143	-3.3	1.10308249	0.389	39400
390	43044.58	6.24	42.98	89	-0.151	-1.8	0.485447156	0.385	34280
444	49586.25	9.36	52.18	85.59	-0.277	17.6	1.474520593	0.313	53910

753	83075.18	8.88	42.77	99.38	0.318	12.7	0.92244835	0.171	153340
267	31616.4	8.95	53.02	93.07	-0.436	6.8	1.616720715	0.354	45090
216	24919.7	5.63	46.46	98.7	-0.279	-3.9	1.401964296	0.605	39400
222	25205.81	8.64	44.51	84.41	-0.432	4.1	1.172286101	0.45	33000
213	24224.1	6.1	41.4	100.7	-0.182	-1.7	0.679953824	0.454	30440
215	25112.96	5.74	41.82	92.84	-0.39	-3.7	2.15461442	0.542	41960
252	28893.1	7.85	40.72	91.27	-0.32	3.8	0.547764443	0.315	31720
229	27079.29	6.26	48.15	93.19	-0.35	-1.2	0.328859498	0.48	38690
225	26110.91	5.27	35.35	91.82	-0.245	-7.6	1.099857916	0.67	30440
413	44846.05	6.27	34.19	80.94	-0.128	-1.2	-0.210865644	0.175	43810
808	91829.22	7.57	45.01	74.59	-0.471	5.3	0.074829825	0.402	113510
480	52929.72	5.81	50.75	92.19	-0.156	-6.3	0.788409611	0.346	50070
637	69321.2	6.27	40.56	83.58	-0.241	-2.3	0.473911369	0.336	69130
465	50942.5	5.57	36.31	100.3	0.075	-9.1	1.18124174	0.235	47510
615	68344.81	5.44	46.76	85.92	-0.239	-16.4	0.390192582	0.32	88760
455	51728.48	5.6	49.87	92.48	-0.162	-8.5	0.434496106	0.281	65290
496	55799.42	6.45	47.74	94.03	-0.194	-0.7	0.417239355	0.316	73540
270	30278.43	9.24	37.61	89.59	-0.147	9.3	1.186383173	0.416	10810
435	48122.24	5.77	38.49	96.37	-0.034	-7.5	1.037726173	0.259	53200
440	49480.12	9.24	34.52	85.98	-0.333	16.8	0.45707948	0.358	55900
499	56846.42	8.52	35.27	92.28	-0.122	6	1.384199876	0.19	45800
517	57643.79	8.59	35.49	93.13	0.072	6.7	1.729119888	0.23	84210
472	53785.47	9.12	41.73	89.87	-0.074	10	1.312806278	0.14	55190
508	57926.9	8.74	38.62	88.29	-0.263	8.2	1.459025046	0.168	81790
482	55175.94	9.19	55.31	90.79	-0.227	12.3	1.278095802	0.1	66000
506	56932.98	9.06	44.08	94.55	-0.064	9.8	1.369793642	0.183	65290
522	59425.32	9.31	34.88	98.81	-0.076	15.7	1.137186263	0.146	60310
541	61347.53	8.76	35.62	86.52	-0.193	10.3	0.953382086	0.169	60880
495	56569.39	8.12	41.09	94.53	-0.036	4.2	1.390142197	0.109	53910
515	58570.39	9.12	34.71	102.95	-0.09	14.4	1.058443703	0.17	69130
187	21411.9	8.43	24.15	77.17	0.139	2.6	1.949221695	0.489	17210
447	41506.64	4.69	31.06	74.12	0.195	-8.1	0.425824918	0.473	7680
302	31646.73	4.62	36.49	91.09	0.114	-11.9	-0.768212514	0.658	5120

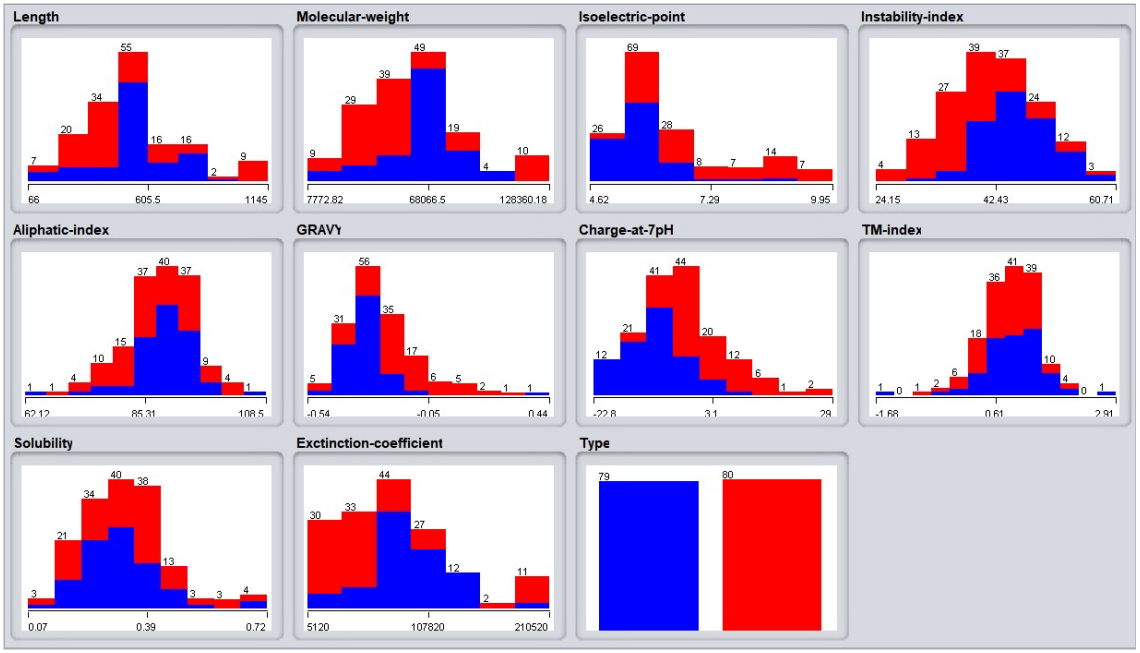


Figure 32: Display of the attribute distribution across terpene synthases and non terpene synthases

In this study, we have used 66% percentage split testing on the data set. We calculated the overall accuracy, true positive rate (TP), false positive rate (FP), precision, recall, Mathew's correlation coefficient (MCC) [A8], and receiver operating characteristic (ROC) along with the confusion matrices of the attributes, threshold curves and attribute vise visualization of the classification (Figure 32).

Table 17- Summary of SMO model developed

Correctly Classified Instances	52	96.2963 %
Incorrectly Classified Instances	2	3.7037 %
Kappa statistic	0.926	
Mean absolute error	0.037	
Root mean squared error	0.1925	
Relative absolute error	7.4048 %	
Root relative squared error	38.475 %	
Total Number of Instances	54	

SMO model developed shows an appreciable 96.2963 % accuracy. This means that there is a 96.2963 % chance that the model accurately predicts the input to be a terpene synthase or a non-terpene synthase. Table 17 shows summary of SMO model developed.

Table 18- Summary of Random Forest model developed

Correctly Classified Instances	51	94.4444 %
Incorrectly Classified Instances	3	5.5556 %
Kappa statistic	0.8889	
Mean absolute error	0.1646	
Root mean squared error	0.229	
Relative absolute error	32.9145 %	
Root relative squared error	45.7909 %	
Total Number of Instances	54	

The random forest model (Table18) gives an accuracy of 94.4444 % percent for appropriate classification with 51 correctly classified instances and 3 incorrectly classified ones out of a total of 54.

Table 19-Perfromance of SMO model developed by class

	TP Rate	FP Rate	Precision	Recall	MCC	ROC Area	Class
	1.000	0.071	0.929	1.000	0.929	0.964	Terpene
	0.929	0.000	1.000	0.929	0.929	0.964	Non terpene
Weighted Avg.	0.963	0.034	0.966	0.963	0.929	0.964	

Table 20- Performance of the Random Forest developed by class

	TP Rate	FP Rate	Precision	Recall	MCC	ROC Area	Class
	0.962	0.071	0.926	0.962	0.889	0.992	Terpene
	0.929	0.038	0.963	0.929	0.889	0.992	Non terpene
Weighted Avg.	0.944	0.054	0.945	0.944	0.889	0.992	

Both SMO and random forest models showed significantly low FP rates (0.034, 0.054) and high precision, TP rate, recall, MCC hovering around 0.9. The closer these values are to 1, the more accurate and reliable the results are (Table 19 and 20).

Table 21-Confusion matrix of the SMO model

Classified as→	Terpene	Non terpene
Terpene	26	0
Non terpene	2	26

Table 22-Confusion matrix of the random forest model

Classifed as →	Terpene	Non terpene
Terpene	25	1
Non terpene	2	26

The confusion matrix summarizes the accuracy by illustrating how many entries were classified under the appropriate class and how many were not (Table 21 and 22).

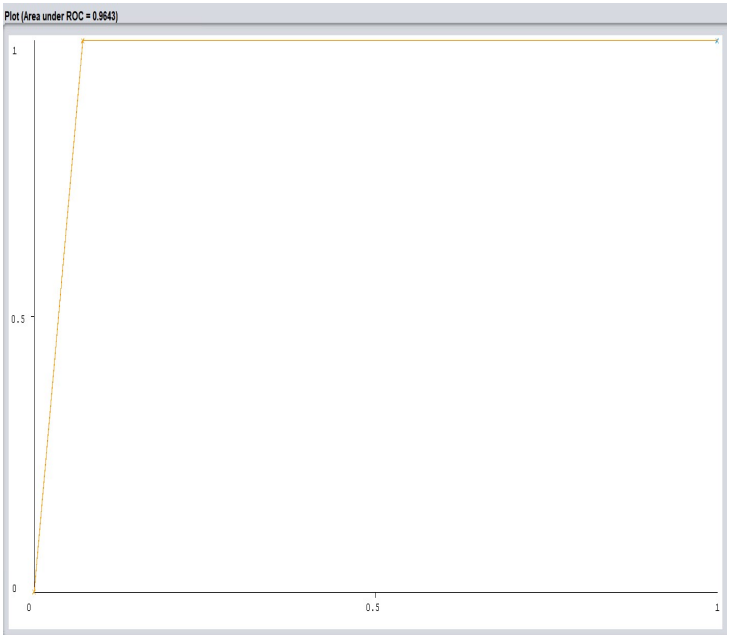


Figure 33: ROC threshold curve of terpene class as per SMO. With an area under 0.9643

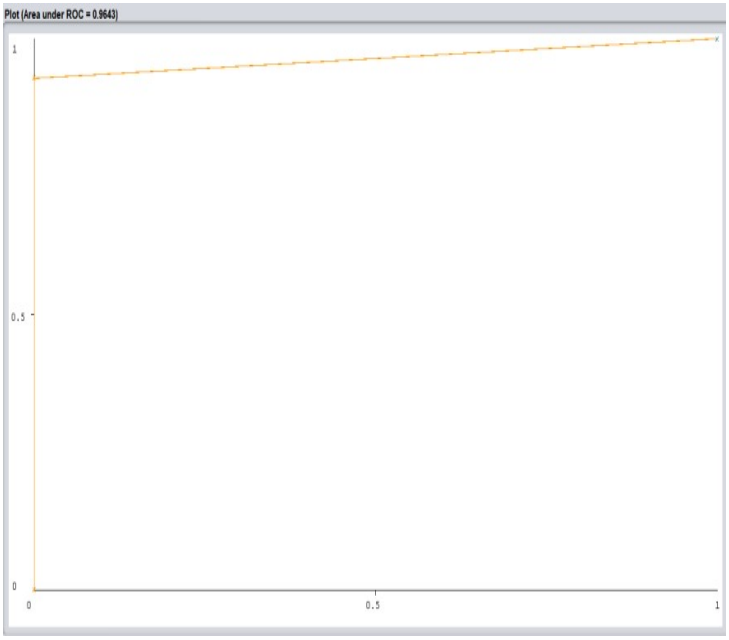


Figure 34: ROC threshold curve of non-terpene class as per SMO. With an area under 0.9643

