CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Genome-Wide Identification and Functional Characterization of *NLP* Gene Family in *Glycine max*

by

Qurat ul Ain

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Health and Life Sciences
Department of Bioinformatics and Biosciences

2026

CERTIFICATE OF APPROVAL

Genome-Wide Identification and Functional Characterization of *NLP* Gene Family in *Glycine max*

by

Qurat ul Ain

(MBS241010)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Haris Khurshid	NARC, Islamabad
(b)	Internal Examiner	Dr. M. Asad Anwar	CUST, Islamabad

Dr. Sohail Ahmad Jan
Thesis Supervisor
August, 2026

Dr. Syeda Marriam Bakhtiar
Head
Dept. of Bioinfo. & Biosciences
August, 2026

Dr. Sahar Fazal
Dean
Faculty of Health & Life Sciences
August, 2026

Author's Declaration

I, **Qurat ul Ain** hereby state that my MS thesis titled "**Genome-Wide Identification and Functional Characterization of *NLP* Gene Family in *Glycine max***" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Qurat ul Ain)
Registration No: MBS241010

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Genome-Wide Identification and Functional Characterization of *NLP* Gene Family in *Glycine max***" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Qurat ul Ain)

Registration No: MBS241010

Acknowledgement

All the praises and thanks be to Allah, Who has guided me to this, and never could I have found guidance, had it not but Allah had guided me. In times of darkness, You, Allah, are my guiding light. Thank you for illuminating my path.

I would like to express my deepest gratitude to the following individuals, without whom this thesis would not have been possible: My supervisor, Dr. Sohail Ahmad Jan, for their unwavering guidance, invaluable insights, and continuous support throughout the entire research process. Their expertise and encouragement have been instrumental in shaping the direction of this study. I am grateful for their valuable feedback and constructive criticism, which greatly contributed to the improvement of this thesis. A special thanks to Dr. Sami Ullah and Dr. Muhammad Asad Anwar, whose moral support made this research project possible. Their help not only facilitated the execution of the study but also allowed me to attend conferences and engage with the academic community. I am thankful for their assistance and resources, which played a crucial role in the successful completion of this research.

I would also like to thank my mother, Kosaur Parveen, brother, Dr. Ghulam Murtaza, and sister Ghulam Fatima, for their unwavering support and understanding during the challenging moments of this academic journey. Their encouragement and belief in me were my constant motivators. Lastly, to Miss Maliha Fatima, who generously contributed her time in formatting my work. I am deeply thankful to everyone.

(Qurat ul Ain)

Abstract

Transcription factors present in the NODULE-INCEPTION-LIKE proteins (*NLPs*) family are regarded as vital regulators of plant response towards nitrogen, its uptake, and molecular structure. Although there have been abundant descriptions available regarding the structures of *NLPs* across several plant genera, there is a lack of knowledge regarding the same in soybean plants. In this study, an exhaustive computational approach has been adopted for the identification of *GmNLPs* in the economically important plant *Glycine max*. Twelve unique types of *GmNLPs* were observed within this plant, showing great similarities to the structure of *Arabidopsis thaliana* in terms of physical and chemical properties. Two conserved regions, PB1 and RWP-RK, were identified in all genesAn interesting observation was that the average length of *GmNLPs* sequences was significantly larger than that of *Arabidopsis thaliana*. However, it was revealed from ProtParam analysis that the overall structure of *GmNLPs* was hydrophilic in nature due to their negative Grand Average of Hydropathicity values. On the other hand, structural modeling enabled the establishment of the positioning of *GmNLPs* in terms of their dynamism within the cell nucleus. Protein-protein interaction analysis indicated that the interacting proteins of the majority of *GmNLPs* were associated with the transportation of biomolecules. Notably, the protein-protein interaction study suggested a strong coordination between *GmNLPs* and nitrogen-responsive genes such as nitrate reductase, emphasizing the significant regulation of NLPs by nitrogen supply. Phylogenetic analysis indicated a definite presence of an evolutionarily stratified group, and it was established that *GmNLPs* share structures with nitrogen regulatory elements of *Oryza sativa, Brassica napus, Zea mays,* and *Arabidopsis thaliana*. This is essential in establishing ancestry and is significant for structural genetic materials in increasing metabolic efficiency through nitrogen use efficiency.

Keywords: NODULE-INCEPTION-LIKE proteins, transcription factors, conserved domains, protein-protein interactions, computational approach..

