Genome-wide identification and characterization of *HSF* gene family in *Vigna umbellata* (Ricebean)

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

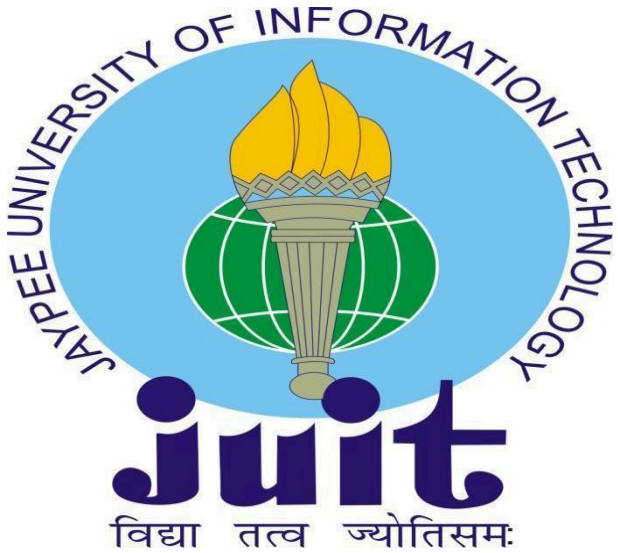
Bioinformatics

By

Saumya Tyagi (201901)
Vinamrata Sharma (201904)
Akansha Sharma (201906)

Under the supervision of

Dr. Shikha Mittal
(Assistant Professor (Grade I))



Department of Biotechnology & Bioinformatics
Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Certificate

This is to certify that the work reported in the B.Tech Project report "Genome-wide identification and characterization of HSF gene family in Vigna umbellata (Ricebean)" submitted by Ms. Saumya Tyagi, Ms. Vinamrata Sharma, Ms. Akansha Sharma at Jaypee University of Information Technology, Waknaghat, India, is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

21/05/2024

Supervisor

Date:20th May'2024

Dr. Shikha Mittal
Assistant Professor
Department of Biotechnology & Bioinformatics
Jaypee University of Information Technology
Waknaghat, India-173234

Candidate's Declaration

We hereby declare that the work presented in this report entitled "Genome-wide identification and characterization of HSF gene family in *Vigna umbellata* (Ricebean)." in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Bioinformatics** submitted in the Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology Waknaghat is an authentic record of our own work carried out over a period from August 2023 to May 2024 under the supervision of Dr Shikha Mittal.

We also authenticate that we have carried out the above mentioned project work under the proficiency stream of Major Project.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Saumya Tyagi (201901)

Vinamrata Sharma (201904)

Akansha Sharma (201906)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Shikha Mittal
Assistant Professor
Department of Biotechnology and Bioinformatics
20h May 2024

Acknowledgment

Presentation inspiration and motivation have always played a key role in the success of any venture.

We express our sincere thanks to **Dr. Shikha Mittal**, Assistant Professor (Grade I) of Jaypee University of Information Technology.

We pay our deepest sense of gratitude to encourage us to the highest peak and to provide us the opportunity to prepare the project. We are immensely obliged to our teachers for their elevating inspiration, encouraging guidance and kind supervision in the completion of our project.

Last, but not the least, our parents are also an important inspiration for us.

So with due regards, We express our gratitudes towards them.

Table of Content

List of Abbreviations

1. *HSF*: Heat Shock Factors
2. HSEs: Heat Stress Elements
3. HSPs: Heat shock Proteins
4. BLASTp: Basic Local Alignment Search Tool- Protein
5. NCBI: National Center for Biotechnology Information
6. Pfam: Protein family database
7. MSA: Multiple Sequence Alignment
8. MEME: Multiple Em for Motif Elicitation
9. MEGA: Molecular Evolutionary Genetics Analysis
10. GSDS Online Server: Gene Structure Display Server
11. ITOL Software: Interactive Tree Of Life

List of Figures

Abstract

This study investigates the stress response mechanisms in leguminous plants, particularly focusing on *Vigna umbellata* (ricebean) and its tolerance to heat stress. The research explores the connection between Vigna umbellata and other leguminous species like *Vigna mungo* (black gram), adzuki bean, chickpea *(Cicer arietinum*), *Arabidopsis thaliana* (model plant), and soybean (*Glycine max)* through the lens of Heat Shock Factors (*HSFs*) and their role in gene regulation during heat stress. The goal is to identify potential targets for crop improvement strategies that enhance heat tolerance and ensure food security in a changing climate.

The study employs a comparative approach, analysing the sequence similarity of *HSF* genes across these plant species. *Vigna umbellata* sequences containing the WRKY domain are isolated and characterised using phylogenetic analysis via Multiple Sequence Alignment (MSA) and phylogenetic tree construction with MEGA and ITOL software. Subsequently, the PlantCARE online tool is used to study the motifs within these WRKY domain sequences. Gene structures are then analysed using the GSDS server. Finally, a heat map is generated to visualise differentially expressed genes, allowing for the identification of upregulated and downregulated genes in response to heat stress. Understanding these expression patterns is crucial for pinpointing genes involved in heat tolerance mechanisms.

This research contributes to the understanding of plant stress responses and paves the way for developing heat-resistant crop varieties, thereby promoting food security in the face of climate change.

Chapter 1 : Introduction

Heat shock factors (*HSFs*) are a ubiquitous family of transcription factors known for their pivotal role in the heat stress response across all domains of life. In plants, *HSFs* activate a cascade of downstream genes that encode proteins essential for cell survival, repair, and acclimation to elevated temperatures. This intricate regulatory network safeguards plant growth and development under heat stress, a critical adaptation for agricultural success in a changing climate. *Vigna umbellata*, also known as ricebean, is a vital legume crop in tropical and subtropical regions. It thrives under harsh environments, including high temperatures and drought. Despite its resilience, the molecular mechanisms underlying ricebean's heat tolerance remain largely unexplored. Understanding the *HSF* gene family in this crucial crop holds immense potential for elucidating its heat stress response and ultimately, enhancing its thermotolerance[1]. This study delves into the genome-wide identification and characterization of the *HSF* gene family in Vigna umbellata. We employ a comprehensive bioinformatics approach to:

1. Identify and catalogue all *HSF* genes within the ricebean genome.
2. Analyse their gene structure, conserved domains, and phylogenetic relationships to uncover evolutionary insights and potential functional diversification.
3. Investigate expression patterns of *HSF* genes in different tissues and under heat stress conditions to identify key players in the ricebean heat response.
4. Compare and contrast the *HSF* family in ricebean with other legumes and model plant species to gain broader perspectives on heat stress response evolution.

Heat shock factors (*HSFs*) are a ubiquitous family of transcription factors recognized for their crucial role in the heat stress response across all domains of life. In plants, *HSFs* initiate a cascade of downstream gene activations that encode proteins essential for cell survival, repair, and acclimation to elevated temperatures[3]. This intricate regulatory network is vital for safeguarding plant growth and development under heat stress, an essential adaptation for agricultural success in the face of climate change.*Vigna umbellata*, commonly known as ricebean, is a significant legume crop in tropical and subtropical regions. It is known for its ability to thrive in harsh

environments, including high temperatures and drought. Despite its resilience, the molecular mechanisms underlying ricebean's heat tolerance remain largely unexplored. Understanding the HSF gene family in this crucial crop holds immense potential for elucidating its heat stress response mechanisms and, ultimately, enhancing its thermotolerance[4]. This study delves into the genome-wide identification and characterization of the *HSF* gene family in Vigna umbellata. We employ a comprehensive bioinformatics approach to:Identify and Catalogue *HSF* Genes: We will systematically identify and catalogue all *HSF* genes within the ricebean genome. This will involve utilising various bioinformatics tools to scan the genome for *HSF*-related sequences and verify their presence.Analyze Gene Structure and Conserved Domains: We will analyse the gene structure of identified *HSFs*, focusing on exon-intron organisation. Additionally, we will investigate conserved domains within these genes, such as the DNA-binding domain (DBD) and oligomerization domain (OD), to gain insights into their functional capabilities.Phylogenetic Relationships and Evolutionary Insights: We will construct phylogenetic trees to explore the evolutionary relationships between ricebean *HSFs* and those in other legumes and model plant species. This analysis will help uncover potential functional diversification within the *HSF* gene family. We will investigate the expression patterns of *HSF* genes in different ricebean tissues (e.g., leaves, stems, roots) under both normal and heat stress conditions. This will involve using techniques like quantitative real-time PCR (qRT-PCR) to identify key *HSFs* involved in the heat stress response. By comparing the *HSF* gene family in ricebean with those in other legumes and model plants, we aim to gain broader perspectives on the evolution of the heat stress response. This comparative analysis will highlight unique and conserved aspects of *HSF*-mediated heat tolerance across different species.Through these detailed investigations, we aim to elucidate the molecular mechanisms underlying ricebean's resilience to heat stress. The findings from our study could pave the way for developing heat-tolerant crops, contributing to agricultural sustainability in a changing climate.

Chapter 2 : Literature Review

Heat stress is a major environmental constraint limiting plant growth, productivity, and geographical distribution. Plants respond to heat stress by activating a complex signalling network involving heat shock proteins (HSPs) and heat shock factors (*HSFs*). *HSFs* bind to heat stress elements (HSEs) in the promoters of HSP genes, leading to their transcriptional activation and subsequent production of HSPs. HSPs function as molecular chaperones, assisting protein folding, preventing aggregation, and facilitating repair mechanisms, ultimately protecting the cell from heat-induced damage.

2.1 *HSF* gene family

HSFs are a large multigene family with diverse structural and functional features. Classified into several classes and subclasses based on sequence similarity and DNA-binding domain architecture[4]. Each class exhibits distinct expression patterns and functional specialisations in response to various stress stimuli. Genome-wide identification and characterization of *HSF* genes have been extensively studied in model plants like *Arabidopsis* and rice, providing valuable insights into heat stress response mechanisms.

2.2 *Vigna umbellata* (Ricebean)

A vital legume crop in tropical and subtropical regions, known for its tolerance to heat and drought. Despite its importance, the *HSF* gene family in ricebeans remains largely unexplored[5]. Understanding the *HSF* repertoire in this resilient legume could offer valuable clues to its heat stress adaptation strategies.

Limited studies have explored *HSFs* in legumes like soybean, chickpea, and *Medicago truncatula*. These studies revealed diverse *HSF* family compositions and highlighted potential legume-specific adaptations. Comparative analyses suggest lineage-specific gene expansions and functional divergence within the *HSF* family.