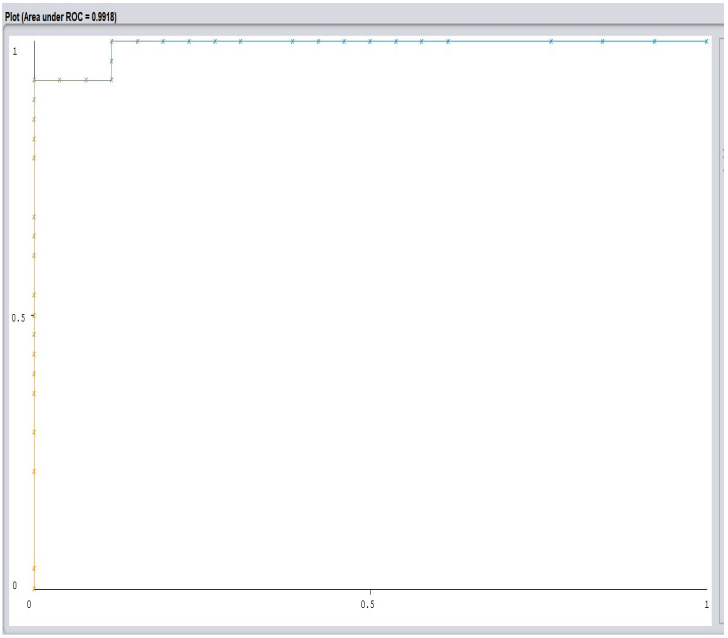


Figure 35: ROC threshold curve of terpene class as per random forest. With an area under 0.9918

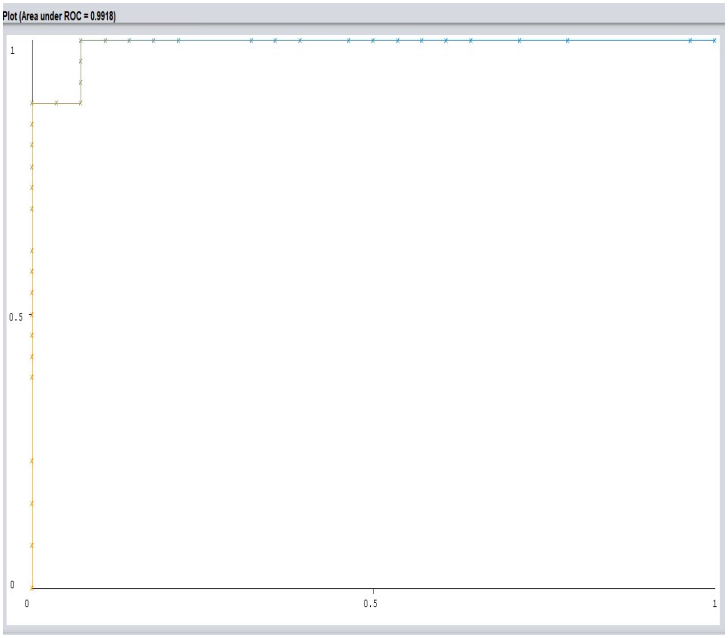


Figure 36: ROC threshold curve of non-terpene class as per random forest. With an area under 0.9918

ROC area (Plot of False positive rate (X) vs True positive rate (Y)) closer to 1 indicates the most accurate classifier. Both models developed (SMO and random forest) show excellent classifier accuracy indicated by the ROC areas under the curve (Figure 33 to 36).

In this study we presented a way to predict the gene clusters in plants through machine learning approach. Softwares and servers utilized in present study provides the way for prediction of gene clusters in variety of the plants. There are several existing tools that predict metabolic gene clusters from microbes, such as Cluster Finder (Cimermancic et al., 2014), SMURF (Khaldi et al., 2010), and antiSMASH (Weber et al., 2015). However, they are trained on predicting gene clusters from bacteria and fungi, which predominantly make polyketides, nonribosomal peptides, and sugar derivatives (Cimermancic et al., 2014). On the other hand, plants predominantly make terpenoids, lignan, and alkaloids (Gunatilaka, 2008; Wink, 2010). Most of the methods which are developed so far mainly focused on the bacteria and fungi metabolic gene clusters. Key challenges and opportunities existed in extension of these tools in more complex genomes. The development of (machine-learning) algorithms that predict BCGs will allows more powerful predictions of the natural product structural diversity encoded in diverse BGCs in plants. plantiSMASH, a antiSMASH derivative uses pHMMs and CD-Hit clustering for plant specific cluster prediction [Kautsar et al., 2017] and PhytoClust employs a hidden Markov model (HMM) search algorithm to

detect genes that are specific for certain types of known clusters, i.e. the tool is confined to already characterized gene cluster types [Töpfer et al., 2017]. In our present work, we have tried to devised an algorithm based on the physiochemical properties of the BGCs forming terpenes synthases using SMO and Random Forest models. These techniques were found to be 96 and 94 percent accurate respectively.

CHAPTER: 7
CONCLUSION

CONCLUSION

The hypothesis that the various signature gene products and the corresponding tailoring gene products of the gene clusters of *Oryza sativa Japonica* and *Indica* undergo protein-protein interaction was substantiated by the reliable docking results as discussed above. The dendrogram generated for each signature gene product displays the evolutionary relationship with similar proteins from other species or the predicted/hypothetical proteins suggested by the algorithm.

The models developed with the help of both classifiers (RF and SMO) as a part of this study to predict BGCs in plants, mainly based on the distribution of physicochemical features of the respective protein products of the said BGCs (terpene synthases and non-terpene synthases) showed significantly positive and accurate results of classification. From just protein sequence information, machine learning methods and classifiers (SMO and RF models) have been created to predict domain swapping at the genome level. The two techniques were found to be 96 percent and 94 percent accurate; respectively. Research in the field of plant secondary metabolic gene clusters has seen exponential growth over the past few years. It is only poised to extend farther as more discoveries of and about gene clusters emerge at an accelerated rate, thanks to high throughput screening methods. Combining systematic genome mining and functional analysis of candidate clusters along with artificial intelligence and machine learning makes the discovery of new pathways, enzymes, and chemistries well within the realm of possibility. The impact that this accelerated development can have on the way we approach everything from farming to drug discovery to nutrition will be unprecedented. Application of such technologies into staple, major food crop industry could mean a significant quality of life change for the world as a whole through better farming and nutrition.

CHAPTER: 8
FUTURE ASPECTS

FUTURE ASPECTS

Advances in genomics and bioinformatics have re-energized natural product research, allowing it to become a more targeted and systematic undertaking based on genomic data. Rapid and low-cost sequencing technologies, along with sophisticated bioinformatics tools, have expanded our understanding genomic and structural diversity, revealing practically limitless natural product discovery possibilities.

The metabolite biosynthesis pathways are fundamentally a part of the metabolic network of their hosts, which are complex processes involving internal sensitive and sophisticated catalysis and control as well as exterior interactions with the metabolism milieu. As a result, identifying and optimizing biosynthetic routes typically entails complex data interpretation and design activities that cannot be completed without the use of computational tools. The discovery of metabolites through bioinformatics study of BGCs is very promising and has been experimentally confirmed numerous times. However, identifying even a single gene in pathways where biosynthetic genes are not grouped can be difficult. When comparing omics data, there are sometimes too many hits to test experimentally, and ranking algorithms can be unreliable. Retro biosynthesis provides an alternate technique, however the fundamental design guidelines, such as quantifying enzyme candidates for certain reactions based on promiscuity and specificity, are still unclear. Automation combined with synthetic biology and machine learning can speed up the design-build-test cycle and help generate models with more accuracy, in addition to incorporating new scientific results and upgrading algorithms for natural pathway identification or artificial pathway development. Catalysis and regulatory components of biosynthetic pathways can be optimized for increased production. For heterologous biosynthesis, successful expression and acceptable performance of biosynthetic enzymes are essential, but effective regulation can balance the metabolic flow and improve biosynthesis efficiency. Factors that may affect the productivity of target molecules can be improved individually with the use of computational techniques. In addition to developing improved models and algorithms to make the process more accurate and efficient, scoring approaches to priorities the rate-limiting factor for improvement in a specific pathway design would be highly valuable.

Apart from genome-guided reinvestigation of well-known producers, the discovery of unusual sources such higher species promises to yield fascinating results. Similarly, studying natural products in their natural environment (as mediators of pairwise or complex organismal interactions) will not only aid in the understanding of natural product functions,

but will also lead to the discovery of novel drug candidates based on secondary metabolite ecological functions. The ability to predict natural product activity solely based on BGC sequences will allow researchers to priorities clusters that are most likely to produce molecules with the desired activity, reducing the number of times natural product discovery bottlenecks must be overcome in order to find a new active compound.

More powerful predictions of natural product structural diversity encoded in diverse BGCs will be possible thanks to the development of (machine-learning) algorithms that predict substrate specificities of key enzymes like terpene synthases and the systematic construction of pHMMs for automated subclassification of complex enzyme families like cytochrome P450s and glycosyltransferases. Furthermore, a complete evolutionary genomic examination of gene clustering, including BGC birth, death, and change processes, will help us better understand how BGCs support natural product diversification during evolution. As more plant BGCs are empirically defined, the algorithms will advance in tandem with the knowledge gathered, allowing for the development of more thorough class-specific cluster recognition methods; also, it will become obvious what constitutes a true BGC. Finally, as scientists decipher the complexity of tissue-specific and variably timed gene expression in plant biosynthetic pathways, we will get a better understanding of how to use coexpression data to forecast biosynthetic pathways. As a result, a better understanding of evolution's remarkable successes in generating an enormous diversity of powerful bioactive molecules will hopefully enable biological engineers to mimic nature's strategies and deliver a slew of new molecules for agricultural, cosmetic, dietary, and clinical applications.

CHAPTER 9:
REFERENCES

REFERENCES

1. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.
2. Bharadwaj, R., Kumar, S. R., Sharma, A., & Sathishkumar, R. (2021). Plant metabolic gene clusters: evolution, organization, and their applications in synthetic biology. *Frontiers in Plant Science*, 1573.
3. Bienert, S., Waterhouse, A., de Beer, T. A., Tauriello, G., Studer, G., Bordoli, L., &Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. *Nucleic acids research*, *45*(D1), D313-D319.
4. Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., & Weber, T. (2013). antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research*, *41*(W1), W204-W212.
5. Boutanaev, A. M., Moses, T., Zi, J., Nelson, D. R., Mugford, S. T., Peters, R. J., & Osbourn, A. (2015). Investigation of terpene diversification across multiple sequenced plant genomes. *Proceedings of the National Academy of Sciences*, *112*(1), E81-E88.
6. Boycheva, S., Daviet, L., Wolfender, J. L., & Fitzpatrick, T. B. (2014). The rise of operon-like gene clusters in plants. *Trends in plant science*, *19*(7), 447-459.
7. Castillo, D. A., Kolesnikova, M. D., & Matsuda, S. P. (2013). An effective strategy for exploring unknown metabolic pathways by genome mining. *Journal of the American Chemical Society*, *135*(15), 5885-5894.
8. Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H., & Gray, J. J. (2011). Benchmarking and analysis of protein docking performance in Rosetta v3. 2. *PloS one*, *6*(8), e22477.
9. Cheng, F., & Cheng, Z. (2015). Research progress on the use of plant allelopathy in agriculture and the physiological and ecological mechanisms of allelopathy. *Frontiers in plant science*, *6*, 1020.
10. Chu, H. Y., Wegel, E., & Osbourn, A. (2011). From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *The Plant Journal*, *66*(1), 66-79.
11. Chu, H. Y., Wegel, E., & Osbourn, A. (2011). From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *The Plant Journal*, *66*(1), 66-79.

12. Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Brown, L. C. W., Mavrommatis, K., &Birren, B. W. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, *158*(2), 412-421.
13. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, *221*(4612), 709-713.
14. Conway, K. R., & Boddy, C. N. (2012). ClusterMine360: a database of microbial PKS/NRPS biosynthesis. *Nucleic acids research*, *41*(D1), D402-D407.
15. Craig, R. A., & Liao, L. (2007). Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices. *Bmc Bioinformatics*, *8*(1), 1-12.
16. de Albuquerque, M. B., dos Santos, R. C., Lima, L. M., de Albuquerque Melo Filho, P., Nogueira, R. J. M. C., Da Câmara, C. A. G., & de Rezende Ramos, A. (2011). Allelopathy, an alternative tool to improve cropping systems. A review. *Agronomy for Sustainable Development*, *31*(2), 379-395.
17. Dixon, R. A., Achnine, L., Deavours, B. E., &Naoumkina, M. (2006). Metabolomics and gene identification in plant natural product pathways. In *Plant Metabolomics* (pp. 243-259). Springer, Berlin, Heidelberg.
18. Duhovny, D., Nussinov, R., & Wolfson, H. J. (2002, September). Efficient unbound docking of rigid molecules. In *International workshop on algorithms in bioinformatics* (pp. 185-200). Springer, Berlin, Heidelberg.
19. Dutartre, L., Hilliou, F., &Feyereisen, R. (2012). Phylogenomics of the benzoxazinoid biosynthetic pathway of Poaceae: gene duplications and origin of the Bx cluster. *BMC Evolutionary Biology*, *12*(1), 1-19.
20. Fan, P., Wang, P., Lou, Y. R., Leong, B. J., Moore, B. M., Schenck, C. A., Combs, R., Cao, P., Brandizzi, F., Shiu, S. H., & Last, R. L. (2020). Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity. *eLife*, *9*, e56717. https://doi.org/10.7554/eLife.56717
21. Field, B., & Osbourn, A. E. (2008). Metabolic diversification—independent assembly of operon-like gene clusters in different plants. *Science*, *320*(5875), 543-547.
22. Field, B., Fiston-Lavier, A. S., Kemen, A., Geisler, K., Quesneville, H., & Osbourn, A. E. (2011). Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proceedings of the National Academy of Sciences*, *108*(38), 16116-16121.