Contents

List of Figures

List of Tables

Abbreviations

CDPKs	calcium-dependent protein kinases
GRAVY	Grand Average of Hydropathicity
GSDS	Gene Structure Display Server
ITOL	Interactive Tree of Life
NLP	Nodule Inception Like Proteins
NUpE	Nitrogen Uptake Efficiency
NUtE	Nitrogen Utilization Efficiency
PB1	Phox and Bem1

Chapter 1

Introduction

Gene families consist of related genes sharing a common ancestor and similar biological functions, evolving through gene duplication to help plants adapt to environmental changes. The Nodule Inception-Like Protein family is vital for nitrogen signaling and plant development, containing transcription factors with essential domains for DNA and protein binding. *NLP* proteins, prevalent in plants, regulate nitrogen-responsive genes and are particularly critical in legumes for nodulation and nitrogen fixation, enhancing nitrogen use efficiency, an essential macronutrient for plant growth. While nitrogen is abundant in the atmosphere, its scarcity in soil necessitates effective assimilation by plants to support biological functions. Plants cannot utilize atmospheric nitrogen directly and depend on forms such as nitrate and ammonium [1].

Nitrogen is a crucial macronutrient for plant growth, playing key roles in the formation of amino acids, proteins, nucleic acids, and chlorophyll. While abundant in the atmosphere as N_2, plants utilize bioavailable forms like nitrate (NO_3^-) and ammonium (NH_4^+) from the soil or fertilizers. Effective nitrogen assimilation is vital for crop productivity and sustainable agriculture. *NLPs* in plants act as transcription factors that regulate the expression of nitrogen-responsive genes, facilitating processes such as nitrate uptake. These proteins control the genes for nitrate transporters and reductases, integrating nitrogen into essential

biomolecules. In legumes, *NLPs* support symbiotic nitrogen fixation, enhancing nitrogen use efficiency and improving soil fertility [2].

Nitrogen uptake in plants involves intricate regulatory processes that link external nitrogen sources with internal metabolic functions. Plants possess signaling networks to sense nitrogen forms like nitrate and ammonium, optimizing absorption and growth. Activated *NLPs* influence genes essential for the nitrate assimilation pathway, enabling nitrogen incorporation into vital biomolecules. Additionally, *NLPs* interact with various signaling pathways, coordinating nutrient management, root architecture, and stress responses. This regulation enhances nitrogen efficiency, reducing the need for fertilizers and environmental risks, thereby supporting sustainable agriculture. Understanding *NLP* dynamics can lead to improved crop varieties capable of higher nitrogen use efficiency, crucial for future food demands [3].

Nitrogen signaling in plants is interlinked with hormonal and stress response networks, primarily through *NLP* transcription factors. These factors help coordinate nitrogen utilization with plant growth and environmental adaptation. For instance, *NLP* interactions with auxin and cytokinin influence root architecture and nodule formation, with cytokinin promoting cortical cell division for nodule organogenesis and auxin affecting root branching [4]. During abiotic stresses like drought, *NLP*-mediated networks adjust to prioritize nitrogen for survival over growth. ABA signaling plays a role in this integration, linking nutrient sensing with stress responses. This regulatory framework maintains nitrogen homeostasis, ensuring that nitrogen-responsive genes are expressed according to the plant's developmental needs, nutrient levels, and environmental challenges. Understanding is extremely important for the development of better nitrogen-utilizing crops that can withstand climate change [5].

Ineffective nitrogen use leads to over-fertilization, resulting in environmental pollution and economic setbacks, therefore, improving the efficiency of nitrogen utilization is essential for sustainable agriculture. *NLPs* play a vital role and are characterized by conserved domains, essential for protein interactions and bindigs. *NLP* genes are important for sensing nitrate and regulating nitrogen-responsive

gene metabolism. They significantly contribute to nodulation and nitrogen fixation in legumes, making them critical for sustainable farming. While *NLP* genes, particularly *NLP6* and *NLP7*, are recognized for their role in nitrate perception, their detailed functional characterization in some legumes, particularly soybean, is inadequate. The exploration of these transcription factors is essential for improving nitrogen assimilation and crop yield [6].

Glycine max, also known as soybean, is a crucial leguminous crop renowned for its high protein i.e., 35-40% and 18- 22% oil content, vital for human and animal nutrition, as well as industrial applications like biofuels and cooking oils. Its unique ability to fix atmospheric nitrogen through a symbiotic relationship with rhizobia enhances soil fertility, minimizes reliance on synthetic fertilizers, and supports sustainable agriculture [7]. Economically, soybeans serve as a primary livestock feed and contribute to diverse food products, especially in vegetarian diets. Their adaptability to various climatic and soil conditions and improvements through biotechnology further encourage their cultivation. However, challenges such as pests, soil degradation, and nitrogen efficiency must be addressed to secure their role in food security and sustainable development [8].