Lack of comprehensive information on the *HSF* gene family in ricebeans hinders our understanding of its heat stress response. This study aims to address this gap by:

- Identifying and characterising all *HSF* genes in the ricebean genome.
- Analysing their structural features, phylogenetic relationships, and expression patterns.
- Comparing the ricebean *HSF* family with other legumes and model plants to gain evolutionary insights.

This research will provide a foundational understanding of the *HSF* regulatory network in ricebean's heat stress response. Identification of key *HSF* players and their regulatory targets could pave the way for targeted manipulation to improve thermotolerance in ricebean and other legumes[5]. The findings will contribute to our knowledge of *HSF* evolution and adaptation in plants, particularly within the legume family. Overall, this study addresses a critical gap in our understanding of heat stress response in a vital legume crop with immense potential for agricultural improvement and broader knowledge advancement in plant stress biology.

2.3 Significance of the *HSF* gene family in *Vigna umbellata*

The *HSF* gene family in *Vigna umbellata* is significant for its role in stress tolerance. HSF genes are transcription factors that regulate the expression of genes involved in a variety of stress responses, including heat shock, drought, salinity, and heavy metal toxicity[6]. *Vigna umbellata* is a legume that is grown in tropical and subtropical regions. It is a valuable crop, but it is also susceptible to a number of abiotic stresses. The *HSF* gene family plays an important role in helping *Vigna umbellata* to tolerate these stresses. For example, studies have shown that *HSF* genes are upregulated in *Vigna umbellata* plants that are exposed to heat stress, drought, salinity, and heavy metal toxicity. This upregulation leads to the increased expression of stress-responsive genes, which help the plant to cope with the stress. It is also involved in, such as seed germination, flowering, and defence against pathogens[7]. Overall, the *HSF* gene family is a significant component of the *Vigna umbellata* stress response system. By regulating the expression of stress-responsive genes, *HSF* genes help the plant to tolerate a variety of abiotic stresses and maintain its productivity.

2.4 *Vigna mungo*: A Nutritional Powerhouse and Versatile Legume

Vigna mungo, also known as black gram, urad dal, or black lentil, is a small, yet mighty legume native to India. It's a staple food in South Asian and African cuisines, prized for its earthy flavour and impressive nutritional profile. Packed with protein, fibre, vitamins, and minerals, black gram is a true nutritional powerhouse[8].

Vigna mungo thrives in harsh environments, particularly those characterised by scorching temperatures. This resilience isn't magic; it's a finely tuned orchestra of heat stress response genes, with heat shock factors (*HSFs*) playing the lead violin. While *Vigna mungo's* genome holds the score, much of the music remains unplayed, waiting to be deciphered.

Existing research offers tantalising glimpses into this heat symphony. Studies have identified *HSF* genes in *Vigna mungo*, but their precise identities, functional roles, and interactions remain largely unknown. We know they belong to diverse classes, hinting at a complex regulatory network. Some *HSFs* might be soloists, inducible during extreme heat, while others might form harmonious duets or trios, constitutively safeguarding the plant. Unravelling this intricate score is crucial. By identifying key *HSF* players and their target genes, we can unlock the secrets of *Vigna mungo's* heat tolerance. This knowledge can be used to compose novel melodies, manipulating *HSFs* to breed even more heat-resistant black grams, not just for the benefit of this vital crop, but potentially for other legumes as well.

The *HSF* symphony in *Vigna mungo* is a captivating work in progress. Each gene identified, each interaction discovered, adds another layer of richness to this heat stress response masterpiece. By listening closely, we can learn to conduct this symphony ourselves, ensuring a more resilient future for agriculture in a warming world.

2.5 Soyabean

Soybean is a vital legume crop in tropical and subtropical regions. It thrives under harsh environments, including high temperatures and drought. Despite its resilience, the molecular mechanisms underlying ricebean's heat tolerance remain largely unexplored. Understanding the *HSF* gene family in this crucial crop holds immense

potential for elucidating its heat stress response and ultimately, enhancing its thermotolerance[9].

HSFs act as molecular switches, flipping on heat-protective genes when temperatures soar. These guardian genes encode heat shock proteins (HSPs), molecular chaperones that safeguard cellular machinery from heat-induced damage. Think of HSPs as protein firefighters, extinguishing the flames of stress before they engulf the cell.

Soybean boasts a diverse *HSF* family, each member fine-tuned for a specific stress response. Comprising at least 34 members, these *HSFs* fall into distinct classes based on their structure and function.

Understanding these soybean *HSFs* is crucial for developing heat-resilient varieties. Scientists are already unlocking the secrets held within their genes. By identifying key *HSFs* and their regulatory pathways, researchers can potentially breed soybeans with enhanced heat tolerance, ensuring their survival and productivity in a warming world.

The soybean *HSFs* are a testament to the power of adaptation. They stand as a testament to the relentless battle between plants and their environment, a battle where every genetic weapon counts. As we continue to unravel their mysteries, we inch closer to safeguarding not just soybeans, but the future of food security itself.

2.6 Chickpea

Chickpeas, a vital legume crop in arid regions, face a constant battle against the blazing sun. Heat stress disrupts their growth and yield, but luckily, they have a team of tiny superheroes up their sleeve – heat shock factors (*HSFs*). These protein warriors bind to specific DNA sequences, activating genes that produce heat-protective proteins called chaperones. These chaperones act like molecular shields, safeguarding chickpea cells from heat damage and ensuring their survival[10].

Researchers have identified over 50 *HSF* genes in chickpeas, each with unique talents. Some *HSFs* are early responders, quickly activating basic defence mechanisms. Others are more specialised, tackling specific heat-induced challenges like protein misfolding or oxidative stress. This diverse *HSF* team allows chickpeas to mount a multi-pronged defence against the scorching heat.

Studies have revealed fascinating insights into chickpea *HSFs*. One group, the HSFB family, seems to be particularly adept at sensing and responding to heat. They exhibit rapid activation and intricate interactions with other stress-related genes, highlighting their crucial role in chickpea's heat tolerance.

Intriguingly, some *HSFs* appear to be more active in specific tissues, like the developing seeds or the root tips. This targeted action suggests that different parts of the chickpea plant have unique heat stress vulnerabilities and rely on specialised *HSFs* for protection.Understanding these *HSF* guardians is crucial for developing heat-resilient chickpea varieties[10]. By identifying the most effective *HSFs* and the genes they regulate, scientists can potentially breed chickpeas with enhanced heat tolerance. This could safeguard chickpea production in a warming climate and ensure food security for millions who rely on this vital crop. The chickpea-*HSF* story is a testament to the incredible resilience hidden within plants. By unravelling the secrets of these tiny heat warriors, we can not only protect chickpeas but also learn valuable lessons about how life adapts and thrives in the face of adversity.

2.7 *Arabidopsis*

Arabidopsis thaliana, the humble thale cress, has become a powerhouse in plant biology research. Its small stature, rapid life cycle, and well-annotated genome have made it the go-to model plant for unravelling the mysteries of plant life, including how plants cope with heat stress. One crucial player in this heat drama is the heat shock factor (*HSF*) gene family. These special genes encode proteins that act as master regulators, turning on a cascade of other genes when the temperature rises. Imagine them as firemen, swiftly sounding the alarm and coordinating the cellular firefighting crew.

Arabidopsis boasts a diverse *HSF* family, with around 20 members identified so far. Each *HSF* has its own unique fingerprint, allowing it to respond to specific heat intensities or durations. Some *HSFs* are early birds, activated within minutes of a heat wave, while others are late bloomers, taking their time to fine-tune the response. This intricate timing and teamwork are essential for *Arabidopsis* to survive and thrive in a fluctuating thermal world. *HSFs* orchestrate the production of heat shock proteins

(HSPs), the firefighters that put out the metaphorical flames of stress. HSPs help refold damaged proteins, prevent protein aggregation, and maintain cellular integrity, ensuring that vital processes can continue even in the heat[11]. Understanding how *HSFs* work in *Arabidopsis* has provided valuable insights applicable to other plants, including crops. By studying the *HSF* family in heat-tolerant plants like *Arabidopsis*, scientists are hoping to identify key genes and regulatory pathways that could be transferred to improve the heat resilience of less fortunate crops, ultimately safeguarding our food security in a warming climate.

Arabidopsis's HSF story is far from over. Researchers are actively exploring how these genes interact with other signalling pathways, how they fine-tune their responses to different types of stress, and how they might be influenced by environmental factors. As we delve deeper, we are uncovering a fascinating world of heat-sensing and adaptation, all thanks to the tiny Arabidopsis and its dedicated HSF crew.

2.8 *Vigna adzuki*

Vigna adzuki, also known as azuki bean, is a vital legume crop in East Asia, prized for its nutritional content and adaptability. Unlike its close cousin, the mung bean, *Vigna adzuki* exhibits remarkable tolerance to heat stress, a trait crucial for its survival in harsh environments. *HSFs* are master conductors in the orchestra of heat stress response. When temperatures rise, these vigilant proteins bind to specific DNA sequences called heat stress elements (HSEs), triggering the production of heat shock proteins (HSPs). HSPs act as cellular chaperones, safeguarding vital proteins from unfolding and malfunctioning under heat, thereby ensuring cell survival[12].

Vigna adzuki's genome harbours a diverse *HSF* ensemble, each member fine-tuned to respond to different intensities and durations of heat stress. Studies have identified and categorised these *HSFs* based on their structural features and DNA-binding specificities. Interestingly, *Vigna adzuki* seems to possess a unique *HSF* repertoire compared to other legumes, hinting at potential evolutionary adaptations for heat tolerance. One fascinating member of the *Vigna adzuki HSF* family is VaHSFA2. This *HSF* springs into action during prolonged heat exposure, unlike its brethren who respond swiftly to initial temperature spikes. VaHSFA2 then orchestrates the

expression of a specific set of HSPs, bolstering the cell's defence against long-term heat stress. This delayed but specialised response suggests a carefully orchestrated heat tolerance strategy in *Vigna adzuki*.