23. Frey, M., Chomet, P., Glawischnig, E., Stettner, C., Grün, S., Winklmair, A., ... & Simcox, K. (1997). Analysis of a chemical plant defense mechanism in grasses. *Science*, *277*(5326), 696-699.
24. Frey, M., Kliem, R., Saedler, H., &Gierl, A. (1995). Expression of a cytochrome P450 gene family in maize. *Molecular and General Genetics MGG*, *246*(1), 100-109.
25. Gao, C., Mulder, D., Yin, C., & Elliot, M. A. (2012). Crp is a global regulator of antibiotic production in Streptomyces. *MBio*, *3*(6).
26. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. (2005). *Protein Identification and Analysis Tools on the ExPASy Server;* (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press, pp. 571-607.
27. Ghosh, B., Ali, M.N. &Gantait, S. (2016). Response of Rice under Salinity Stress: *A Review Update. J. Res. Rice*, 4(2): 2–9.
28. Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., & Baker, D. (2003). Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology*, *331*(1), 281-299.
29. Gunatilaka, A. L. (2008). Natural Products in Plants: Chemical Diversity. *Wiley Encyclopedia of Chemical Biology*, 1-17.
30. Guo, L., Qiu, J., Ye, C., Jin, G., Mao, L., Zhang, H., ... & Lin, Z. (2017). Echinochloa crus-galli genome analysis provides insight into its adaptation and invasiveness as a weed. *Nature communications*, *8*(1), 1-10.
31. Gupta, C., Ramegowda, V., Basu, S., & Pereira, A. (2021). Using Network-Based Machine Learning to Predict Transcription Factors Involved in Drought Resistance. *Frontiers in Genetics*, *12*.
32. Haralampidis, K., Bryan, G., Qi, X., Papadopoulou, K., Bakht, S., Melton, R., & Osbourn, A. (2001). A new class of oxidosqualenecyclases directs synthesis of antimicrobial phytoprotectants in monocots. *Proceedings of the National Academy of Sciences*, *98*(23), 13431-13436.
33. Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R., & Warwicker, J. (2017). Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics (Oxford, England)*, *33*(19), 3098–3100. https://doi.org/10.1093/bioinformatics/btx345

34. Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.
35. Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832-844.
36. Itkin, M., Heinig, U., Tzfadia, O., Bhide, A. J., Shinde, B., Cardenas, P. D., ... &Tikunov, Y. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, *341*(6142), 175-179.
37. Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, *3*(3), 318-356.
38. Jacob, F., Perrin, D., Sánchez, C., Monod, J., & Edelstein, S. (2005). The operon: a group of genes with expression coordinated by an operator. CR Acad. Sci. Paris 250 (1960) 1727-1729. *Comptesrendusbiologies*, *328*(6), 514-520.
39. Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, Asdrubal Lopez (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 480, 189-215, doi.org/10.1016/j.neucom.2019.10.118.
40. Jonczyk, R., Schmidt, H., Osterrieder, A., Fiesselmann, A., Schullehner, K., Haslbeck, M., ... & Frey, M. (2008). Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of Bx6 and Bx7. *Plant physiology*, *146*(3), 1053-1063.
41. Kautsar, S. A., Duran, H. G. S., &Medema, M. H. (2018). Genomic identification and analysis of specialized metabolite biosynthetic gene clusters in plants using PlantiSMASH. In *Plant Chemical Genomics* (pp. 173-188). Humana Press, New York, NY
42. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic acids research*, *45*(W1), W55-W63.
43. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic acids research*, *45*(W1), W55-W63.

44. Khaldi, N., Seifuddin, F. T., Turner, G., Haft, D., Nierman, W. C., Wolfe, K. H., &Fedorova, N. D. (2010). SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology*, *47*(9), 736-741.
45. Khanh, T. D., Chung, M. I., Xuan, T. D., &Tawata, S. (2005). The exploitation of crop allelopathy in sustainable agricultural production. *Journal of Agronomy and Crop Science*, *191*(3), 172-184.
46. Khush, G. S. (2001). Green revolution: the way forward. *Nature reviews genetics*, *2*(10), 815-822.
47. Koonin, E. V. (2009). Evolution of genome architecture. *The international journal of biochemistry & cell biology*, *41*(2), 298-306.
48. Kretschmann,E., Fleischmann,W. and Apweiler,R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, 17, 920–926.
49. Krokida, A., Delis, C., Geisler, K., Garagounis, C., Tsikou, D., Peña-Rodríguez, L. M., ... &Papadopoulou, K. K. (2013). A metabolic gene cluster in L otus japonicus discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytologist*, *200*(3), 675-690.
50. Kurkcuoglu, Z., Koukos, P. I., Citro, N., Trellet, M. E., Rodrigues, J. P. G. L. M., Moreira, I. S., &Xue, L. C. (2018). Performance of HADDOCK and a simple contact-based protein–ligand binding affinity predictor in the D3R Grand Challenge 2. *Journal of computer-aided molecular design*, *32*(1), 175-185.
51. Lear, S., Cobb, S.L (2016). Pep-Calc.com: a set of web utilities for the calculation of peptide and peptoid properties and automatic mass spectral peak assignment. *J Comput Aided Mol Des,* 30**,** 271–277. https://doi.org/10.1007/s10822-016-9902-7
52. Lewi, P. J. (1994). 3.3 Receptor Mapping and Phylogenetic Clustering. *Methods and Principles in Medicinal Chemistry*, 131.
53. Li, J., & Wong, L. (2002). Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*, *18*(5), 725-734.
54. Li, J., Cocker, J. M., Wright, J., Webster, M. A., McMullan, M., Dyer, S., ... & Gilmartin, P. M. (2016). Genetic architecture and evolution of the S locus supergene in Primula vulgaris. *Nature plants*, *2*(12), 1-7.

55. Liu, L. W., Lu, C. T., Wang, Y. M., Lin, K. H., Ma, X., & Lin, W. S. (2022). Rice (Oryza sativa L.) Growth Modeling Based on Growth Degree Day (GDD) and Artificial Intelligence Algorithms. *Agriculture*, *12*(1), 59.
56. Mackay, J., Dean, J. F., Plomion, C., Peterson, D. G., Cánovas, F. M., Pavy, N.,&Vinceti, B. (2012). Towards decoding the conifer giga-genome. *Plant molecular biology*, *80*(6), 555-569.
57. Matsuba, Y., Nguyen, T. T., Wiegert, K., Falara, V., Gonzales-Vigil, E., Leong, B., ... &Usadel, B. (2013). Evolution of a complex locus for terpene biosynthesis in Solanum. *The Plant Cell*, *25*(6), 2022-2036.
58. McCallum, C. M., Comai, L., Greene, E. A., &Henikoff, S. (2000). Targeted screening for induced mutations. *Nature biotechnology*, *18*(4), 455-457.
59. Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., & Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research*, *39*(suppl_2), W339-W346.
60. Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., & Cruz-Morales, P. (2015). Minimum information about a biosynthetic gene cluster. *Nature chemical biology*, *11*(9), 625-631.
61. Mugford, S. T., Louveau, T., Melton, R., Qi, X., Bakht, S., Hill, L.,& Osbourn, A. (2013). Modularity of plant metabolic gene clusters: a trio of linked genes that are collectively required for acylation of triterpenes in oat. *The Plant Cell*, *25*(3), 1078-1092.
62. Nim, S., Jeon, J., Corbi-Verge, C., Seo, M. H., Ivarsson, Y., Moffat, J., & Kim, P. M. (2016). Pooled screening for antiproliferative inhibitors of protein-protein interactions. *Nature chemical biology*, *12*(4), 275-281.
63. Nützmann, H. W., & Osbourn, A. (2014). Gene clustering in plant specialized metabolism. *Current opinion in biotechnology*, 26, 91-99.
64. Nützmann, H. W., Scazzocchio, C., & Osbourn, A. (2018). Metabolic gene clusters in eukaryotes. *Annual Review of Genetics*, *52*, 159-183.
65. Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., &Vicedomini, R. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, *497*(7451), 579-584.
66. Osbourn, A. (2010). Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends in Genetics*, *26*(10), 449-457.

67. Osbourn, A., Papadopoulou, K. K., Qi, X., Field, B., &Wegel, E. (2012). Finding and analyzing plant metabolic gene clusters. In *Methods in enzymology* (Vol. 517, pp. 113-138). Academic Press.
68. Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J., & Osbourn, A. E. (1999). Compromised disease resistance in saponin-deficient plants. *Proceedings of the National Academy of Sciences*, *96*(22), 12923-12928.
69. Peng, S., Laza, R. C., Visperas, R. M., Khush, G. S., Virk, P., & Zhu, D. (2004, September). Rice: progress in breaking the yield ceiling. In *Proceedings of the 4th international crop science congress* (Vol. 26).
70. Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.
71. Prasad, B., Garg, A., Takwani, H., & Singh, S. (2011). Metabolite identification by liquid chromatography-mass spectrometry. *TrAC Trends in Analytical Chemistry*, *30*(2), 360-387.
72. Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R., & Osbourn, A. (2004). A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proceedings of the National Academy of Sciences*, *101*(21), 8233-8238.
73. Qi, X., Bakht, S., Qin, B., Leggett, M., Hemmings, A., Mellon, F., ... & Melton, R. (2006). A different function for a member of an ancient and highly conserved cytochrome P450 family: from essential sterols to plant defense. *Proceedings of the National Academy of Sciences*, *103*(49), 18848-18853.
74. Qin, F.-J., Sun, Q.-W., Huang, L.-M., Chen, X.-S., & Zhou, D.-X. (2010). Rice SUVH histone methyltransferase genes display specific functions in chromatin modification and retrotransposon repression. *Molecular Plant,* 3(4), 773–782.
75. Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.
76. Rocha, E. P. (2008). The organization of the bacterial genome. *Annual review of genetics*, *42*, 211-233.
77. Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, *5*(4), 725-738.
78. Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, *5*(4), 725-738.

79. Sakamoto, T., Miura, K., Itoh, H., Tatsumi, T., Ueguchi-Tanaka, M., Ishiyama, K., ... &Miyao, A. (2004). An overview of gibberellin metabolism enzyme genes and their related mutants in rice. *Plant Physiology*, *134*(4), 1642-1653.
80. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., & Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic acids research*, *33*(suppl_2), W363-W367.
81. Schonlau M, Zou RY (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29. doi:10.1177/1536867X20909688
82. Sethy, P. K., Barpanda, N. K., Rath, A. K., & Rajpoot, S. C. (2021). Rice (Oryza Sativa) panicle blast grading using support vector machine based on deep features of small CNN. *Archives of Phytopathology and Plant Protection*, *54*(15-16), 1001-1013.
83. Shimura, K., Okada, A., Okada, K., Jikumaru, Y., Ko, K. W., Toyomasu, T., ... & Koga, J. (2007). Identification of a biosynthetic gene cluster in rice for momilactones. *Journal of Biological Chemistry*, *282*(47), 34013-34018.
84. Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B., & Peters, R. J. (2009). CYP76M7 is an ent-cassadiene C11α-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *The Plant Cell*, *21*(10), 3315-3325.
85. Takos, A. M., Knudsen, C., Lai, D., Kannangara, R., Mikkelsen, L., Motawia, M. S., &Møller, B. L. (2011). Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in Lotus japonicus and suggests the repeated evolution of this chemical defence pathway. *The Plant Journal*, *68*(2), 273-286.
86. Tobler, J. B., Molla, M. N., Nuwaysir, E. F., Green, R. D., & Shavlik, J. W. (2002). Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. *Bioinformatics*, *18*(suppl_1), S164-S171.
87. Töpfer, N., Fuchs, L. M., & Aharoni, A. (2017). The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic acids research*, *45*(12), 7049–7063. https://doi.org/10.1093/nar/gkx404
88. Upadhyay, A. K., & Sowdhamini, R. (2016). Genome-wide prediction and analysis of 3D-domain swapped proteins in the human genome from sequence information. *PloS one*, *11*(7), e0159627.
89. Vaughan, D. A., Lu, B. R., &Tomooka, N. (2008). The evolving story of rice evolution. *Plant science*, *174*(4), 394-408.