Glycine max can convert atmospheric nitrogen into a usable form through symbiotic nitrogen fixation, facilitated by a mutualistic relationship with rhizobia in the soil. This process is crucial for reducing reliance on synthetic fertilizers, enhancing soil fertility, and ensuring sustainable crop production. The interaction begins when legume roots secrete flavonoids, attracting rhizobia that produce nod factors to initiate nodule formation. Within these nodules, rhizobia differentiate into nitrogen-fixing bacteroids, utilizing the nitrogenase enzyme complex to convert nitrogen into ammonia, supported by the plant ATP. The resulting ammonia is transformed into amino acids, critical for plant growth. Nodule formation efficiency directly impacts crop yield and can vary due to genetic, environmental factors, or rhizobial strain specificity [9]. Understanding nodulation mechanisms is vital for optimizing soybean breeding for nitrogen efficiency. However, there exists a gap in genome-wide studies on *NLP* genes in soybean, which highlights

the urgent need for further research to inform breeding strategies and enhance crop production.

The proposed study aims to conduct a genome-wide analysis of *NLP* genes in soybean, focusing on their structural features, evolutionary relationships, and role in nitrogen signaling. Advances in bioinformatics and computational biology facilitate large-scale genomic analysis with high precision. *In silico* methods enable the discovery and functional prediction of genes, employing techniques like sequence alignment and phylogeny to elucidate gene families' biological roles. By utilizing bioinformatics tools, we can identify gene members, analyze conserved motifs, and predict functions within nitrogen signaling pathways, thus enhancing the understanding of plant genomics and improving crop traits.

1.1 Problem Statement

The *NLP* gene family and its member genes have been well studied in several plants, but insufficient information is known about the *NLP* gene family in soybeans.

1.2 Aim and Objectives

The goal of this study is to identify and thoroughly characterize the *NLP* gene family in soybeans based on advanced bioinformatics methods. The objectives of the study include:

i. To identify *NLP* gene family members in the soybean using bioinformatics tools.

ii. To analyze physicochemical properties, gene structure, conserved domains, and protein interactions of *NLP* proteins.

iii. To access the phylogenetic relationship of the soybean *NLP* gene family with other economically important crop species.

1.3 Scope of Study

The *NLP* gene family in soybeans is the subject of a genome-wide investigation in this work utilizing bioinformatic tools for gene structure, conserved domains, interactions, phylogenetics, and function prediction. The findings enhance understanding of nitrogen signaling pathways, aiming to improve nitrogen use efficiency (NUE) in agriculture. By providing insights into transcriptional regulators responsible for nitrogen metabolism, the research supports crop improvement strategies that reduce reliance on synthetic fertilizers, thus mitigating environmental impacts and contributing to sustainable agricultural practices. The implications extend to food security, supporting the productivity of leguminous crops and aligning with goals for environmental conservation in response to population growth and climate change.

Chapter 2

Literature Review

2.1 Soybean

Soybean (*Glycine max*), is one of the most economically vital oilseed legumes grown globally. It serves as a major source of plant-based protein (40–42%) and high-quality cholesterol-free oil (18–22%) rich in unsaturated fatty acids, alongside essential minerals and complex carbohydrates, as shown in figure 2.1. It forms symbiotic relationships with nitrogen-fixing bacteria, enhancing soil fertility and reducing chemical fertilizer dependency. Economically, it supports livestock feed, biofuels, and food products. Its adaptability to various climates promotes widespread cultivation. The genetic and molecular mechanisms regulating nitrogen in soybean are vital for improving crop productivity and sustainability [10].

Beyond its culinary and nutritional roles as a functional food, soybean possesses therapeutic potential, exhibiting positive impacts on conditions such as hyperglycemia, hyperlipidemia, and hypertension. The plant's metabolic profiling highlights rich concentrations of bioactive components like isoflavones, polyphenols, and low-molecular-weight peptides that exhibit diverse antioxidant and immunomodulatory traits. Conducting a comprehensive genome-wide identification of structural gene families within soybean offers critical insight into improving its

Nitrogen Use Efficiency, consolidating its position as an indispensable agricultural and industrial staple [11].