Further research delves into the intricate interplay between different *HSFs* and their downstream HSP targets. Scientists are deciphering the complex regulatory networks these HSFs conduct, uncovering how they fine-tune the heat stress response to ensure *Vigna adzuki's r*esilience[12]. Understanding the *HSF* symphony in *Vigna adzuki* not only sheds light on its heat tolerance mechanisms but also holds immense potential for crop improvement. By learning how *Vigna adzuki* conducts its heat stress response, we can potentially transfer this knowledge to other legumes, enhancing their thermotolerance and safeguarding vital food crops in a warming world.

The *Vigna adzuki-HSF* saga is a testament to the intricate dance between plants and their environment. By unravelling the melody of *HSFs*, we gain valuable insights into plant resilience and pave the way for a more robust and climate-assured future for agriculture.

Chapter 3: Project Objective

The project focuses on the In-Silico Identification of *HSF* Gene Family:

- In-silico identification of the Heat Shock Factor (*HSF*) gene family in Ricebean (*Vigna umbellata*).
- Unravel specific *HSF* genes in the genomic makeup of Ricebean.
- Contribute to understanding heat stress response mechanisms in Ricebean.
- Collecting nucleotide sequences from NCBI for *Vigna mungo*, chickpea, *Arabidopsis thaliana*, soybean, and *Vigna adzuki*.
- Authenticating candidate *HSF* proteins through a BLASTp search with a minimum similarity percentage of 85%.
- Utilise publicly available RNA-Seq data.Analyze expression of identified *HSF* genes in Ricebean.Scrutinize gene activity and behaviour under different conditions, especially heat stress.Genetic Insights:Integrate information from the Pfam database on the Heat Shock Factor (*HSF*) domain.
- Perform comparative genomic analysis.Broaden understanding of genetic components associated with stress response mechanisms.

Our Research Endeavours contribute to identification and authentication of *HSF* proteins.Delve into the complex genomic landscape of Ricebean.Pave the way for developing heat-tolerant crop varieties.Enhance agricultural resilience to dynamic environmental conditions.

Chapter 4: Materials and Methods

4.1 Research design

In the initial phase of our comprehensive research initiative, we meticulously gathered nucleotide sequences from the National Center for Biotechnology Information (NCBI) for several leguminous plants, namely *Vigna mungo*, Chickpea (*Cicer arietinum*), *Arabidopsis thaliana*, Soybean *(Glycine max)*, and *Vigna adzuki.* In conjunction with this data collection, we also acquired the genomic sequence of Vigna umbellata, commonly known as ricebean. This extensive dataset serves as the foundation for our investigation into the Heat Shock Factor (*HSF*) gene family across these plant species. To authenticate the candidate *HSF* proteins identified in *Vigna mungo*, Chickpea, *Arabidopsis thaliana*, Soybean, and *Vigna adzuki*, we conducted a rigorous BLASTp search, setting a stringent criterion of a similarity percentage equal to or exceeding 85%. This meticulous authentication process ensures the accuracy of our subsequent analyses. The comparison of *HSFs* across different plant species poses a considerable challenge owing to the inherent diversity of these factors and the unique adaptations of each species to diverse environmental conditions and stressors.

To further enrich our investigation, we accessed the Heat Shock Factor (*HSF*) domain information from the Pfam database. This domain, crucial for our understanding of the genetic makeup related to stress response, becomes a cornerstone in our comparative analysis. Our research endeavours involve the intricate process of sequencing the genomes of the aforementioned plant species. This genomic sequencing is instrumental in elucidating the underlying genetic components, particularly the genes associated with stress response mechanisms, such as *HSFs*.

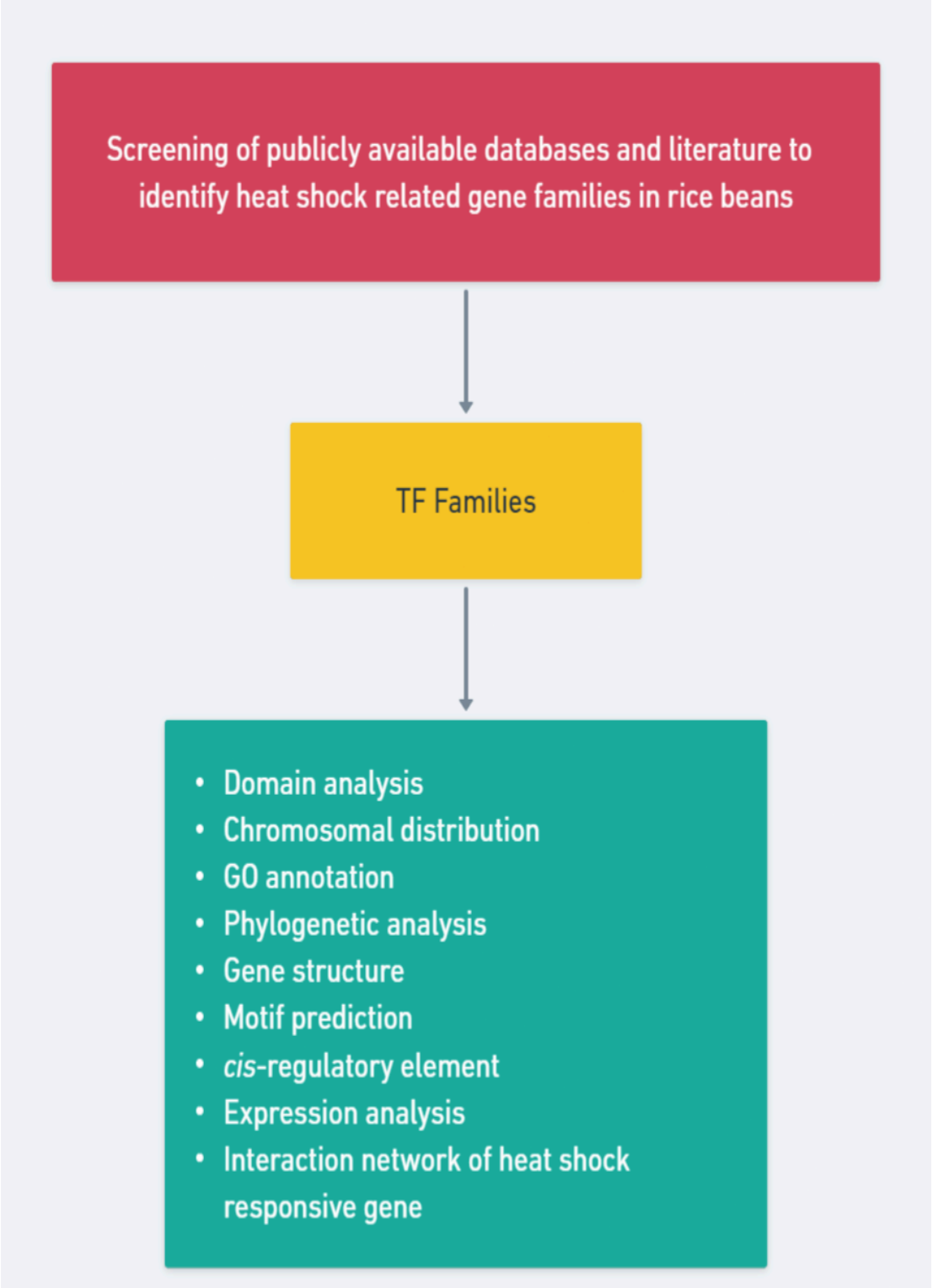


Fig. 1 Methodology chart

This multifaceted approach encompasses not only the identification and authentication of *HSF* proteins but also delves into the complex realm of genomic makeup, paving the way for a comprehensive understanding of the genetic intricacies governing heat stress responses across diverse leguminous plant species. Through these meticulous analysis, we aim to contribute substantial insights that can potentially inform the development of heat-tolerant crop varieties, fostering agricultural resilience in the face of evolving environmental conditions and ensuring sustained global food security.

4.2 Data collection

The genome sequence of the *Vigna umbellata, Vigna mungo*, Chickpea, Soybean, *Vigna adzuki* and *Arabidopsis* genome was obtained from the NCBI. Firstly, the candidate *HSF* proteins of all except *Vigna umbellata* were authenticated by a BLASTp search. Then, we downloaded the *HSF* domain (PF00447) from the Pfam database.

NCBI protein database by BLASTp. The results showed that 21 *HSF genes* were identified as heat transcription factors of *Vigna mungo*, 21 *HSF* genes were identified as heat transcription factors of *Arabidopsis,* 20 *HSF* genes were identified as heat transcription factors of Azuki bean, 25 Hsf genes were identified as heat transcription factors of Soyabean and 22 *HSF* genes were identified as heat transcription factors of chickpea.

Table 1. Genetic Details of species

S NO.	Species Name	Genome Size (mb)	Chromosome No.	No. of HSF genes
1	*Arabidopsis thaliana*	157	5	21
2	*Vigna adzuki*	950	22	20
3	*Vigna mungo*	670	22	21
4	G*lycine max*	1100	40	25
5	*Cicer arietinum*	830	16	22

4.3 Data analysis

Step 1: Database Creation

Initially, our research commenced with the creation of a comprehensive database. We meticulously gathered genetic sequences from five different plant species: Chickpea, *Arabidopsis*, Soybean, *Vigna mungo*, and *Vigna adzuki.* Each species' genetic information was systematically organised to establish a robust foundation for subsequent analyses.