90. Von Rad, U., Hüttl, R., Lottspeich, F., Gierl, A., & Frey, M. (2001). Two glucosyltransferases are involved in detoxification of benzoxazinoids in maize. *The Plant Journal*, *28*(6), 633-642.
91. Wang, X. K., & Li, R. H. (1997). Determination and classification of subspecies of Asian rice and their inter-subspecies hybrids. *Chin Sci Bull*, *42*, 2596-2603.
92. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R.,& Lepore, R. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*(W1), W296-W303.
93. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, et al. (2015). antiSMASH 3.0: a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* 43: W237–W243.
94. Wilderman, P. R., Xu, M., Jin, Y., Coates, R. M., & Peters, R. J. (2004). Identification of syn-pimara-7, 15-diene synthase reveals functional clustering of terpene synthases involved in rice phytoalexin/allelochemical biosynthesis. *Plant Physiology*, *135*(4), 2098-2105.
95. Wink, M. (2010). Introduction: biochemistry, physiology and ecological functions of secondary metabolites. *Annual plant reviews volume 40: Biochemistry of plant secondary metabolism*, 1-19.
96. Winzer, T., Gazda, V., He, Z., Kaminski, F., Kern, M., Larson, T. R., ... & Walker, C. (2012). A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*, *336*(6089), 1704-1708.
97. Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *AcmSigmod Record*, *31*(1), 76-77.
98. Xu, M., Galhano, R., Wiemann, P., Bueno, E., Tiernan, M., Wu, W., & Peters, R. J. (2012). Genetic evidence for natural product-mediated plant–plant allelopathy in rice (Oryza sativa). *New Phytologist*, *193*(3), 570-575.
99. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, *117*(3), 1496-1503.
100. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature methods*, *12*(1), 7-8.
101. Yi, G., Sze, S. H., & Thon, M. R. (2007). Identifying clusters of functionally related genes in genomes. *Bioinformatics*, *23*(9), 1053-1060.

102. Zheng, Y., Szustakowski, J. D., Fortnow, L., Roberts, R. J., &Kasif, S. (2002). Computational identification of operons in microbial genomes. *Genome research*, *12*(8), 1221-1230.

ISSN 0974-3618 (Print)
0974-360X (Online)

www.rjptonline.org

RESEARCH ARTICLE

Genome-wide Identification and Annotation of metabolite producing Gene Clusters in Rice Genome

Himanshu Singh[1], Nooman Siddique[1], Atul Kumar Upadhyay[2]
[1]School of Bioengineering and Biosciences, Lovely Professional University, Punjab
[2]Department of Biotechnology, Thapar University, Punjab
*Corresponding Author E-mail: **atul.20483@lpu.co.in**

ABSTRACT:
Plants are able to produce numerous types of metabolites, which can be utilized for drug development, and various other purposes like resistance to a pest, abiotic stresses, and disease. Recently it was discovered that genes, which are responsible for the production of these metabolites, are arranged in operon-like gene clusters. These gene clusters are co-expressed and regulated by the same set of regulatory elements. Identification of such gene clusters has tremendous application in synthetic biology. Advancement in genome information, genome mining, and analysis tools have placed us at a commanding position that will provide us a technique to modify the gene for large production of the specialized chemicals. Rice being stapled food for the majority of the human population is chosen in the present study to find out the gene clusters responsible for the production of secondary metabolites. We have reported 39 gene clusters on 12 chromosomes of Oryza sativa group Japonica. Few of the selected metabolite producing gene clusters is a saccharide, lignin, terpene, alkaloid etc. There were several gene clusters for putative metabolites, which we have annotated in this study.

KEYWORDS: Gene Cluster, Plant Metabolites, Alkaloid, biosynthetic gene cluster, Saccharides.

INTRODUCTION:
Plants metabolites are used from ancient time for human health and drug development purpose. The major bottleneck in utilization of these metabolites is slow production rate of these chemicals. Recently it was found that genes, which are responsible for the production of these specialized chemicals, are arranged in clusters. Gene cluster are a group of two or more genes which is found in the DNA of any organism that collectively work in encoding similar proteins or enzymes (Yi, Sze, and Thon, 2007). Gene clusters work in various ways and they usually have the same promoter region by which all of them get expressed. Secondary metabolic pathways consist of genes for — signature enzymes that make the scaffold of the secondary metabolite, along with genes for— tailoring enzymes that carry out subsequent modifications to this scaffold (Osbourn 2010).

Received on 31.09.2019 Modified on 11.10.2019
Accepted on 12.12.2019
Research J. Pharm. and Tech. 2020; 13(4):1744-1746.
DOI: 10.5958/0974-360X.2020.00314.5

Examples of plant signature enzymes are terpene synthases (for terpenes), chalcone synthases (for flavonoids), and CYP79 family enzymes for cyanogenic glucosides (D'Auria and Gershenzon, 2005; Osbourn, 2010). Examples of tailoring enzymes include oxidoreductases, methyltransferases, acyltransferases, and glycosyltransferases. A good candidate gene cluster will contain genes encoding a signature enzyme and tailoring enzymes. Given that these genes contribute to a common pathway, they are co-expressed, although this is not always the case (Takos et al., 2011).

Rice (*Oryza sativa*) is of family poaceae and genus *oryza* with over 20 cultivated wild species. Rice is the staple food of more than 3 billion people worldwide. After wheat, rice is considered second in the most important crops of the world (Ghosh et al., 2016). We are targeting the rice plant for identifying the gene clusters to enhance the nutrient quality.

METHODOLOGY:
We are inspecting and learning about Gene Cluster found in the genome of Rice plant. We first took the whole genome sequences of *Oryza sativa* of Japonica cultivar (Wang et al., 2013) and then took all of the 12 chromosomes and ran it in a online tool called plantismash.

Plantismash is a tool that carries out automated identification, annotation as well as the expression analysis of a plants biosynthetic gene clusters or BCG in short. Plantismash works in identifying the gene clusters by looking for all the genes predicted to encode biosynthetic enzymes, including those required for tailoring of the scaffold. This it is a very useful tool and helps in the discovery of candidate BGC (Kemmerer, Lei, and Wu, 1991) and the powerful visualization of co-expression data it provides allows their prioritization present a key technological step in the route towards high-throughput genome mining of plants natural products and also helps us understand various pathways in any plant. In our case we chose to find information about rice and got various results in terms of chromosomes. We have performed detailed annotation of the gene clusters for the presence of different motifs by using Prosite database.

RESULTS:

We got several gene clusters (39) of rice on various chromosomes and those gene clusters were differentiated as alkaloid, saccharide, terpenes, putative etc., based on the type of gene and the product that particular sequence in the genome (Qin, Sun, Huang, Chen, and Zhou, 2010). We inspect various clusters based on their function based on genes they contain. We took the gene clusters responsible for synthesizing alkaloids as they have application and value in various industries including Pharmaceuticals, food, etc. Alkaloids are basic (alkali-like) compounds that contain nitrogen which is derived from amino acid metabolism (Samanani, Liscombe, and Facchini, 2004). The biosynthesis of alkaloid usually follows complex pathway and includes stereospecific steps. Different alkaloids from plants have been known to be of very high value as they have various properties important to human health eg, they are antimalarial, antiasthma, anticancer, analgesic, etc properties. Many have also been in use in traditional as well as modern medicine and have been starting points of drug discovery. (Stöckigt, Barleben, Panjikar, and Loris, 2008) Thus identifying and learning ways to study them is going to be very helpful to a lot of fields.

Table1: *Gene clusters of Oryza sativa Japonica* group (chromosome-1 to chromosome-12)

	Sr. No.	Gene cluster	Size (kb)	Core Domains
Chromosome 1	1.	Saccharide	71.44	2OG-FeII_Oxy, DIOX_N, Peptidase_S10, UDPGT_2
	2.	Lignan- Polyketide	70.90	Chal_ sti_synt_C, Chal_sti_synt_N, Dirigent, p450
	3.	Saccharide	82.51	Aminotran_1_2, UDPGT_2
	4.	Saccharide	72.22	UDPGT_2, p450
	5.	Alkaloid	33.28	Bet_v_1, Epimerase, Methyltransf_11
Chromosome 2	1.	Saccharide	139.97	Glycos_transf_1, p450
	2.	Saccharide- Polyketide	211.17	Chal_sti_synt_C, UDPGT_2, p450
	3.	Terpene	369.98	COesterase, Terpene_synth, Terpene_synth_C, p450
Chromosome 3	1.	Lignan- Saccharide	97.55	Cellulose_synt, Dirigent, Methyltransf_11, UDPGT_2
	2.	Saccharide	64.12	Amino_oxidase, UDPGT_2, adh_short
Chromosome 4	1.	Terpene	212.71	Terpene_synth, Terpene_synth_C, adh_short_C2, p450
	2.	Saccharide- Alkaloid	360.51	Cu_amine_oxid, UDPGT_2, adh_short
	3.	Saccharide	169.20	Peptidase_S10, UDPGT_2
	4.	Terpene	334.35	2OG-FeII_Oxy, Terpene_synth, Terpene_synth_C
	5.	Saccharide	42.28	Peptidase_S10, UDPGT_2
	6.	Terpene	61.50	Terpene_synth, Terpene_synth_C, Transferase
	7.	Lignan	82.15	2OG-FeII_Oxy, DIOX_N, Dirigent, Methyltransf_7
Chromosome 5	1.	Saccharide	207.12	2OG-FeII_Oxy, DIOX_N, Transferase, UDPGT_2
Chromosome 6	1.	Putative	71.58	2OG-FeII_Oxy, DIOX_N
	2.	Putative	105.71	Peptidase_S10, Transferase, adh_short_C2
	3.	Saccharide	165.15	Transferase, UDPGT_2
	4.	Polyketide	133.31	Chal_sti_synt_C, p450
Chromosome 7	1.	Lignan	86.46	Aminotran_1_2, Dirigent
	2.	Lignan- Saccharide	88.18	Aminotran_1_2, Dirigent, Glycos_transf_1
	3.	Lignan	86.37	CO-esterase, Dirigent, p450
Chromosome 8	1.	Saccharide- Terpene	127.02	Methyltransf_2, Terpene_synth, Terpene_synth_C, UDPGT_2
	2.	Lignan- Alkaloid	132.28	Bet_v_1, Dirigent, Epimerase
	3.	Putative	83.82	COesterase, adh_short
Chromosome 9	1.	Saccharide	99.62	AMP-binding, UDPGT_2, p450
	2.	Putative	150.15	COesterase, Peptidase_S10, adh_short
Chromosome 10	1.	Saccharide	141.74	Transferase, UDPGT_2, p450
	2.	Lignan- Saccharide	432.20	Dirigent, UDPGT_2, p450
	3.	Polyketide	141.94	Acetyltransf_1, COesterase, Chal_sti_synt_C, Epimerase
	4.	Polyketide	139.12	Amino_oxidase, Chal_sti_synt_C, GMC_oxred_C, GMC_oxred_N
Chromosome 11	1.	Alkaloid	41.98	HMGL-like, Str_synth, p450
	2.	Lignan	130.12	Dirigent, Peptidase_S10
	3.	Saccharide	468.44	2OG-FeII_Oxy, UDPGT_2, adh_short, adh_short_C2
Chromosome 12	1.	Lignan	323.86	Dirigent, Methyltransf_2, p450
	2.	Saccharide	67.79	Glycos_transf_1, p450

The annotation of gene clusters were performed and it was found that on the Chromosome 1, Cluster 1, the most similar sequence that was obtained was Proteosome subunit alpha- type. (Query coverage: 7%. Similarity: 89%.), Chromosome 1, cluster 4, the most similar sequence that was obtained was **Peroxidase precursor**. (Query coverage: 8%. Similarity: 96%.), Chromosome 3, cluster 1, the most similar sequence that was obtained was **Isoflavanone**. (Query coverage: 9%. Similarity: 87%.), Chromosome 4, Cluster 2: Most similar sequence that was obtained was **Carboxylesterase**. (Query coverage: 17%. Similarity: 100%.), Chromosome 5, Cluster 2: Most similar sequence that was obtained was **Proteosome subunit alpha-type**. (Query coverage: 8%. Similarity: 88%.), Chromosome 6, Cluster 2: Most similar sequence that was obtained was **Amyloplastic**. (Query coverage: 13%. Similarity: 94%.), Chromosome 9, Cluster 2: Most similar sequence that was obtained was **prx 89 gene**. (Query coverage: 21%. Similarity: 83%.), Chromosome 10, Cluster 6: Most similar sequence that was obtained was **Hydroxyl Isoflavanone**. (Query co

One of the predicted clusters is of 71 Kb and contains four genes producing different protein domains viz., 2OG- FeII Oxy, DIOX_N, Peptidase_S10, UDPGT_2. Detailed analysis of theses genes was performed. For DIOX_N, motif search in the prosite database resulted four motifs (Figure 1).