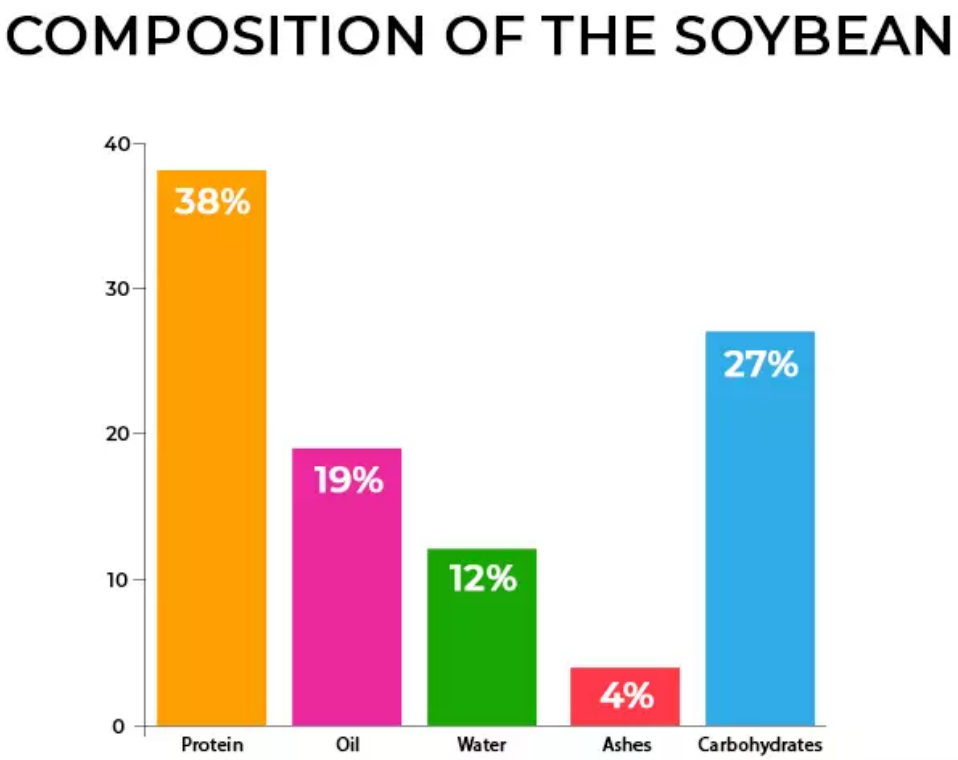


FIGURE 2.1: Composition of soybean. Soybean is composed of 38% protein, 19% oil, 12% water, 27% carbohydrates and minute quantity of ashes [11]

2.2 Morphology and Taxonomy of Soybean

Soybean is a dicotyledonous leguminous plant characterized by a well-developed morphological structure that supports its growth, reproduction, and nitrogen-fixing ability. The plant consists of a strong taproot system with lateral roots that facilitate nutrient and water absorption. Specialized root nodules are formed through symbiotic interaction with rhizobia bacteria, enabling biological nitrogen fixation. The stem is erect and branched, providing structural support and transport of nutrients. Leaves are trifoliolate, broad, and green, playing a major role in photosynthesis. Soybeans produce small flowers that develop into pods, which typically contain two to four seeds, as shown in figure 2.2. These seeds are rich in protein and oil, making Soybean economically valuable [12].

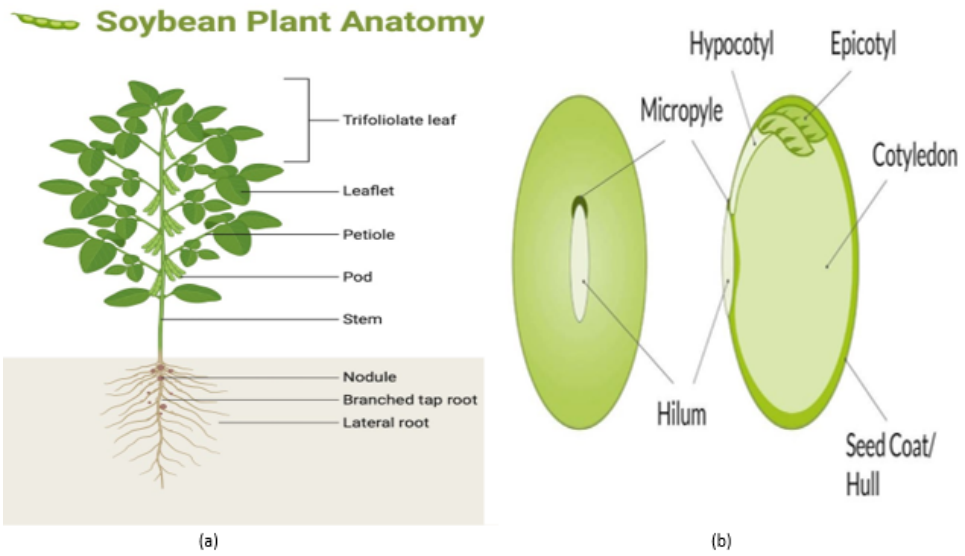


FIGURE 2.2: Anatomy and seed structure of soybean. The soybean is an erect, branching annual legume of the pea family that can reach over 2 meters in height. Its seed consists of cotyledon, micropyle, and hilum, etc [12].