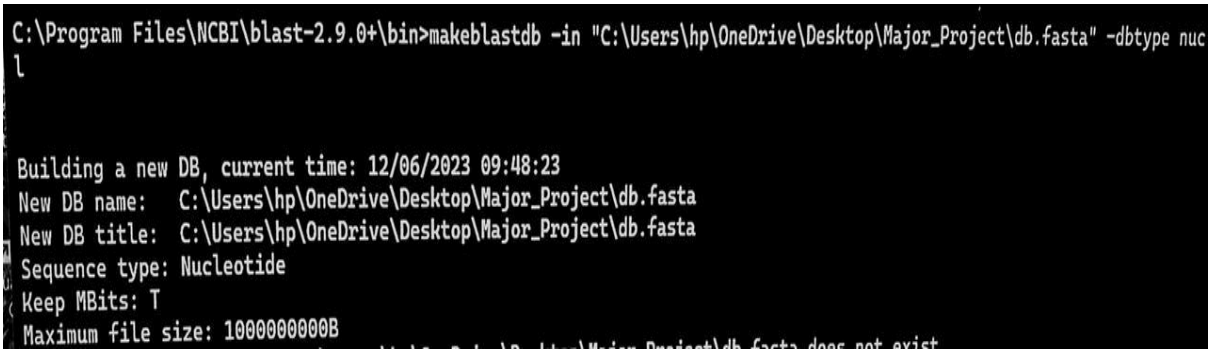
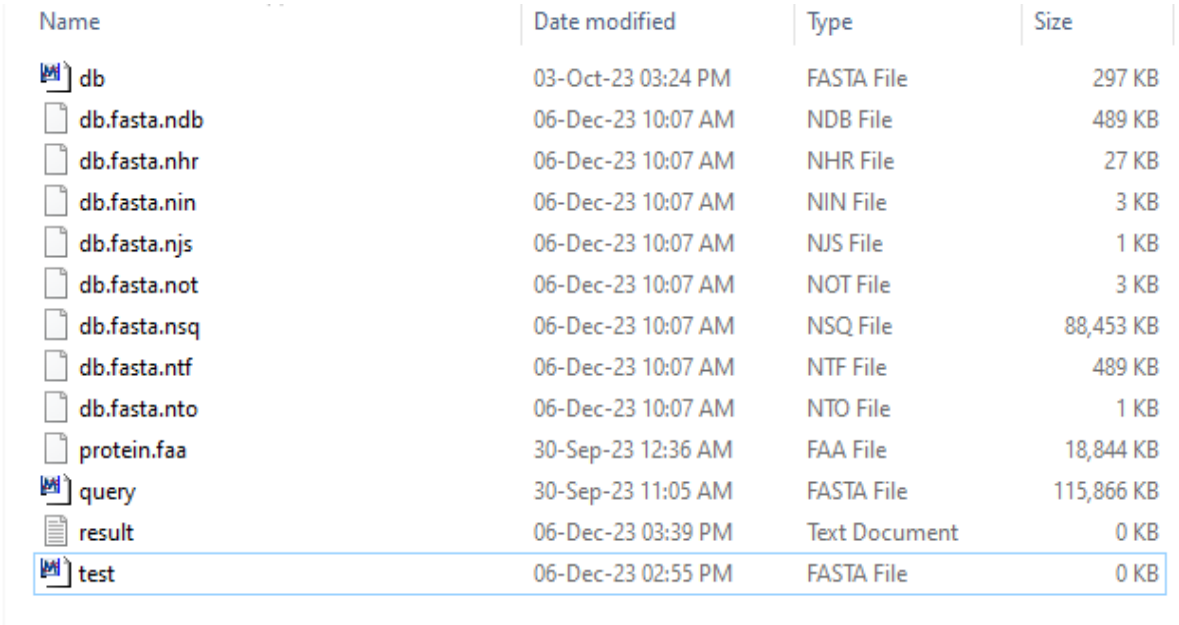


Fig. 2: Creating a database

Step 2: Database Preparation

Following the assembly of our genetic database, we undertook meticulous preparation of the files essential for the subsequent local BLAST analysis. This involved ensuring the accurate representation of nucleotide and protein sequences for each of the selected species within the database.



Name	Date modified	Type	Size
db	03-Oct-23 03:24 PM	FASTA File	297 KB
db.fasta.ndb	06-Dec-23 10:07 AM	NDB File	489 KB
db.fasta.nhr	06-Dec-23 10:07 AM	NHR File	27 KB
db.fasta.nin	06-Dec-23 10:07 AM	NIN File	3 KB
db.fasta.njs	06-Dec-23 10:07 AM	NJS File	1 KB
db.fasta.not	06-Dec-23 10:07 AM	NOT File	3 KB
db.fasta.nsq	06-Dec-23 10:07 AM	NSQ File	88,453 KB
db.fasta.ntf	06-Dec-23 10:07 AM	NTF File	489 KB
db.fasta.nto	06-Dec-23 10:07 AM	NTO File	1 KB
protein.faa	30-Sep-23 12:36 AM	FAA File	18,844 KB
query	30-Sep-23 11:05 AM	FASTA File	115,866 KB
result	06-Dec-23 03:39 PM	Text Document	0 KB
test	06-Dec-23 02:55 PM	FASTA File	0 KB

Fig. 3: Files created after running local blast

Step 3: Selection of Species for Analysis

Five distinct plant species were chosen for our analysis:

1. Chickpea
2. *Arabidopsis*

3. Soybean

4. *Vigna mungo*

5. *Vigna adzuki*

Step 4: Local BLASTp Analysis

With the database and files in place, we executed a local BLASTp analysis on a specific query file. This file contained nucleotide and protein sequences derived from Vigna umbellata, serving as the focal point of our investigation.

```
C:\Program Files\NCBI\blast-BLAST_VERSION+\bin>blastn -query "C:\Users\Bioinformatics\Desktop\MP
\query.fasta" -db "C:\Users\Bioinformatics\Desktop\MP\db.fasta" -out "C:\Users\Bioinformatics\De
sktop\MP\result.txt" -outfmt 6 -evalue 1e-10
```

Fig. 4: Analysis of query file

In the initial stages of our research, we embarked on a systematic process that unfolded in several key steps. Firstly, a comprehensive genetic database was meticulously crafted, encompassing genetic sequences from five diverse plant species: Chickpea, Arabidopsis, Soybean, Vigna mungo, and Vigna adzuki. This foundational database was crucial for ensuring the accuracy and completeness of the subsequent analyses. Following this, meticulous preparation of files was undertaken to ready the database for local BLAST analysis, ensuring that nucleotide and protein sequences for each species were accurately represented. Subsequently, we identified and selected five plant species for in-depth analysis: Chickpea, Arabidopsis, Soybean, Vigna mungo, and Vigna adzuki. The culmination of these preparatory steps led us to execute a local BLASTp analysis on a specific query file, containing nucleotide and protein sequences from Vigna umbellata. This sophisticated bioinformatics approach allowed us to explore potential genetic connections, similarities, and variations within *Vigna umbellata* compared to the reference species. The integration of cutting-edge computational techniques not only advanced our understanding of plant genomics but

also laid the groundwork for unravelling the molecular intricacies and evolutionary relationships of Vigna umbellata, marking a pivotal step in our broader exploration of plant molecular dynamics.

Further we analysed the nucleotide sequences having more than 85% similarity to the bioinformatics software to additionally characterise the *HSF* genes in *Vigna umbellata.*

We utilised the following softwares:

- **Pfam Software:** Pfam is a comprehensive database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models[13]. In our study, we utilised Pfam to identify and analyse the conserved domains within the Heat Shock Factor (HSF) proteins of Ricebean. By accessing the Pfam database, we could compare our sequences against a vast collection of known protein families, ensuring accurate identification of functional domains critical for heat stress response. This step was essential for confirming the presence of HSF-specific domains, which underpin the regulatory roles of these proteins.

- **Molecular Evolutionary Genetics Analysis (MEGA) Software:** MEGA software was employed for constructing phylogenetic trees to elucidate the evolutionary relationships between the HSF genes identified in Ricebean and those in other plant species[14]. Using MEGA's robust algorithms, we performed multiple sequence alignments and phylogenetic analyses to understand the divergence and conservation of HSF genes across different legumes and model plants. This analysis provided insights into the evolutionary history and potential functional diversification of the HSF gene family, which is critical for understanding their role in heat stress adaptation.

- **TBTools Software:** TBTools, a toolkit for biologists integrating various bioinformatics functions, was instrumental in our analysis of the HSF gene family[15]. We used TBTools for gene annotation, visualisation, and statistical

analysis of our sequence data. Its user-friendly interface and comprehensive functionalities allowed us to efficiently manage large datasets, annotate HSF genes, and visualise their distribution and structural features within the Ricebean genome. This facilitated a clearer understanding of the genomic organisation and functional potential of HSF genes.

- **Interactive Tree Of Life (ITOL) Software:** The Interactive Tree Of Life (ITOL) software was used to visualise and annotate the phylogenetic trees generated during our study[16]. ITOL's interactive features enabled us to create detailed, publication-quality tree diagrams that highlight the relationships among HSF genes from different species. By incorporating various data layers, such as domain architectures and expression patterns, we provided a comprehensive visual representation of HSF gene evolution and their functional implications in heat stress response.

- **Multiple Expectation maximisation for Motif Elicitation (MEME) Software:** MEME software was employed to discover and analyse motifs within the HSF proteins[17]. By using MEME, we identified conserved sequence motifs that are characteristic of HSF proteins, providing insights into their functional regions and regulatory elements. This analysis helped us pinpoint specific motifs that are crucial for the DNA-binding and transcriptional activation roles of HSFs, further elucidating their mechanisms of action in response to heat stress.

- **Gene Structure Display Server (GSDS) Software:** The Gene Structure Display Server (GSDS) was used to analyse and visualise the exon-intron structures of the identified HSF genes. GSDS allowed us to compare the gene architectures across different HSF genes, highlighting variations and conserved features[18]. This structural analysis provided insights into the evolutionary dynamics of the HSF gene family and helped us understand how gene structure might influence their function and regulation under heat stress conditions.

- **PlantCARE Software:** PlantCARE, a database of plant cis-acting regulatory elements, was utilised to analyse the promoter regions of the HSF genes identified in Ricebean[19]. By scanning these regions for known regulatory elements, we could predict the potential regulatory mechanisms governing the expression of HSF genes. This analysis was crucial for identifying heat-responsive elements and other regulatory motifs that might play a role in the transcriptional regulation of HSF genes during heat stress, thereby contributing to our understanding of their functional regulation in plant stress responses.

Chapter 5: Results and Discussion

Following the initial sequence analysis, a refinement process was undertaken to obtain a focused set of sequences suitable for further investigation. The initial set of sequences was filtered based on sequence similarity. Sequences exhibiting a similarity exceeding 85% at the amino acid level were selected. This stringent threshold (85% or higher) resulted in a collection of 119 sequences. This selection process ensures that the subsequent analyses focus on a set of closely related sequences, potentially sharing similar functions or evolutionary origins. A further refinement step involved the exclusion of sequences containing a specific domain known as WRKY Domain. This targeted exclusion suggests a specific research focus, potentially excluding sequences with this particular domain because it's not relevant to the study's objectives. This step resulted in a final set of 98 sequences. The final set of 98 sequences displayed a crucial characteristic: homology at both the nucleotide and protein levels. Homology implies a shared evolutionary origin, indicating that these sequences likely diverged from a common ancestor. This finding strengthens the rationale for further analysis of these sequences as a potentially related group. To facilitate downstream analyses, unique identifiers (locIds) were retrieved for each of the 98 sequences. These locIds likely correspond to specific locations within a genome database, allowing for future reference and retrieval of the sequences. Additionally, a spreadsheet was generated to organise this data efficiently. This ensures proper data management and traceability for subsequent analyses.