VTTYARFTYYPPcprPELVYGlkPHTDN--SVLTVLLLDKHVGGLQLLKDGrwldI--PV
LTNELLVVAGD----EielfallgvadHeqvfmapVHRVVT-SERERMSVVMFYQP

Predicted features:

DOMAIN	225	331	Fe2OG dioxygenase
METAL	249		Iron
METAL	251		Iron
METAL	313		Iron
BINDING	322		2-oxoglutarate

Figure 1: Description of different motifs present in the DIOX_N protein.

CONCLUSIONS:

Rice is one of the most important model crop as well as one of the most consumed grains in the world. Rice plant is able to produce the various kinds of specialized chemical, which may be utilized for drug development. The major problem associated with use of secondary metabolites is slow production rate of these specialized chemicals.

Present study provides the information that rice genomes contain numerous gene clusters. Each gene cluster expression is governed mostly by one promoter region. Further investigation will provide us ways for regulations of the gene expression, thus understanding various features in the rice genome can help us study different pathways. This work has also helped get more information regarding the annotation of the rice genome. We have found around 39 probable gene clusters for production of metabolites such as saccharides, alkaloid, polyketide, and terpene. Further analysis will provide us opportunity to manipulate the rice genome for large production of the desired chemicals.

REFERENCES:

1. Yi, G., Sze, S.-H., and Thon, M. R. (2007). Identifying clusters of functionally related genes in genomes. Bioinformatics (Oxford, England), 23(9), 1053–1060. https://doi.org/10.1093/bioinformatics/btl673
2. Osbourn, A. (2010). Secondary metabolic gene clusters: Evolutionary toolkits for chemical innovation. Trends in Genetics, 26, 449–457
3. D'Auria, J. C., and Gershenzon, J. (2005). The secondary metabolism of Arabidopsis thaliana: Growing like a weed. Current Opinion in Plant Biology, 8, 308–316.
4. Ghosh, B., Ali, M.N. and Gantait, S. (2016). Response of Rice under Salinity Stress: A Review Update. J. Res. Rice, 4(2): 2–9.
5. Osbourn, A., and Field, B. (2009). Operons. Cellular and Molecular Life Sciences, 66, 37555–37575.
6. Wang, S., Takahashi, H., Kajiura, H., Kawakatsu, T., Fujiyama, K., and Takaiwa, F. (2013). Transgenic rice seeds accumulating recombinant hypoallergenic birch pollen allergen Bet v 1 generate giant protein bodies. Plant and Cell Physiology, 54(6), 917–933. https://doi.org/10.1093/pcp/pct043
7. Kemmerer, E. C., Lei, M., and Wu, R. (1991). Isolation and molecular evolutionary analysis of a cytochrome c gene from Oryza sativa (rice). Molecular Biology and Evolution, 8(2), 212–226.
8. https://doi.org/10.1093/oxfordjournals.molbev.a040639 Qin, F.-J., Sun, Q.-W., Huang, L.-M., Chen, X.-S., and Zhou, D.-X. (2010). Rice SUVH histone methyltransferase genes display specific functions in chromatin modification and retrotransposon repression. Molecular Plant, 3(4), 773– 782. https://doi.org/10.1093/mp/ssq030
9. Samanani, N., Liscombe, D. K., and Facchini, P. J. (2004). Molecular cloning and characterization of norcoclaurine synthase, an enzyme catalyzing the first committed step in benzylisoquinoline alkaloid biosynthesis. The Plant Journal : For Cell and Molecular Biology, 40(2), 302—313. https://doi.org/10.1111/j.1365-313x.2004.02210.x
10. Stöckigt, J., Barleben, L., Panjikar, S., and Loris, E. A. (2008). 3D-Structure and function of strictosidine synthase-- the key enzyme of monoterpenoid indole alkaloid biosynthesis. Plant Physiology and Biochemistry : PPB, 46(3), 340—355. https://doi.org/ 10.1016/j.plaphy.2007.12.011
11. Wen, J., Vanek-Krebitz, M., Hoffmann-Sommergruber, K., Scheiner, O., and Breiteneder, H. (1997). The potential of Betv1 homologues, a nuclear multigene family, as phylogenetic markers in flowering plants. Molecular Phylogenetics and Evolution, 8(3), 317—333. https://doi.org/10.1006/mpev.1997.0447
12. Ritika Chauhan, Nidhi Singh, Jayanthi Abraham. Bioactivity and Molecular Docking of Secondary Metabolites produced by Streptomyces xanthochromogenes JAR5. Research J. Pharm. and Tech. 8(3): Mar., 2015; Page 300-309.
13. Prabhjot Kour, Anupam Tiwari. Medicinal Values of Secondary Metabolites of *Withania Somnifera*. Research J. Pharm. and Tech 2018; 11(7): 3167-3170.
14. Savita More, Vijay Raje, Namita Phalke, Sarika Lokhande. Bioinformatics – An Emerging Field. Asian J. Res. Pharm. Sci. 2018; 8(4): 185-191.
15. Neeraj Upmanyu, Surya Gupta, B N Prakash, Gopal Garg, P Mishra. Cheminformatics: A New Approach. Research J. Pharm. and Tech. 1(1): Jan.-Mar. 2008; Page 2-5.

ISSN 0975-2366
DOI: https://doi.org/10.31838/ijpr/2021.13.02.324

Research Article

In Silico Study to Establish Molecular Interaction between Plant Gene Clusters to Improve Metabolite Production

HIMANSHU SINGH[1], VINEETH CHANGARANGATH[1], VIKAS KAUSHIK*[1]
[1]Department of Bioinformatics, School of Bioengineering and Biosciences, Lovely Professional University, Phagwara, Punjab
*Corresponding Author
Email ID: himanshu.11691@lpu.co.in, vineeth.ccc@gmail.com, vikas31bt@gmail.com
Received: 12.01.21, Revised: 11.02.21, Accepted: 10.03.21

ABSTRACT
Organized microbial gene expression under the concept of operons is a well-established concept. An operon-esque gene arrangement in the plant kingdom has been brought to the limelight with the help of recent developments in genetics, biochemistry and bioinformatics. We aim to explore the interactions between various gene products (enzymes) from the gene clusters of the 12 chromosomes in Oryza sativa Japonica and Oryza sativa Indica. Sequence information of the all the reviewed signature and tailoring genes were retrieved. The 3-dimentional structures of these proteins were predicted using Bioinformatics tools. The interactions among various predicted structures were ascertained by molecular docking techniques. The hypothesis that the various signature gene products and the corresponding tailoring gene products of the gene clusters of Oryza undergo protein-protein interaction was substantiated by the reliable molecular docking results reveals perfect interaction which leads to production of metabolites. The dendrogram generated for each signature gene products displays the evolutionary relationship with similar proteins from other species or the predicted/hypothetical proteins suggested by the algorithm.

Keywords: Oryza sativa, Enzyme, Molecular Docking, and Evolutionary relationship

INTRODUCTION

The endless possibilities that have presented themselves due to the exponential developments in the field of plant biology and natural product discovery by way of plant metabolite gene cluster discovery include improved farming (allelopathic interactions), drug discovery, better nutrition and lastly but not limited to synthetic biology [1].

Belonging to the family Poaceae and genus Oryza, rice (Oryza sativa) has over 20 wild varieties cultivated around the world of which the two most cultivated varieties are Oryza sativa and Oryza glaberrima. Oryza sativa is grown globally while Oryza glaberrima has its origins dating back to 3500 years ago in West Africa. With n=12 chromosome number, rice species' are observed to be either diploid or triploid. Oryza sativa or Oryza glaberrima L. are diploid species (2n = 24). The genome of Asian cultivated rice was the first among all crop genomes to be completely sequenced. Over 3 billion people consume rice as a staple food which provides recommended quantities of niacin and zinc. With a digestibility of around 88 percent, the rice protein is regarded as the richest biological protein. Wheat takes the mantle as the most important food crop while rice comes in at a close second [2]. Two or more genes which are working together for the production of a protein or enzyme, in the DNA of a particular organism are collectively termed as a gene cluster [3]. The genes in a cluster usually get expressed with the same promoter region. In some gene clusters where there is formation of enzymes. There are steps in which the formation of enzyme takes place. First there is the synthesis of enzymes known as tailoring enzyme which helps in catalyzing as well as accelerating the processes in order to form the main enzyme which is bigger and more complex molecule known as the Signature enzyme. Gene clusters existing in all the 12 chromosomes of Oryza sativa Japonica (Japan rice) have been characterized and are known to consist of a signature gene and one or multiple tailoring genes [4]. It is reasonable to assume that the interactions between the enzymes result in the enhancement of the overall cluster expression. This hypothesis, if proven right, can open an avenue for improving the expression of specific clusters within a given plant species [5]. The pertaining data which includes the evolutionary relationships can be used to device an algorithm that predicts the gene clustering of any given species of plants. Research in the field of plant secondary metabolic gene clusters has seen exponential growth over the past few years. It is

only poised to extend farther as more discoveries of and about gene clusters emerge at an accelerated rate, thanks to high throughput screening methods. Combining systematic genome mining and functional analysis of candidate clusters along with artificial intelligence and machine learning makes the discovery of new pathways, enzymes and chemistries well within the realm of possibility. In this project we are aiming to develop a universal cluster prediction algorithm using data sets obtained through interaction studies of gene products and their evolutionary relationships.

METHODOLOGY

Signature and tailoring genes were selected from the many gene clusters within the 12 chromosomes of Japonica and Indica varieties. Sequences of chromosomes of Japonica and Indica were obtained from the NCBI and further fed to the Plantismash software. Details of signature and tailoring genes were obtained from the analysis of Plantismash result data. A simple search with the name of the gene product and the species name was performed on UniProtKB. Out of the search results obtained, the ones which were annotated and manually reviewed were selected for the purposes of reliability. The FASTA sequences of the selected gene products (enzymes/proteins) were extracted from UniProtKB which was in turn used as the input for Swiss-Model, TrRosetta and I-TASSER servers to predict the 3D structures. The obtained 3D structures were then loaded to the patch dock server to conduct molecular docking and interaction analysis. Finally, phylogenetic trees were developed by utilizing the BLAST TREE widget on pBLAST. Flow chart is being given below in figure-1.

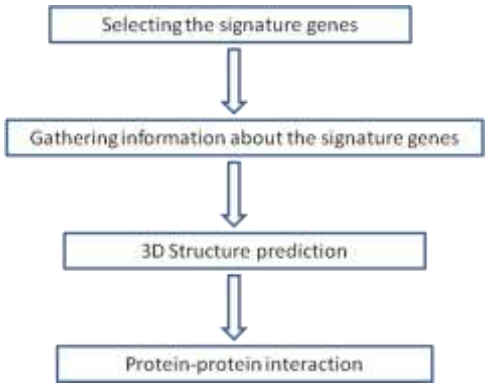


Fig-1 Flow chart for methodology

RESULT AND DISCUSSION

A. Selecting the signature genes

A gene from each cluster was randomly selected as a potential signature candidate. Different genes were selected as signatures from clusters that produce same metabolites. In total we have selected 12 genes.

B. Gathering information about the signature genes

The signature gene products were looked up in UniProtKB. Information such as the function, gene name, FASTA sequence, family, superfamily and pdb (protein data bank) structure were gathered. 11 signature genes out of the 39 had annotated UniProtKB profiles. Namely: Acetyl transferase, Alcohol dehydrogenase , Cellulose synthase, Chalcone synthase, Carboxyl esterase, DIOX N, Epimerase, Methyl transferase, Cytochrome P450, Peptidase S10, UDPGT(Glucuronosyltransferase), for Oryza sativa japonica. Alcohol dehydrogenase, chalcone synthase, methyl transferase, epimerase, glycosyl transferase, terpene synthase, acetyl transferase, amino oxidasem amino transferase, Sqs psy, Cellulase synthase were the annotated gene products belonging to Oryza sativa indica.

C. 3-D structure prediction

Oryza stativa japonica

None of the signature gene products had 3D structures available in PDB. Swiss Model server was used for the prediction of all possible 3D conformations of these proteins by virtue of homology. Due to some un-foreseen errors, the models predicted using Swiss-Model was fragmented. Ab-Initio protein modeling using trRosetta [6][7] and I-TASSER web server were used to circumvent this issue [8][9][10]. TM Scores for selected enzymes was successfully interpreted: Acetyltransferase : Estimated TM-score: 0.753, Alcohol dehydrogenase : Estimated TM-score: 0.776, Cellulose synthase : Estimated TM-score = 0.70±0.12; C-score=-0.14,Chalcone synthase : Estimated TM-score: 0.815, COesterase : Estimated TM-score: 0.803, DIOXN : Estimated TM-score: 0.729, Epimerase : Estimated TM-score: 0.788, Methyl transferase : Estimated TM-score: 0.759, P450 : Estimated TM-score: 0.694, Peptidase S10 : Estimated TM-score: 0.706, UDPGT : Estimated TM-score: 0.476, trRosetta server predicts a total of 5 possible 3D structures for a given sequence (**Figure 2**). The quality of the predicted structures can be deduced from the Template Modeling scores(TM score). TM score signifies similarity which is akin to Root Mean Square Deviation (RMSD). A 100 percent identical match between two given structures is indicated by a TM score of 1. The structure of Cellulose synthase predicted through I-Tasser web server is validated by another parameter called the C-score. The quality of models predicted by I-Tasser server can be judged based on the C-score which is a confidence score for quality. Significance of threading template alignments and

the convergence parameters of the structure assembly simulations are considered for C-score estimation. C-score values lie between (-5, 2).A higher C-score signifies that the model is of high confidence and a lower score indicates otherwise.