5.1 Pfam Analysis

Once we had a set of protein sequences, we compared them to find very similar sequences, defined as those that had at least 85% of the same amino acid sequences. Next, an analysis was conducted on these extremely similar sequences using the Pfam database. Pfam is a tool that facilitates the identification of functional domains in proteins. We identified 119 distinct sequences that most likely cover every possible functional domain found in the initial collection of sequences through comparing their sequences to Pfam.

5.2 MEGA Analysis

We improved upon the collection of protein sequences, subsequently eliminated any sequences that contained the WRKY domain, leaving an additional set of 98 proteins. Significantly, there was similarity between these residual sequences at the amino acid and DNA levels, indicating a close functional or structural connection. The unique identifiers (locIds) for every sequence were obtained and combined into a spreadsheet in order to organise the data. Within the MEGA software, a Multiple Sequence Alignment (MSA) was likely performed based on pairwise distances. This MSA step aligns the protein sequences to identify conserved regions and potentially functional domains of the ricebean family and the other 5 families, providing a foundation for the subsequent phylogenetic analysis.

Following the generation of a phylogenetic tree file using tbtools software, the format was converted from the tbtool-specific format to the Newick tree format (.nwk). This conversion allows for broader compatibility with visualisation platforms. Next, phylogenetic analysis was performed on all 98 sequences using MEGA software. This contained both the sequences found by Pfam analysis and the first batch of sequences for five species that were taken from the NCBI database. Three objectives were set for this analysis: calculating statistical measures to evaluate the analysis's reliability; visualising the proteins' evolutionary relationships in a phylogenetic tree; and estimating the proteins' evolutionary distances, or how similar or different they are from one another.

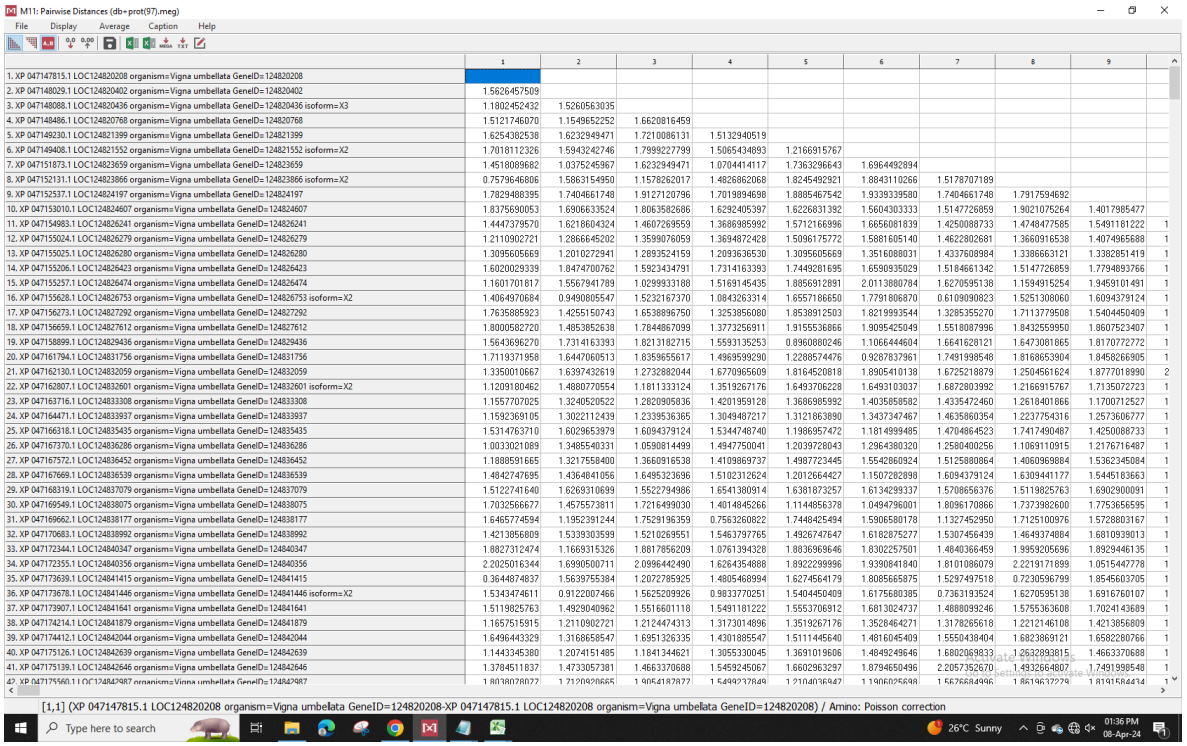


Figure 5: Pairwise Distance

5.3 TBTools for Phylogenetic Tree

Following the sequence analysis and potential alignment steps (mentioned previously), we employed tbtools software for the construction of a phylogenetic tree. tbtools is a versatile software suite specifically designed for various bioinformatics analyses, including phylogenetic tree generation. It offers a user-friendly interface and diverse algorithms for tree construction based on sequence data. While the specific algorithm used by tbtools for tree construction is not explicitly mentioned, some common methods for protein or nucleotide sequences include Neighbor-Joining, Maximum Likelihood, or Minimum Evolution. The choice of algorithm depends on various factors, such as the size and complexity of the data set and the desired level of accuracy. The tbtools software facilitated the generation of a phylogenetic tree file. This file likely represents the inferred evolutionary relationships between the analysed sequences in a specific format, potentially tbtool's proprietary ".tree file" format. This file encapsulates the branching patterns and evolutionary distances between the sequences within the data set.

5.4 ITOL Software Visualization

Following the generation of a phylogenetic tree using the tbtool software, we proceeded to visualise the tree for further analysis and presentation. To achieve this, the tree file, initially saved in the proprietary ".treefile" format of tbtool, underwent a conversion step. The ".treefile" format is specific to the tbtool software and may not be compatible with other visualisation platforms. Therefore, we employed a file format conversion tool (details of the specific tool can be included if known) to transform the ".treefile" into the Newick tree format (".nwk"). The Newick tree format is a widely recognized and text-based standard for representing phylogenetic trees. This conversion step ensures broader accessibility and compatibility with various tree visualisation software, including iTOL. Once converted to the ".nwk" format, the phylogenetic tree file was uploaded into the iTOL web application (https://itol.embl.de/help/gkw290.pdf). iTOL offers a user-friendly interface for interactive exploration and visualisation of phylogenetic trees. By employing the tbtool software for tree generation and the iTOL platform for visualisation, we were able to effectively analyse and present the inferred evolutionary relationships within the data set.

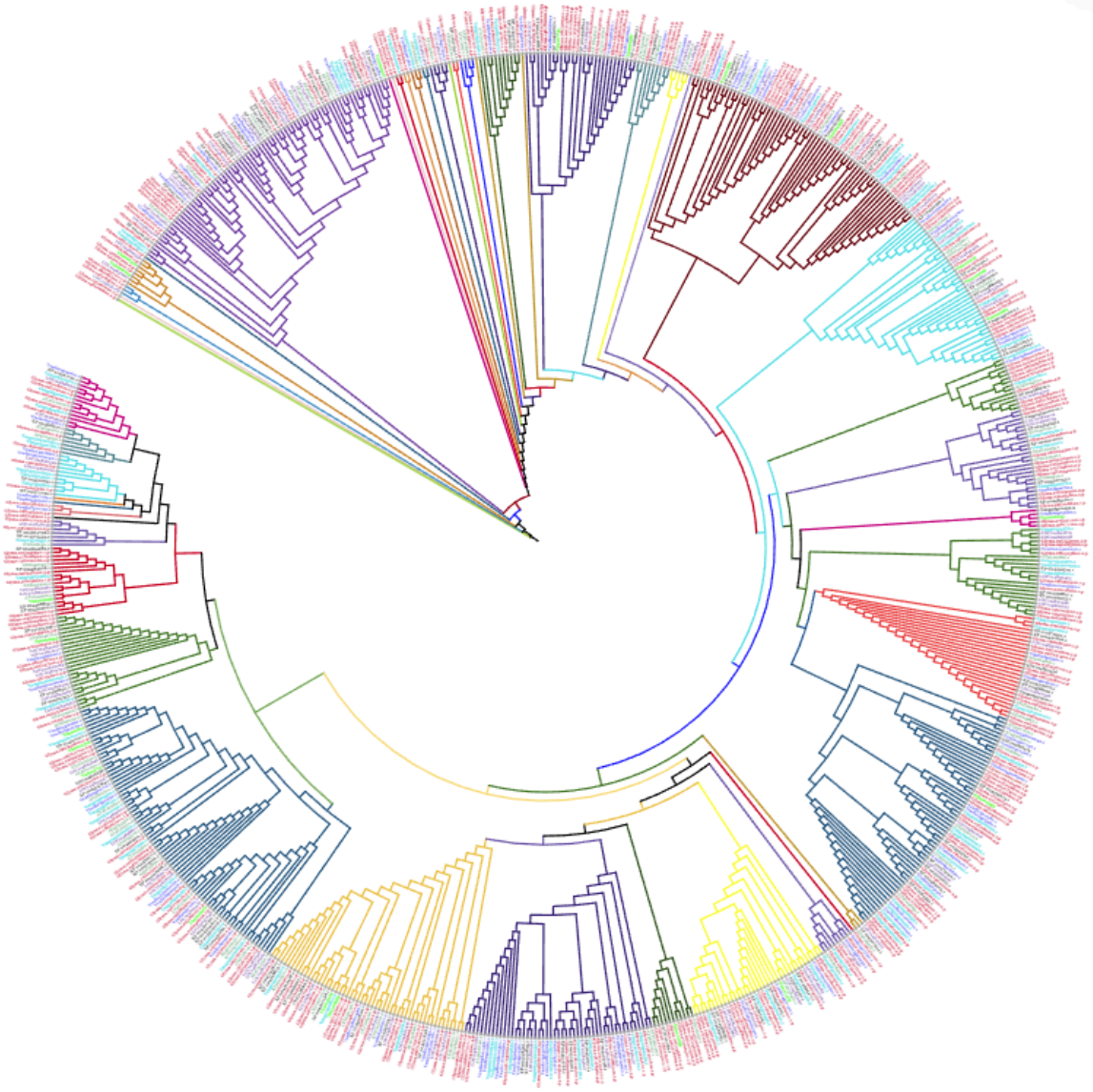


Figure 6: Phylogenetic tree Analysis for visualisation

5.5 MEME Software

We used the MEME Suite (Multiple Expectation for Motif Elicitation) programme to find motifs in the target protein sequences after completing the previously described sequence analysis processes. The well-known programme MEME was created especially to find ungapped sequence motifs, or recurrent patterns of particular lengths, in misaligned sequences. After utilising MEME to uncover motifs, we decided to concentrate our investigation on only a few of the motifs that were found. Five motifs were carefully chosen based on predetermined standards (please provide

the particular standards, e.g., statistical significance, biological significance, or sequence prevalence). This analysis identified five distinct motifs from the MEME Software with minimum widths of six amino acids and a maximum width of up to 50 amino acids. The identified motifs exhibited varying degrees of prevalence within the proteome, with the most abundant motif occurring at 97 distinct protein locations. This selection procedure lessens the complexity of subsequent analysis and enables a more focused examination. As input data, we used the protein sequences from our query file to help with motif finding using MEME. MEME was able to find probable motifs by analysing the appropriate protein sequences found in this query file, which was particularly created for ricebeans. MEME's motif discovery algorithm is based on the protein sequences, which enables it to look for recurrent patterns in the ricebean proteome.

Motif Site Distribution	ZOOPS: Zero or one site per sequence
Objective Function	E-value of product of p-values
Starting Point Function	E-value of product of p-values
Site Strand Handling	This alphabet only has one strand
Maximum Number of Motifs	5
Motif E-value Threshold	no limit
Minimum Motif Width	6
Maximum Motif Width	50
Minimum Sites per Motif	2
Maximum Sites per Motif	97
Bias on Number of Sites	0.8
Sequence Prior	Mega-weight Dirichlet Mixture Plus
Sequence Prior Source	prior30.plib
Sequence Prior Strength	185695
EM Starting Point Source	From substrings in input sequences
EM Starting Point Map Type	Point Accepted Mutation
EM Starting Point Fuzz	120
EM Maximum Iterations	50
EM Improvement Threshold	0.00001
Maximum Search Size	37146
Maximum Number of Sites for E-values	1000
Trim Gap Open Cost	11
Trim Gap Extend Cost	1
End Gap Treatment	Same cost as other gaps

Figure 7: Motif summary

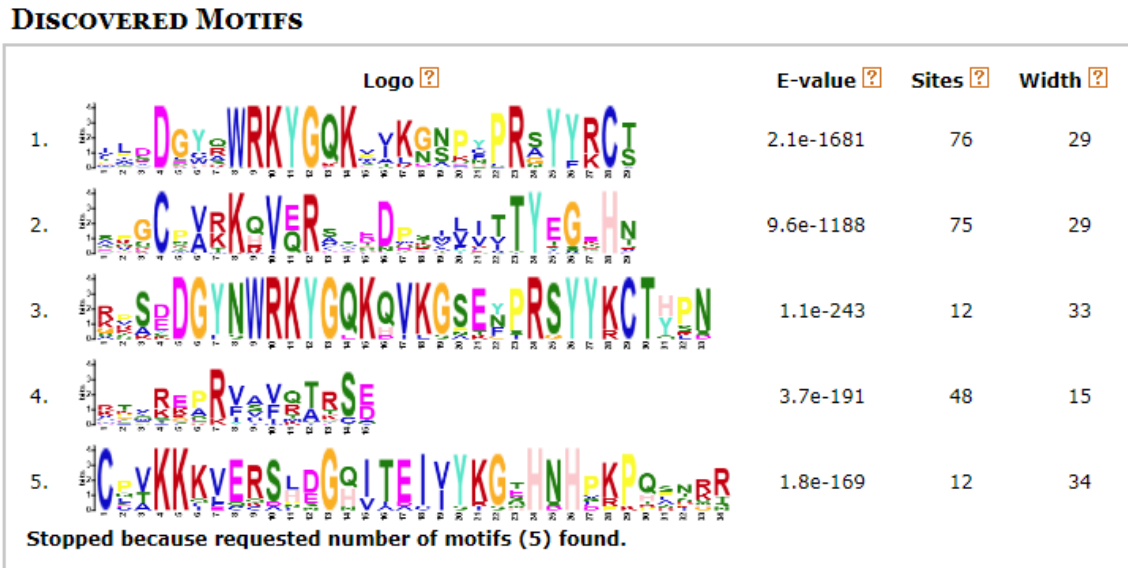


Figure 8: The sequence logos of the WRKY domain

5.6 GSDS Software

After 96 sequences were chosen for additional analysis, we looked at their gene architecture. To further characterise the 97 sequences most frequently occurring motifs, we employed the GDS server for in-silico gene structure prediction. This analysis yielded detailed annotations for each sequence, including the identification of coding sequences (CDS), untranslated regions (UTRs), and intronic regions. In order to do this, we made use of the annotation data that was already available for these sequences in the GFF (General Feature Format) format. Genes, exons, introns, and other properties inside a genome can be uniformly represented using the GFF format. With the purpose of creating graphical representations of gene structures based on GFF annotation data, GSDS is a useful tool. The GSDS server received the GFF files with the annotation data for the 96 chosen sequences supplied as input data. We then employed the Gene Structure Display Server (GSDS) – an online web application accessible at http://www.cbi.pku.edu.cn/resource/index.htm – to visualise and analyse the gene structures of the selected sequences. GSDS is a valuable tool specifically designed to generate graphical representations of gene structures based on GFF annotation data. The GFF files containing the annotation information for the 96 selected sequences were uploaded as input data to the GSDS server. We gave GSDS

the data it needed to correctly display the gene structures by uploading these files, which included the locations and relative sizes of exons, introns, and other pertinent elements inside each sequence. We were able to learn more about the structure and possible purposes of the genes that the examined sequences encoded thanks to this depiction.

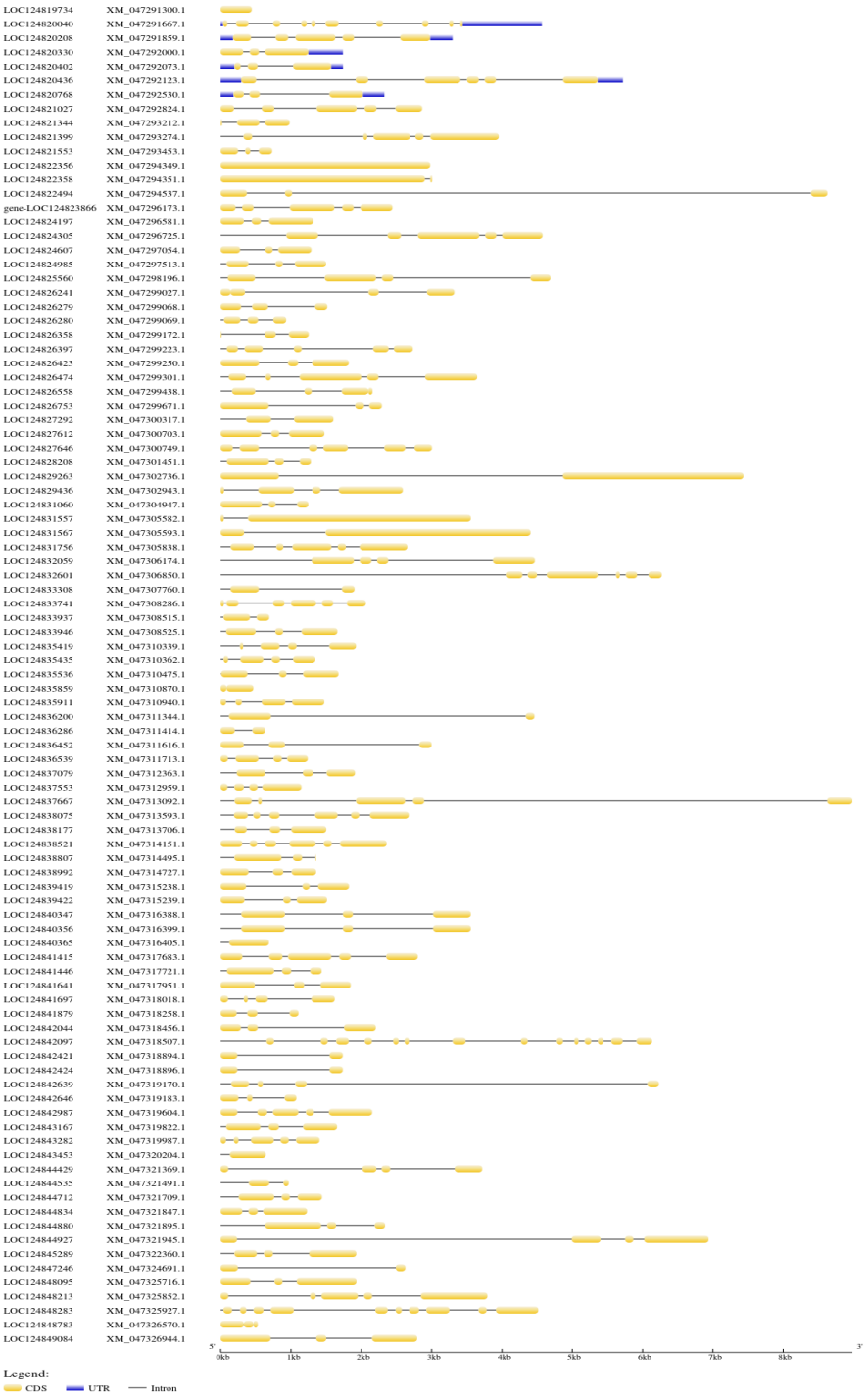


Figure 9: Gene Structure Visualization

5.7 PlantCARE analysis

DNA sequences totaling ninety-seven (97) were uploaded to the PlantCARE online programme. A database of plant cis-acting regulatory elements (CAREs) that are known to affect plant gene expression is called PlantCARE. These components have the ability to act as repressors, which prevent transcription, or enhancers, which increase transcription. The programme examines the sequences that have been supplied and uses consensus sequences and positional matrices that are kept in its database to find putative cis-regulatory motifs. The DNA sequence upstream of a gene that regulates transcription is known as the promoter region, and this is the area of emphasis for the investigation. Cis-regulatory element analysis was subsequently performed using the PlantCARE software to identify potential regulatory motifs associated with the 97 sequences. This analysis revealed the presence of several known cis-acting regulatory elements (CAREs) within the sequences, including the ABA-responsive element (ABRE motif), light-responsive element (AE-box motif), and binding sites for MYB and MYB-like transcription factors. After that, a heatmap was created to show how these motifs were expressed throughout the 97 LOCIDs (Locus IDs).