Oryza sativa indica

10 out of the 11 proteins had their 3-D structures available in the Swiss Model Repository (SMR) database. The structure of Cellulose synthase was determined using I-Tasser web server which had the following parameters: C score=0.82, TM score=0.82+0.08, RMSD=7.4+4.2 Angstroms. (Figure 3)

D. Interaction study

The proteins produced by the signature genes were assumed to interact with the protein products of their respective tailoring genes of the cluster. To study the said interactions, the signature genes products were docked with the tailoring gene products using PatchDock server. Signature gene products in the docked complex are represented in red colour and the tailoring gene products in green. Negative Atomic Contact Energies (ACE) between the docked proteins indicate the formation of a stable complex [11][12].Docked complexes with the minimum ACE and considerably high docking score are selected since there is no one to one correlation between merely the docking score and the corresponding binding affinity.

E. Docked Complex

Oryza sativa japonica

Various docked complexes was considered for molecular docking studies : Methyltransferase-Cytochrome P450, Alcohol dehydrogenase (ADH) – Glucuronosyltransferase (UDPGT), UDPGT –Cyto P450, CytP450-Chalcone synthase, UDPGT-peptidase S10, Peptidase S10 –ADH, Peptidase S10 –ADH, UDPGT-DIOXN, Methyltransferase-UDPGT, Acetyl transferase- COesterase, Chalcone synthase-UDPGT, Acetyl transferase-Epimerase, DIOX-Methyl transferase, Cyto P450-Chalcone synthase, COesterase-ADH, Coesterase-CytP450, Peptidase S10-Coesterase, and cellulose synthase-UDPGT. Best outcomes were Methyltransferase-Cytochrome P450 (ACE:-776) and Chalcone Synthase cytP450 (ACE:-764) and perfect interaction as represented in **Figure 4** and **Table 1**. Sequences of the signature gene products were fed to pBLAST to obtain their homologs. This assisted in phylogenetic comparisons for selected enzymes and both were found highly conserved and most functional.

Oryza sativa indica

The following 24 possible interactions were ascertained from the gene products.

Chalconesynthase-Glycosetransferase,Chalconesynthase-acetyltransferase,Methyltransferase-epimerase,Methyltransferase-glycosyltransferase,Methyltransferase-cellulasesynthase,Methyltransferase-acetyltransferase,Methyltransferase-terpenesynthase,Epimerase-methyltransferase,Glycosyltransferase-methyltransferase,Glycosyltransferase-chalconesynthase,Terpenesynthase-methyltransferase,Terpenesynthase-adh,Acetyltransferase-aminooxidase,Acetyltransferase-aminotransferase,Acetyltransferase-methyltransferase,Acetyltransferase-chalconesynthase,Aminooxidase-acetyltransferase, Amino oxidase- amino transferase, Amino oxidase-adh, Aminotransferase- aminooxidase, Aminotransferase-acetyltransferase, Aminotransferase-ADH, Aminotran- glycostransf, Cellulosesynthase-methyl transferase. Best outcomes were Methyl transf-acetyl transf (ACE:-705.53) andAcetyl transf- amino oxidase (ACE:-658.49) (**Figure 5**,**Table 2**)

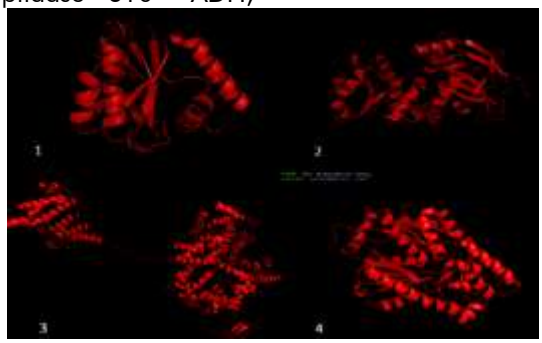


Fig.2: Predicted structures of 1) Acetyltransferase 2) Alcohol dehydrogenase 3) Cellulose synthase 4) Chalcone synthase

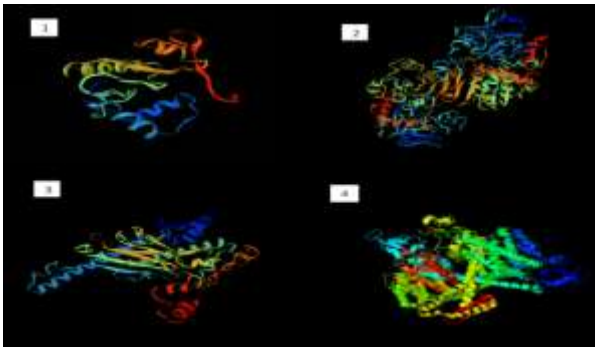


Fig.3: Predicted structures of1) Acetyltransferase 2) ADH 3) Aminooxidase 4) Cellulose synthase

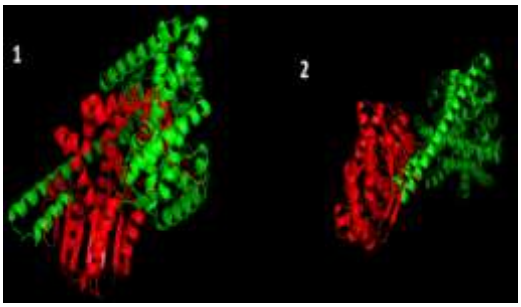


Fig.4: Molecular Docking investigation (least ACE):1) Methyltransferase-CytP4502) Chalcone synthase-CytP450

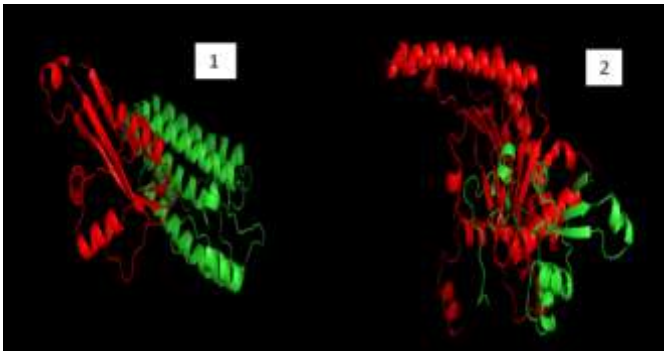


Fig.5: 1) Molecular docking investigation (least ACE) Methyltransferase-Acetyltransferase 2) acetyltransferase-aminooxidase

Table 1. Molecular Docking results for selected complexes from Oryza sativa japonica

Complex	Score	ACE
Methyltransferase-CytP450	13102	-776.10
Chalcone synthase-CytP450	12422	-764.76

Table 2: Molecular Docking results for selected complexes from Oryza sativa indica

Complex	Score	ACE
Methyltransferase-Acetyltransferase	2356.10	-705.53
acetyltransferase-aminooxidase	2754.10	-658.49

CONCLUSIONS

The hypothesis that the various signature gene products and the corresponding tailoring gene products of the gene clusters of Oryza sativa undergo protein-protein interaction was substantiated by the reliable molecular docking results reveals perfect interaction between. The dendrogram generated for each signature gene products displays the evolutionary relationship with similar proteins from other species or the predicted/hypothetical proteins suggested by the algorithm.

ACKNOWLEDGEMENT

All the authors conducted research and write the MS and thoroughly studied, and are thankful towards school of bioengineering and biosciences, Lovely professional university, Phagwara, Punjab for providing finest computational facility for conduction of research. The authors read and approved the final manuscript. The authors hereby declare that they have no conflict of interest. No potential conflict of interest was reported by the author(s).

DECLARATIONS

Ethical clearance: The authors did not perform any experiments on human or animals

Funding: Not applicable

Conflict of interest: The authors hereby declare that they have no conflict of interest.

REFERENCES

1. Nützmann, H. W., & Osbourn, A. (2014). Gene clustering in plant specialized metabolism. Current opinion in biotechnology, 26, 91-99.
2. Yi, G., Sze, S. H., & Thon, M. R. (2007). Identifying clusters of functionally related genes in genomes. Bioinformatics, 23(9), 1053-1060.
3. Qin, F. J., Sun, Q. W., Huang, L. M., Chen, X. S., & Zhou, D. X. (2010). Rice SUVH histone methyltransferase genes display specific functions in chromatin modification and retrotransposon repression. Molecular plant, 3(4), 773-782.
4. Ghosh, B., Md, N. A., & Gantait, S. (2016). Response of rice under salinity stress: a review update. Rice research: open access, 1-8.
5. Chu, H. Y., Wegel, E., & Osbourn, A. (2011). From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. The Plant Journal, 66(1), 66-79.
6. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences, 117(3), 1496-1503.
7. Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics, 57(4), 702-710.
8. Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. Proteins: Structure, Function, and Bioinformatics, 77(S9), 100-113.
9. Roy, A., Yang, J., & Zhang, Y. (2012). COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic acids research, 40(W1), W471-W477.
10. Yang, J., & Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. Nucleic acids research, 43(W1), W174-W181.
11. Duhovny, D., Nussinov, R., & Wolfson, H. J. (2002, September). Efficient unbound docking of rigid molecules. In International workshop on algorithms in bioinformatics (pp. 185-200). Springer, Berlin, Heidelberg.
12. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., & Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic acids research, 33(suppl_2), W363-W367.

CHAPTER 27

In Silico Identification, Analysis, and Prediction Algorithm for Plant Gene Cluster

Himanshu Singh[1], C. Vineeth[1], Bhupender Thakur[1], Atul Kumar Upadhyay[2] and **Vikas Kaushik[1,*]**

[1] *School of Bioengineering and Biosciences, Lovely Professional University, Punjab, India*

[2] *Department of Biotechnology, Thapar University, Punjab, India*

Abstract: The concept/phenomenon of operons, which are organized genes that work in a coordinated way in microbes, is well established. Recent developments in genetics, biochemistry, and bioinformatics have unraveled similar gene arrangements in plants. Here we aim to develop an algorithm/tool which would help us detect and identify biosynthetic gene clusters (BGCs) from any input plant genome. Through this tool, we intend to match or supersede the performance of pre-existing sting tools for BGC prediction, like the popular plantiSMASH. The predictions models were developed using the machine learning tool WEKA using the physicochemical properties as data set to classify between terpene synthases and non-terpene synthases. A set of ten physicochemical properties were selected and their values were predicted for each of the 159 proteins (terpene synthases and non-terpene synthases) Employing the random forest and SMO classifiers, we were able to obtain significantly promising accuracy of over 90 percent with 66 percent percentage split testing. Accurate prediction of BGCs in the plants, especially the major food crops like rice, wheat, and corn revolutionize farming and nutrition for the better.

Keywords: Algorithm, BGC, Mining, PlantiSMASH, Random forest, SMO WEKA.

INTRODUCTION

Metabolite gene cluster discovery techniques are improving at exponential rates which have opened up avenues of endless possibilities in the field of plant biology and natural product discovery. Improved farming (allelopathic interactions), drug discovery, better nutrition, and synthetic biology are only a few of the promising areas [1]. With over 20 wild varieties cultivated around the world, rice(Oryza *Sativa),* a member of the family Poaceae and genus Oryza, is one of the most

* **Corresponding author Vikas Kaushik:** School of Bioengineering and Biosciences, Lovely Professional University, Punjab, India; E-mail: vikas31bt@gmail.com

Dharam Buddhi, Rajesh Singh and Anita Gehlot (Eds.)

popular and staple food crops in the world. *Oryza sativa* and *Oryzaglaberrima*are the most cultivated out of all the rice varieties. Sativa variety is a global favorite while glabberima has been around for over 3500 years, originating in West Africa. Rice species could be diploid or triploid with n=12 and *Oryza sativa* or *Oryzaglaberrima* L. are diploid species (2n = 24). Complete sequencing of the Asian cultivated rice genome has been performed and it was the first food crop to be the whole genome sequenced [2, 3]. A group of genes in the DNA of a particular organism is called a gene cluster when they collectively work in the production of protein or enzyme [4]. These genes are usually regulated by the same promoter region. BGCs produce two types of enzymes in general, signature enzymes produced by the signature genes of the BGC and tailoring enzymes produced by the tailoring genes of the BGC. Tailoring enzymes get produced first which in turn enter a cascade of reactions accelerating and catalyzing the production of the signature enzymes, which is the bigger, complex. In some gene clusters where there is the formation of enzymes. There are steps in which the formation of enzyme takes place. First, there is the synthesis of enzymes known as tailoring enzymes which helps in catalyzing as well as accelerating the processes to form the main enzyme which is a bigger and more complex molecule known as the Signature enzyme [5]. In this project, we are aiming to develop a universal cluster prediction algorithm using terpene synthase gene clusters as the reference and classifying genes into Terpene synthases and non-terpene synthases based on the distribution of selected unique physicochemical properties. Random forest and Sequential minimal optimization (SMO) [6, 7] classifiers were employed for the development of the machine learning model in the WEKA tool.