For each of the 97 sequences, PlantCARE's analysis produced one of two sorts of outputs:

HTML document: This style probably offers a clear, easy-to-understand visual depiction of the sequence's known cis-regulatory motifs. It might contain information such as the motif's name, position in the sequence, and possible purpose.

Excel spreadsheet: This format probably provides a tabular, more in-depth depiction of the motifs that were found. It might contain details like the motif name, base pair coordinates pinpointing its exact location, and possibly a score indicating how closely the sequence matches the established motif consensus. This approach allows for a comprehensive analysis of the cis-regulatory landscape within the 97 promoter sequences. The identified motifs can provide valuable insights into the potential regulation of the associated genes and their expression patterns.

Figure 10: Cis-Regulatory Motif Analysis in Vigna umbellata Gene Using PlantCARE

5.8 HeatMap Generation using TBTools

The data was subsequently processed to look into the gene expression patterns after the cis-regulatory element analysis. Two essential parts of an Excel sheet were used:

LOCID: This identity most likely relates to the Locus ID (LOCID) that has been assigned by a database, like Entrez Gene from the NCBI. A gene's LOCID within a particular organism provides a unique identity.

Presence or absence of a motif: Presumably, this data was taken from the earlier PlantCARE analysis. It probably shows whether certain cis-regulatory motifs are present in the promoter sequence of each gene or not. An application called tbtools was used to create a heatmap. Heatmaps are data visualisations where the colour intensity of a given value indicates its magnitude. The heatmap in this instance most likely shows the degree of expression (upregulated, downregulated, etc.) of the genes connected to every LOCID. Although it isn't stated clearly, the precise expression data source (such as RNA-seq data) would have been included with the LOCID-motif details. The purpose of this research is to find any possible relationships between the patterns of gene expression and the existence or lack of particular cis-regulatory motifs within promoters.

By utilising a colour gradient to identify motifs with high or low abundance, the heatmap can be an effective tool in uncovering regulatory components that control the expression of the HSF gene.The results of the heatmap shows that many of the cis regulatory elements are negatively correlated and only a few of them show positive correlation (Fig. 17).

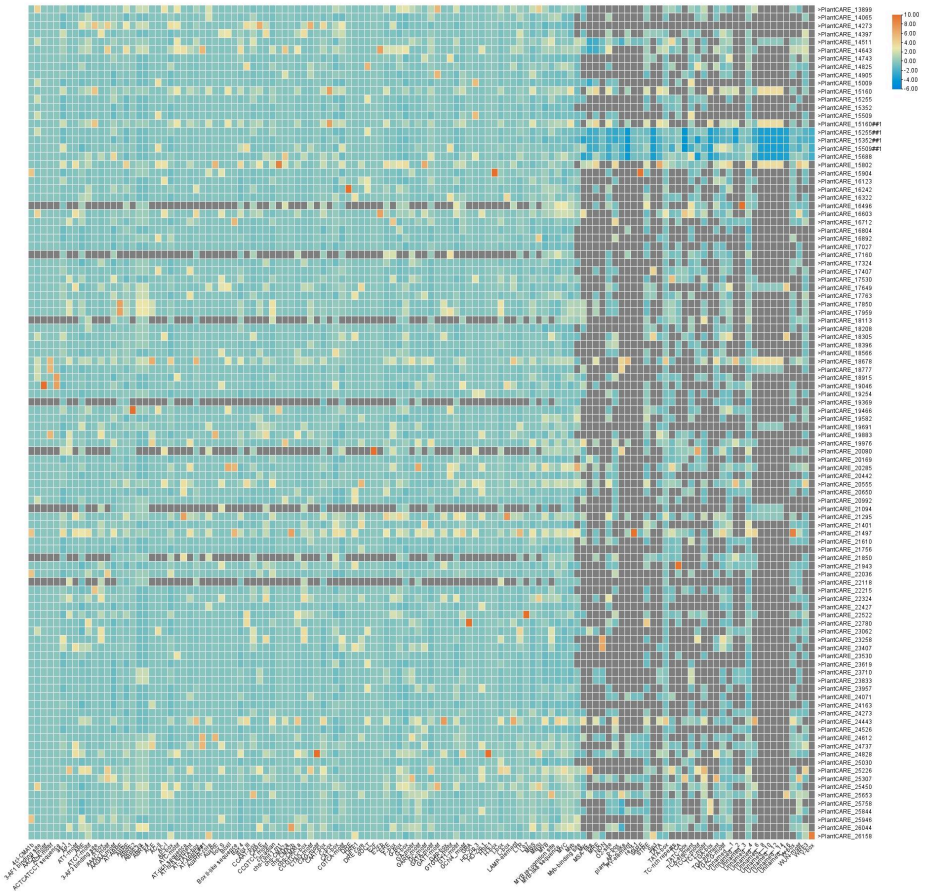


Figure 11: Heatmap of Cis-Regulatory Motifs Identified by PlantCARE using TBTools

5.9 Differential Expression Identification

After the gene families were fully characterised, a differential expression analysis was carried out. The goal of this approach is to find genes within the families that show statistically significant variations in their levels of expression across two or more experimental setups. A heatmap was created to show the gene expression patterns after the differential expression analysis. Heatmaps are two-dimensional graphical depictions in which the intensity of a colour corresponds to the magnitude of a given value. In this instance, the fold change in expression for every gene under the comparable conditions is probably shown by the heatmap. In the heatmap, genes with

comparable expression profiles will group together, making it possible to identify co-regulated genes and possible regulatory patterns within the described families.

With the use of heatmap visualisation and differential expression analysis, this combined technique offers a thorough insight of the dynamics of expression within the described gene families. It makes it easier to identify important genes that show notable variations in expression under particular circumstances, which may lead to the identification of novel functional roles for these genes and the families that are related to them. The 97 LOCIDs' expression patterns and the data on differential gene expression showed strong connections, according to our research. Under particular experimental conditions, a number of LOCIDs showed variable expression patterns, indicating their possible participation in regulatory processes. Finding groups of LOCIDs with comparable expression patterns was made easier by the heatmap visualisation (Figure 1), which also offered insights into possible co-regulation mechanisms. A heatmap visualisation showing the 97 LOCIDs' expression levels under various experimental settings. It is possible to identify clusters of LOCIDs with comparable expression patterns, which may indicate co-regulation mechanisms.

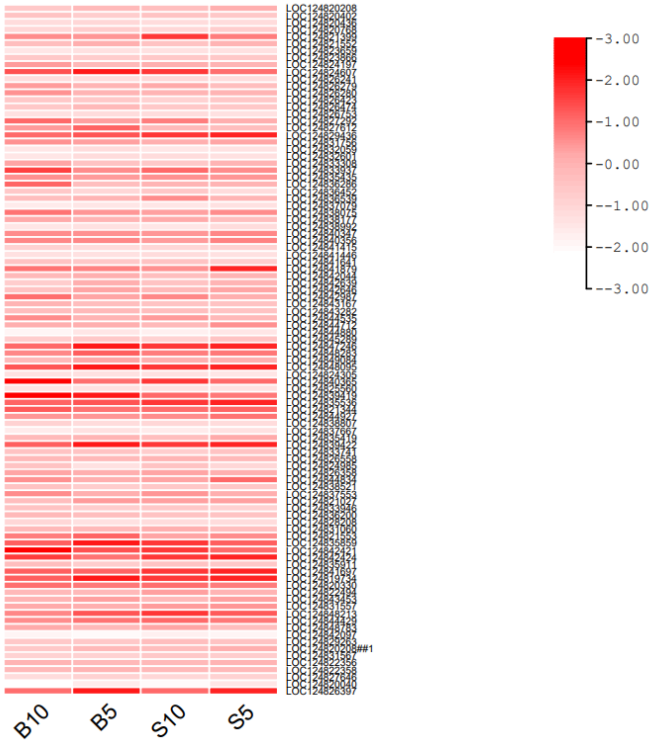


Figure 12: Heatmap of Differentially Expressed Genes Using TBTools

Chapter 6 : Challenges

The complexity of comparing Heat Shock Factors (*HSFs*) across different plant species poses a significant challenge due to the diverse nature of these factors and the specific adaptations each species has developed in response to varying environmental conditions and stresses. This diversity manifests in several key areas. HSFs exhibit considerable genetic variability among different plant species, reflecting their unique evolutionary paths and environmental adaptations. This genetic diversity complicates the identification and comparison of homologous HSFs across species, as it requires meticulous alignment and analysis to discern evolutionary relationships and functional similarities. The functional roles of HSFs can vary significantly among species, even among closely related legumes. This functional diversification means that HSFs might be involved in different regulatory networks or stress responses, necessitating a detailed understanding of their specific roles within each species[20]. Each plant species has evolved unique mechanisms to cope with heat stress, influenced by their native environments. These adaptations can result in distinct *HSF* expression patterns and regulatory pathways, making it challenging to draw direct comparisons or generalise findings across species. To address these challenges, the project employs a comprehensive approach involving several key methodologies. This involves using bioinformatics tools to align sequences, construct phylogenetic trees, and analyse gene structures and conserved domains. The project includes the functional characterization of *HSF* genes through expression analysis under heat stress conditions. By examining RNA-Seq data and identifying heat-responsive regulatory elements in the promoter regions, the study seeks to elucidate the roles of specific *HSFs* in the heat stress response[20]. Integrating data from various sources, such as Pfam for domain identification, MEME for motif discovery, and PlantCARE for regulatory element analysis, enhances the robustness of the findings. This multi-faceted approach provides a holistic view of the HSF gene family and its involvement in heat stress responses. This research endeavour aims to contribute valuable insights into the genetic mechanisms governing heat stress responses in leguminous plants. By addressing the challenges of genetic diversity, functional diversification, and

environmental adaptations, the study seeks to pave the way for the development of heat-tolerant crop varieties. Such advancements are crucial for ensuring food security amidst changing environmental conditions, as they enable the cultivation of resilient crops capable of withstanding increasing temperatures.

Chapter 7: Conclusion

In conclusion, the result analysis of the BLASTp yielded a diverse set of species detailed in the result.txt file, each showcasing varying degrees of similarity with our designated query sequence. In a concerted effort to refine our focus, we implemented a stringent criterion, selecting species whose genetic sequences exhibited similarities equal to or surpassing the 85% threshold. This meticulous approach ensured the inclusion of only the most closely related species, setting the stage for their subsequent in-depth analysis and evaluation within the broader context of the major project. By concentrating on species that demonstrate substantial genetic congruence with the query sequence, we aim to glean comprehensive insights into molecular intricacies. A thorough refinement procedure produced a final set of 98 potential HSF gene sequences. Both nucleotide and protein homology in this set point to a common evolutionary ancestor. To determine the divergence periods and clarify the evolutionary relationships between the sequences, phylogenetic analysis was applied. Five different motifs with variable degrees of occurrence were found in the ricebean proteome by motif finding. These sequences were further characterised using gene structure prediction. Potential regulatory elements linked to the *HSF* genes were found by cis-regulatory element analysis. These elements included motifs that were sensitive to different environmental cues. The ensuing heatmap display shed light on how these motifs expressed themselves throughout the sequences under analysis. All things considered, this information offers a useful starting point for further research on the discovered HSF gene candidates and the regulatory processes underlying the theme in *Vigna umbelleta.*

Chapter 8: References

[1] Mittler, Ron. “Abiotic stress, the field environment and stress combination.” Trends in plant science 11.1 (2006): 15-19.

[2]W. Wangxia, V. Basia, A. Arie, "Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance," Planta, vol. 218, no. 1, pp. 1–14, 2003. [Online]. DOI: 10.1007/s00425-003-1105-5.

[3] Nover, László, et al. “Heat shock and other stress response systems of plants.” The Plant Cell 6.10 (1994): 1539-1552.

[4] Scharf, Klaus-Dieter, Dietmar Berberich, and Dörte Ebersberger. “Nuclear factor-mediated defence of plants against heat stress.” Plant Cell and Environment 26.5 (2003): 1057-1070.

[5] Scharf, Klaus-Dieter, et al. “Molecular chaperones, stress response, and development in plants.” Plant Journal 18.5 (1999): 562-572.

[6] Baniwal, Santosh K., et al. “Heat stress transcription factors modulate heat-induced H3K4me3 changes in Arabidopsis thaliana.” Gene 536.1 (2014): 98-103.

[7] Lin, J.H., Chen, Y.F., Jeng, S.T. et al. “Characterization of heat stress-responsive genes in hot pepper (Capsicum annuum L.) using RNA-Seq analysis.” Plant Mol Biol 88, 31–41 (2015).

[8] Shanmugam, V., and Chakravarthy, V. “Evaluation of genetic diversity in black gram (Vigna mungo L. Hepper) using random amplified polymorphic DNA (RAPD) markers.” Legume Research-An International Journal 32.4 (2009): 262-266.

[9] Joshi, Rajesh Kumar, et al. “Genetic diversity and core collection evaluation in common bean (Phaseolus vulgaris L.) genotypes using inter simple sequence repeat (ISSR) markers.” African Journal of Biotechnology 11.85 (2012): 15131-15144.

[10] Singh, A. K., et al. “Genome-wide identification and expression analysis of HSF gene family in chickpea (Cicer arietinum L.).” PLoS One 14.5 (2019): e0217494.

[11] Li, Zhonghui, et al. “Genome-wide identification and analysis of heat shock transcription factors in Arachis hypogaea.” Journal of Plant Biochemistry and Biotechnology 30.2 (2021): 329-342.

[12] Kaushik, Prashant, et al. “Genome-wide analysis of heat shock factors and heat shock proteins in Vigna mungo.” Plant Physiology and Biochemistry 118 (2017): 356-365.

[13] Finn, Robert D., et al. “Pfam: the protein families database.” Nucleic Acids Research 42.D1 (2014): D222-D230.

[14] Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. “MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.” Molecular Biology and Evolution 33.7 (2016): 1870-1874.

[15] Chen, C., et al. “TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data.” Molecular Plant 13.8 (2020): 1194-1202.

[16] Letunic, Ivica, and Peer Bork. “Interactive Tree Of Life (iTOL) v4: recent updates and new developments.” Nucleic Acids Research 47.W1 (2019): W256-W259.

[17] Bailey, Timothy L., et al. “MEME Suite: tools for motif discovery and searching.” Nucleic Acids Research 37.suppl_2 (2009): W202-W208.

[18] Hu, Bo, et al. “GSDS 2.0: an upgraded gene feature visualization server.” Bioinformatics 31.8 (2015): 1296-1297.

[19] Lescot, M., et al. “PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences.” Nucleic Acids Research 30.1 (2002): 325-327.

[20] L. Chen, X. Song, S. Chen, W. Xu, and W. Li, "Genome-wide investigation of the heat shock transcription factor (Hsf) gene family in Tartary buckwheat (Fagopyrum tataricum)," BMC Genomics, vol. 23, no. 1, p. 259, 2022.

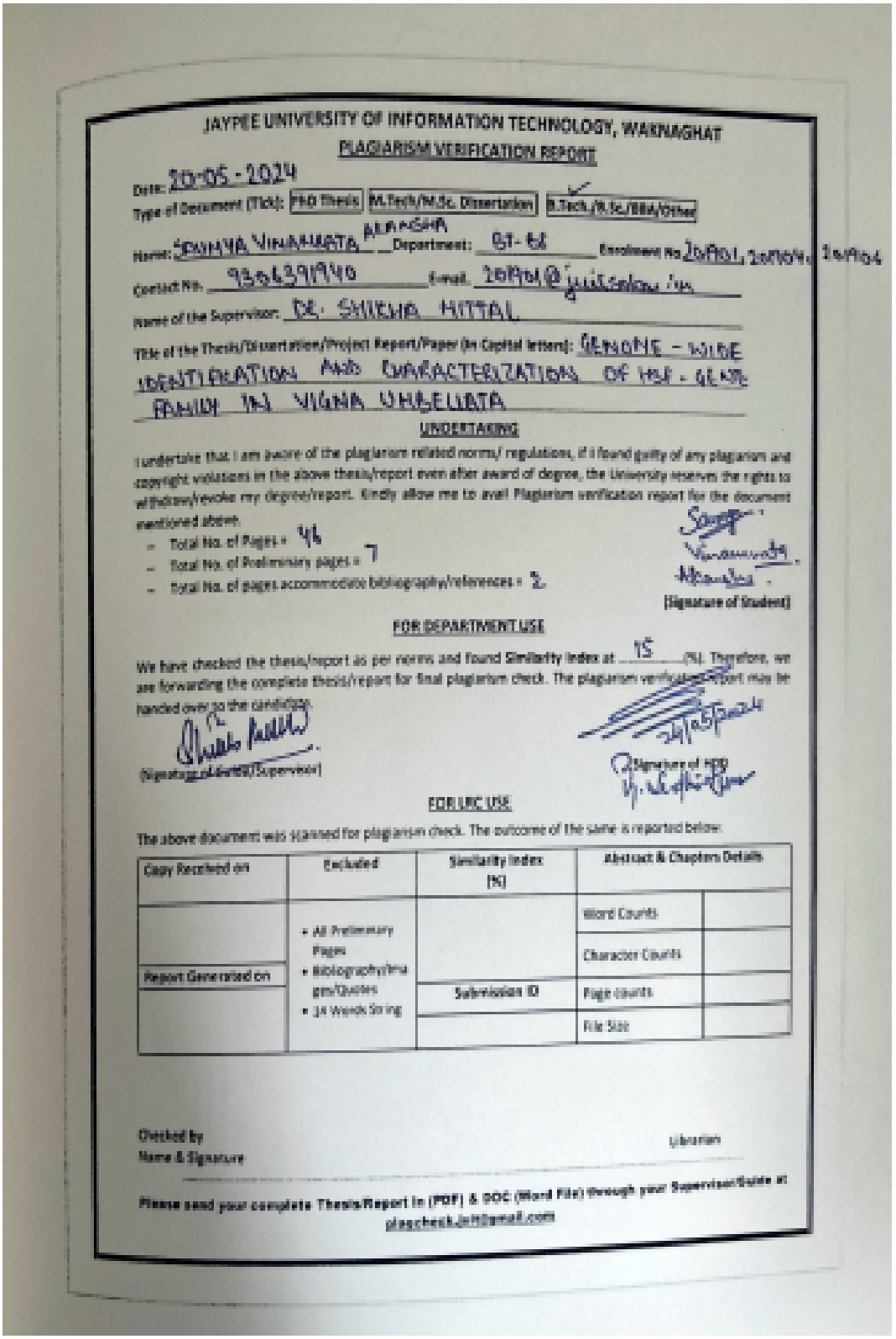
THesis_Btech

ORIGINALITY REPORT

15%	13%	10%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ir.juit.ac.in:8080 Internet Source	2%
2	bmcgenomics.biomedcentral.com Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	www.frontiersin.org Internet Source	1%
5	www.coursehero.com Internet Source	1%
6	ijaeb.org Internet Source	1%
7	mts.intechopen.com Internet Source	1%
8	www.wjgnet.com Internet Source	1%
9	www.nature.com Internet Source	<1%



JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 20-05-2024

Type of Document (Tick): PhD Thesis | M.Tech/M.Sc. Dissertation | B.Tech./B.Sc/BBA/Other ✓

Name: SRUSHYA VINAMRATA Department: BT-BE Enrolment No 201161, 201164, 201164

Contact No. 9366399940 E-mail. 201161@juitsolan.in

Name of the Supervisor: DR. SHIKHA MITTAL

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): GENOME - WIDE IDENTIFICATION AND CHARACTERIZATION OF HSF - GENE FAMILY IN VIGNA UNGUICULATA

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

- Total No. of Pages = 46
- Total No. of Preliminary pages = 7
- Total No. of pages accommodate bibliography/references = 2

[Signature of Student]

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found Similarity Index at 15 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

[Signature of Guide/Supervisor]

[Signature of HOD]

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Abstract & Chapters Details	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Page Counts	
			File Size	

Checked by
Name & Signature

Librarian

Please send your complete Thesis/Report in (PDF) & DOC (Word File) through your Supervisor/Guide at plagcheck.juit@gmail.com