METHODOLOGY

Collection of Data

We have collected FASTA [8] sequences of Terpene synthases and non-terpene synthases (negative control), across the plant kingdom. UniProtKB database was used for the retrieval of the relevant data. Around 80 terpene synthases and non-terpene synthases each were gathered for this study. Fig. (**1**) shows the flow chart.

Selection and Prediction of Features

A total of 10 physicochemical properties were selected. They were Length, Molecular weight, Isoelectric point, Instability index, Aliphatic index, Hydropathicity Charge at pH 7, TMindex, Solubility, Extinction coefficient. Following web servers were utilized for the prediction of these features where the FASTA sequences of the proteins were fed as input.

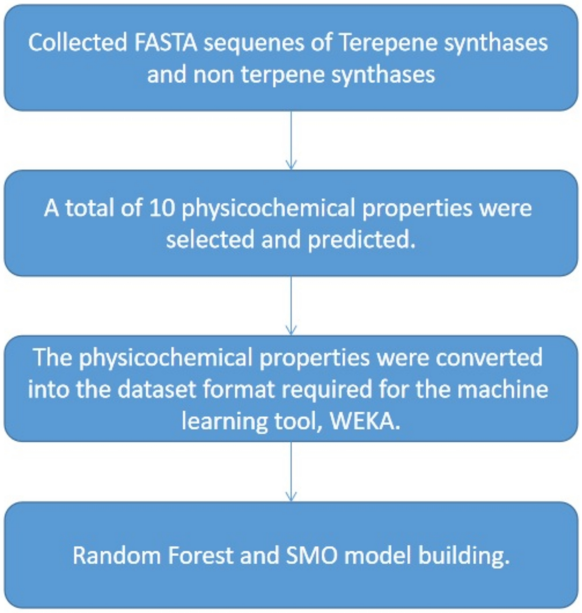


Fig. (1). Flow chart.

- ProtParam by Expasy: Chain length, molecular weight, PI, instability index, aliphatic index, GRAVY (Hydropathicity) https://web.expasy.org/protparam/ [9].
- PROTEIN CALCULATOR v3.4: Charge at pH7 http://protcalc.sourceforge.net/
- TM predictor: TMindex http://tm. life.nthu.edu.tw/
- Protein-sol: Solubility https://protein-sol.manchester.ac.uk/ [10].
- PepCalc.com-Peptide property calculator by INNOVAGEN: Extinction coefficient https://pepcalc.com/

Preparation of Data Set

The physicochemical properties were converted into the dataset format required for the machine learning tool, WEKA [11], which was used in this project, where classes were terpene or non terpene. The data set was saved in .arff format and explored through the WEKA tool for model building.

Model Building

Loaded with several algorithms such as Bayesian Network, SVMLib, Artificial Neural Network (ANN), Nearest Neighbor (IBk), Random Forest, *etc*, Weka is a convenient machine learning tool. For this study, we developed Random forest

and SMO models. Fig. (**2**) shows a Display of the attribute distribution across terpene synthases and non-terpene synthases and Fig. (**3**) shows the J48 decision tree representation of attributes.

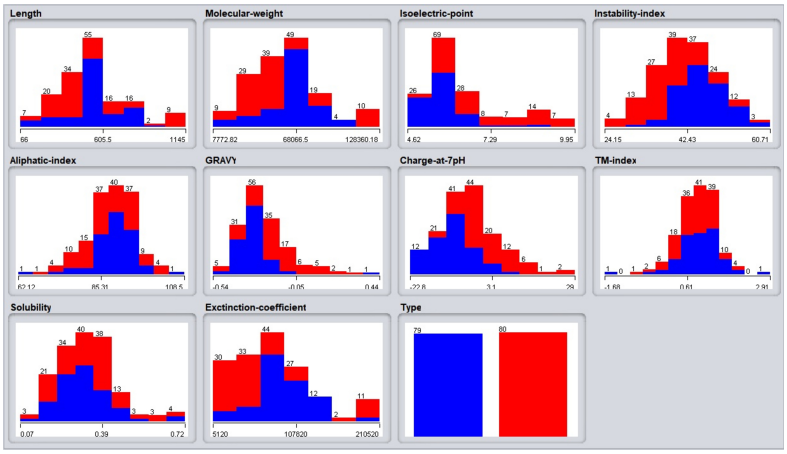


Fig. (2). Display of the attribute distribution across terpene synthases and non-terpene synthases [12].

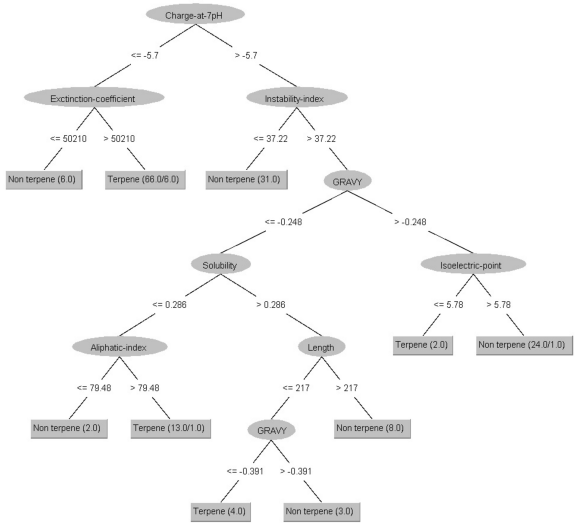


Fig. (3). J48 decision tree representation of attributes [11].

RESULT AND DISCUSSION

Physiochemical values were formatted as per the WEKA tool input requirements. The same ten physiochemical properties were considered for terpene synthases and non-terpene synthases. In this study, we have used 66% percentage split

testing on the data set. We calculated the overall accuracy, true positive rate (TP), false positive rate (FP), precision, recall, Mathew's correlation coefficient(MCC), and receiver operating characteristic (ROC) along with the confusion matrices of the attributes, threshold curves and attribute vise visualization of the classification. Table **1** shows a Summary of the SMO model developed

Table 1. Summary of SMO model developed [12, 13].

Correctly Classified Instances	**52**	**96.29%**
Incorrectly Classified Instances	**2**	**3.70%**
Kappa statistic	**0.92**	
Mean absolute error	**0.037**	
Root mean squared error	**0.19**	
Relative absolute error	**7.40%**	
Root relative squared error	**38.47%**	
Total Number of Instances	**54**	

SMO model developed shows an appreciable 96.2963% accuracy. This means that there is a 96.2963% chance that the model accurately predicts the input to be a terpene synthase or a non-terpene synthase. Table **2** shows the summary of the Random forest model.

Table 2. Summary of Random forest model developed [12, 13].

Correctly Classified Instances	**51**	**94.44%**
Incorrectly Classified Instances	**3**	**5.55%**
Kappa statistic	**0.89**	
Mean absolute error	**0.16**	
Root mean squared error	**0.23**	
Relative absolute error	**32.91%**	
Root relative squared error	**45.79%**	
Total Number of Instances	**54**	

The random forest model gives an accuracy of 94.4444% percent for appropriate classification with 51 correctly classified instances and 3 incorrectly classified ones out of a total of 54. Table **3** shows the Performance of the SMO model developed by the class and Table **4** shows the performance of the Random forest developed by the class.

Table 3. Performance of SMO model developed by class [12, 13].

	TP Rate	FP Rate	Precision	Recall	MCC	ROC Area	Class
	1	0.071	0.929	1	0.929	0.964	Terpene
	0.929	0	1	0.929	0.929	0.964	Non terpene
Weighted Avg.	0.963	0.034	0.966	0.963	0.929	0.964	

Table 4. Performance of the Random forest developed by class [12, 13].

	TP Rate	FP Rate	Precision	Recall	MCC	ROC Area	Class
	0.962	0.071	0.926	0.962	0.889	0.992	Terpene
	0.929	0.038	0.963	0.929	0.889	0.992	Non terpene
Weighted Avg.	0.944	0.054	0.945	0.944	0.889	0.992	

The confusion matrix summarizes the accuracy by illustrating how many entries were classified under the appropriate class and how many were not. Table **5** shows the Confusion matrix of the SMO model and Table **6** shows the confusion matrix of the random forest model.

Table 5. Confusion matrix of the SMO model [12, 13].

Classified as→	Terpene	Non-terpene
Terpene	**26**	**0**
Non-terpene	**2**	**26**

Table 6. Confusion matrix of the random forest model [12, 13].

Classified as →	Terpene	Non-terpene
Terpene	**25**	**1**
Non-terpene	**2**	**26**

CONCLUSION

The models developed with the help of both classifiers (RF and SMO) as a part of this study to predict BGCs in plants, mainly based on the distribution of physicochemical features of the respective protein products of the said BGCs (terpene synthases and non-terpene synthases) showed significantly positive and accurate results of classification. Machine learning approaches, classifiers (SMO and RF models) have been developed to predict domain swapping at the genome level from mere protein sequence information. An accuracy of 96% and 94% were achieved for the two methods, respectively. Research in the field of plant

secondary metabolic gene clusters has seen exponential growth over the past few years. It is only poised to extend farther as more discoveries of and about gene clusters emerge at an accelerated rate, thanks to high throughput screening methods. Combining systematic genome mining and functional analysis of candidate clusters along with artificial intelligence and machine learning discovers new pathways, enzymes, and chemistries well within the realm of possibility. The impact that this accelerated development can have on the way we approach everything from farming to drug discovery to nutrition will be unprecedented. Application of such technologies into staple, major food crop industry could mean a significant quality of life change for the world as a whole through better farming and nutrition.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

All the authors conducted research and write the MS and thoroughly studied, and are thankful towards the school of bioengineering and biosciences, LPU, Phagwara, Punjab for providing the finest computational facility for conduction of research. The authors read and approved the final manuscript. The authors hereby declare that they have no conflict of interest.

REFERENCES

[1] H-W. Nützmann, and A. Osbourn, "Gene clustering in plant specialized metabolism", *Curr. Opin. Biotechnol.,* vol. 26, pp. 91-99, 2014.
[http://dx.doi.org/10.1016/j.copbio.2013.10.009] [PMID: 24679264]

[2] G. Yi, S-H. Sze, and M.R. Thon, "Identifying clusters of functionally related genes in genomes", *Bioinformatics,* vol. 23, no. 9, pp. 1053-1060, 2007.
[http://dx.doi.org/10.1093/bioinformatics/btl673] [PMID: 17237058]

[3] H. Singh, N. Siddique, and A.K. Upadhyay, "Genome-wide Identification and Annotation of metabolite producing Gene Clusters in Rice Genome", *Research Journal of Pharmacy and Technology,* vol. 13, no. 4, p. 1744, 2020.
[http://dx.doi.org/10.5958/0974-360X.2020.00314.5]

[4] F-J. Qin, Q-W. Sun, L-M. Huang, X-S. Chen, and D-X. Zhou, "Rice SUVH histone methyltransferase genes display specific functions in chromatin modification and retrotransposon repression", *Mol. Plant,* vol. 3, no. 4, pp. 773-782, 2010.
[http://dx.doi.org/10.1093/mp/ssq030] [PMID: 20566579]

[5] B. Ghosh, and N. Ali Md, "Response of Rice under Salinity Stress: A Review Update", *Rice Research: Open Access,* vol. 4, 2016.
[http://dx.doi.org/10.4172/2375-4338.1000167]

[6] T.K. Ho, "The random subspace method for constructing decision forests", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 20, no. 8, pp. 832-844, 1998.
[http://dx.doi.org/10.1109/34.709601]

[7] Z-Q. Zeng, H-B. Yu, H-R. Xu, Y-Q. Xie, and J. Gao, "Fast training Support Vector Machines using parallel sequential minimal optimization", *2008 3rd International Conference on Intelligent System and Knowledge Engineering,,* 2008

[8] D.J. Lipman, and W.R. Pearson, "Rapid and sensitive protein similarity searches", *Science,* vol. 227, no. 4693, pp. 1435-1441, 1985.
[http://dx.doi.org/10.1126/science.2983426] [PMID: 2983426]

[9] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, and A. Bairoch, *Protein Identification and Analysis Tools on the ExPASy Server.* The Proteomics Protocols Handbook, 2005, pp. 571-607.
[http://dx.doi.org/10.1385/1-59259-890-0:571]

[10] M. Hebditch, M.A. Carballo-Amador, S. Charonis, R. Curtis, and J. Warwicker, "Protein-Sol: a web tool for predicting protein solubility from sequence", *Bioinformatics,* vol. 33, no. 19, pp. 3098-3100, 2017.
[http://dx.doi.org/10.1093/bioinformatics/btx345] [PMID: 28575391]

[11] J.R. Quinlan, "Combining Instance-Based and Model-Based Learning", *Machine Learning Proceedings,* vol. 1993, pp. 236-243, 1993.

[12] S. Le Crom, F. Devaux, C. Jacq, and P. Marc, "yMGV: helping biologists with yeast microarray data mining", *Nucleic Acids Res.,* vol. 30, no. 1, pp. 76-79, 2002.
[http://dx.doi.org/10.1093/nar/30.1.76] [PMID: 11752259]

[13] A.K. Upadhyay, and R. Sowdhamini, "Genome-Wide Prediction and Analysis of 3D-Domain Swapped Proteins in the Human Genome from Sequence Information", *PLoS One,* vol. 11, no. 7, p. e0159627, 2016.
[http://dx.doi.org/10.1371/journal.pone.0159627] [PMID: 27467780]

Gene cluster identification for secondary metabolite production in *Oryza sativa japonica*

Himanshu Singh, Vikas Kaushik

Department of Biotechnology, School of Bioengineering and Biosciences, Lovely Professional University, Phagwara, Punjab.

Abstract:

Plants are able to produce various kinds of secondary metabolites. These secondary metabolites are a rich source of bioactive compounds that can be utilized as medicinal, therapeutic and agrochemical agents. Rapid developments in computational biology pave the way to find out the mechanism for the production of plant metabolites. The recent development in this field provides information that genes encoded some of these metabolites is arranged in an operon like gene cluster. These clusters are governed by the same regulatory element. Advancement in molecular biology, genome mining, and analysis techniques provides us the opportunity to modify these genes for large scale production of these specialized metabolites. Rice which is a major staple food for the human population was utilized in the present study to find out the gene clusters able to synthesize specialized metabolites. In this present study, we have used plantismash tool to find out the gene clusters in Oryza sativa group japonica and indica. Plantismash tool is able to provide information about the annotation and expression analysis of plant gene clusters. We have found out the 39 gene clusters on 12 chromosomes of Oryza sativa group japonica.

Keywords: Plant Metabolites, BLAST, Gene clusters, Plantismash, *Oryza Sativa*, Secondary Metabolites

Introduction

Plant metabolites are specialized chemicals that provide important ecological functions, protection against various diseases and stresses. Plant secondary metabolites are used for various purposes by humans from ancient times. Bioactive compounds of these metabolites are used for various medicinal and therapeutic purposes. Recent advancement in molecular biological techniques and computational biology provides the opportunity to find the new metabolic pathways which are responsible for the production of these specialized chemicals [1]. It has recently found out that genes that are responsible for the formation of these metabolites are arranged in clusters and governed by the same regulatory element. Gene clusters are a group of two or more genes that are found in the DNA of any organism that collectively works in encoding similar proteins or enzymes [2]. Gene clusters work in various ways and they usually have the same promoter region by which all of them get expressed. It is believed till recent, that genes producing these secondary metabolites are scattered in the genome and express exclusive mutual. Secondary metabolic pathways consist of genes for ― signature enzymes that make the scaffold of the secondary metabolite, along with genes for― tailoring enzymes that carry out subsequent modifications to this scaffold [3]. Examples of plant signature enzymes are terpene synthases (for terpenes), chalcone synthases (for flavonoids), and CYP79 family enzymes for cyanogenic glucosides [3,4]. Examples of tailoring enzymes include oxidoreductases, methyltransferases, acyltransferases, and glycosyltransferases. A good candidate gene cluster will contain genes encoding a signature enzyme and tailoring enzymes. Given that these genes contribute to a common pathway, they may be expected to be tightly coexpressed, although this is not always the case (5).

Rice is the staple food of more than 3 billion people worldwide. A significant quantity of recommended niacin and zinc are provided by rice. The digestibility of rice protein is very high (88 per.) and thus is considered as biologically richest protein. After wheat, rice is considered second in the most important crops of the world [6]. Rice (*Oryza sativa*) is of family Poaceae and genus Oryza with over 20 cultivated wild species. *Oryza sativa* and *Oryza glaberrima* being the most cultivated rice species. Oryza sativa is grown worldwide while *Oryza glaberrima* has been cultivated for about the last 3500 years in West Africa. Rice contains a basic chromosome number of n = 12. The species' can be diploid as well as triploid. Both *Oryza sativa* and *Oryza glaberrima L*. being diploid species (2n = 24). Asian cultivated rice is the first fully sequenced crop genome [7].

Oryza sativa japonica rice varieties having a short plant. Grains of this group are short and round. These are having low amylase content. *Oryza sativa indica* group having tall light green leaves. Grains of this group are long and flat. They are having high amylose content.

Experimentation:

Collection of data:

The rice genome consists of 12 chromosomes. The genome of *Oryza sativa japonica* downloaded from Genbank, NCBI [8].

Plantismash:

It allows the fast far-reaching recognizable proof, explanation and examination of optional metabolite biosynthesis quality bunches over the plant kingdom. It is a particular augmentation of the generally utilized anti-SMASH webserver, customized particularly to target plant genomes. It first takes an input sequence and then searches for gene clusters in it taking standard sequences in the inbuilt database as a reference and shows results when and if gene clusters are found in the input sequence. In this software, we first give the email to which we want our results to be sent. In the second input box, we upload our file of genomic or nucleotide sequence in GenBank or EMBL format. The file as a whole is quite big in size. To avoid and cut off time consumption we can use the third input box. In this, we upload the NCBI accession no. of our desired file. Another option would be to segment the whole genome into their chromosomes and upload each individually. This helps in reducing time consumption significantly. Once the results are obtained we can analyze them for finding gene clusters, their size, location, and core domains [9].

BLAST:

Basic Local alignment search tool is an algorithms utilized for checking the sequence similarity in primary sequence of amino acids and nucleotides of genetic material i.e. DNA or RNA. Sequence similarity search was used to annotate the putative gene clusters present in chromosomes of *Oryza sativa*. [10]

Results and Discussion:

Chromosomes of *Oryza sativa japonica* run on plantismash webserver and the following results were obtained as per Table 1 and Table 2:

Table 1: Gene clusters of *Oryza sativa japonica* (chr-1 to chr-12)

Chromosome	Sr. No.	Gene cluster	Size (kb)	Core Domains
Chromosome 1	1.	Saccharide	71.44	2OG-FeII_Oxy, DIOX_N, Peptidase_S10, UDPGT_2
	2.	Lignan-Polyketide	70.90	Chal_sti_synt_C, Chal_sti_synt_N, Dirigent, p450
	3.	Saccharide	82.51	Aminotran_1_2, UDPGT_2
	4.	Saccharide	72.22	UDPGT_2, p450
	5.	Alkaloid	33.28	Bet_v_1, Epimerase, Methyltransf_11
Chromosome 2	1.	Saccharide	139.97	Glycos_transf_1, p450
	2.	Saccharide-Polyketide	211.17	Chal_sti_synt_C, UDPGT_2, p450
	3.	Terpene	369.98	COesterase, Terpene_synth, Terpene_synth_C, p450
Chromosome 3	1.	Lignan-Saccharide	97.55	Cellulose_synt, Dirigent, Methyltransf_11, UDPGT_2
	2.	Saccharide	64.12	Amino_oxidase, UDPGT_2, adh_short

Chromosome 4	1.	Terpene	212.71	Terpene_synth, Terpene_synth_C, adh_short_C2, p450
	2.	Saccharide-Alkaloid	360.51	Cu_amine_oxid, UDPGT_2, adh_short
	3.	Saccharide	169.20	Peptidase_S10, UDPGT_2
	4.	Terpene	334.35	2OG-FeII_Oxy, Terpene_synth, Terpene_synth_C
	5.	Saccharide	42.28	Peptidase_S10, UDPGT_2
	6.	Terpene	61.50	Terpene_synth, Terpene_synth_C, Transferase
	7.	Lignan	82.15	2OG-FeII_Oxy, DIOX_N, Dirigent, Methyltransf_7
Chromosome 5	1.	Saccharide	207.12	2OG-FeII_Oxy, DIOX_N, Transferase, UDPGT_2
Chromosome 6	1.	Putative	71.58	2OG-FeII_Oxy, DIOX_N
	2.	Putative	105.71	Peptidase_S10, Transferase, adh_short_C2
	3.	Saccharide	165.15	Transferase, UDPGT_2
	4.	Polyketide	133.31	Chal_sti_synt_C, p450
Chromosome 7	1.	Lignan	86.46	Aminotran_1_2, Dirigent
	2.	Lignan-Saccharide	88.18	Aminotran_1_2, Dirigent, Glycos_transf_1
	3.	Lignan	86.37	COesterase, Dirigent, p450
Chromosome 8	1.	Saccharide-Terpene	127.02	Methyltransf_2, Terpene_synth, Terpene_synth_C, UDPGT_2
	2.	Lignan-Alkaloid	132.28	Bet_v_1, Dirigent, Epimerase
	3.	Putative	83.82	COesterase, adh_short
Chromosome 9	1.	Saccharide	99.62	AMP-binding, UDPGT_2, p450
	2.	Putative	150.15	COesterase, Peptidase_S10, adh_short
Chromosome 10	1.	Saccharide	141.74	Transferase, UDPGT_2, p450
	2.	Lignan-Saccharide	432.20	Dirigent, UDPGT_2, p450
	3.	Polyketide	141.94	Acetyltransf_1, COesterase, Chal_sti_synt_C, Epimerase
	4.	Polyketide	139.12	Amino_oxidase, Chal_sti_synt_C, GMC_oxred_C, GMC_oxred_N
Chromosome 11	1.	Alkaloid	41.98	HMGL-like, Str_synth, p450
	2.	Lignan	130.12	Dirigent, Peptidase_S10
	3.	Saccharide	468.44	2OG-FeII_Oxy, UDPGT_2, adh_short, adh_short_C2
Chromosome 12	1.	Lignan	323.86	Dirigent, Methyltransf_2, p450
	2.	Saccharide	67.79	Glycos_transf_1, p450

- Results obtained from *Oryza sativa japonica* have the most occurrence of saccharide and putative gene clusters in chromosomes (1,2,3,4,5,6,8,9,10,11,12) and (8,9,6) respectively.

Analysis: We analyzed results obtained for putative clusters to find similarities in order to identify and predict their functions via Blast.

Japonica group Putatives:

- Chromosome 6, Cluster 1: Most similar sequence that was obtained was Proteosome subunit alpha-type. (Query coverage: 9%. Similarity: 100%.)
- Chromosome 6, Cluster 2: Most similar sequence that was obtained was Houba Copia like Retrotransposon. (Query coverage: 21%. Similarity: 100%.)
- Chromosome 8, Cluster 3: Most similar sequence that was obtained was Carboxylesterase. (Query coverage: 11%. Similarity: 97%.)
- Chromosome 9, Cluster 2: Most similar sequence that was obtained was Glycosyltransferase. (Query coverage: 17%. Similarity: 99%.)

Conclusion:

The gene clusters obtained via plantismash of rice chromosomes 1 to 12 of japonica group. The predominant cluster that was obtained was saccharide while Putative has second-most occurrence. Core domains that have the most occurrences are 2OG-FeII_Oxy, transferases, and peptidases.

References:

1.) Nützmann HW, Huang A, Osbourn A. , *New Phytol*, 2016, 211, 771-89.
2.) Yi, G., Sze, S.-H., & Thon, M. R., *Bioinformatics (Oxford, England)*, 2007, 23, 1053–1060.
3.) Osbourn, A., *Trends in Genetics*, 2010, 26, 449–457.
4.) D'Auria, J. C., & Gershenzon, J., *Current Opinion in Plant Biology*, 2005, 8, 308–316.
5.) Takos AM, Knudsen C, Lai D, Kannangara R, Mikkelsen L, Motawia MS, Olsen CE, Sato S, Tabata S, Jorgensen K, et al., *Plant Journal*, 2011, 68, 273–286.
6.) Ghosh, B., Ali, M.N. & Gantait, S. , *J. Res. Rice*, 2016, 4, 2–9.
7.) C-H Wang, X-M Zheng, Q Xu, X-P Yuan, L Huang, H-F Zhou, X-H Wei, S Ge, *Heredity*, 2014, 112, 489–496.
8.) Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW, Nucleic Acids Res., 2013, 41, 36-42.
9.) Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH, Nucleic Acids Res., 2017, 45, 55-63.
10.) Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, David J. Lipman, Journal of Molecular Biology, 1990, 215, 403–410.