The structural organization of soybean is closely linked to its physiological efficiency, particularly its ability to adapt to different environmental conditions and optimize nitrogen utilization. The taxonomic classification of soybean is given in table 2.1.

TABLE 2.1: Taxonomy of *Glycine max* [13]

Kingdom	**Plantae**
Division	Magnoliophyta
Class	Magnoliopsida
Order	Fabales
Family	Fabaceae
Genus	*Glycine*
Species	*Glycine max*

2.3 Status of Soybean in Pakistan

In Pakistan, soybean is categorized as a non-conventional oilseed crop capable of successful cultivation during both the spring and summer seasons. Production

is primarily situated within specific sub-regions of Punjab, Sindh, and Khyber Pakhtunkhwa, where ecological niches provide favorable conditions. Although the country features highly compatible agro-climatic regions, commercial cultivation has historically faced significant stability issues, leading to an overall reliance on imports [14].

The seed yield potential of indigenous lines varies, yet specialized breeding trials have shown that institutes like NIBGE and NARC can achieve highly competitive grain weights and biological yields under optimized conditions.

The marketing structure of soybeans in Pakistan is dynamically related to the requirements of the local poultry and feed industries, which remain the primary consumers of domestic soy meal. Government initiatives, market information networks, and supportive agricultural policies play a critical role in stabilizing price mechanisms for local growers.

On a global scale, major primary producers such as the United States, Brazil, and Argentina dominate production, shaping market standards and dictating price indexes that influence local transactional systems across Pakistani procurement hubs, as shown in figure 2.3.

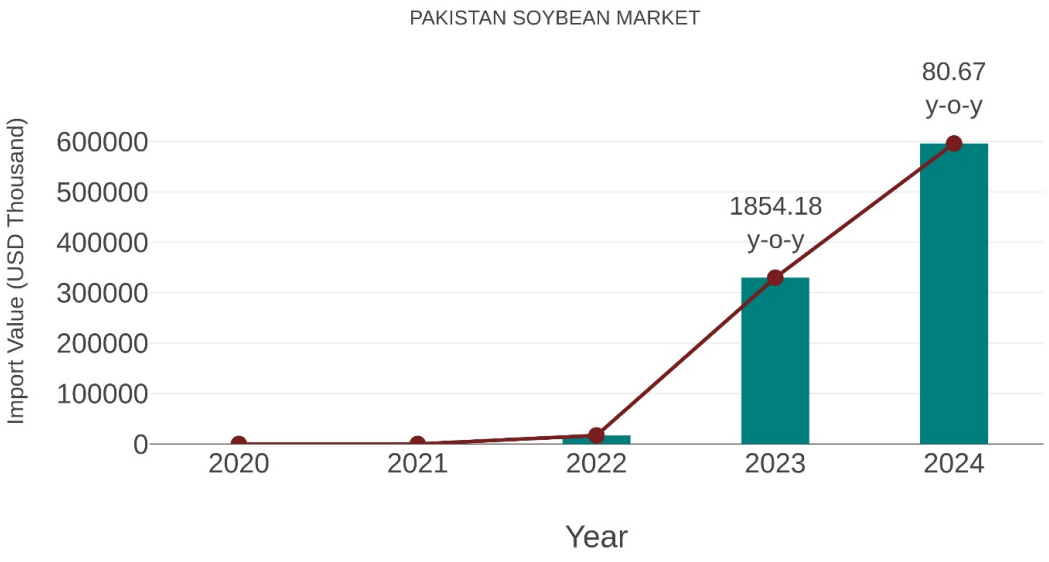


FIGURE 2.3: Pakistan soybean market. Pakistan soybean market has gradually increased from a year-over-year explosion of 1854.18% in 2023. This growth momentum continued into 2024, with an additional 80.67% year-over-year to approach an import value of nearly 600,000 USD [15].

2.4 Worldwide Import and Export of Soybean

Soybean imports constitute a monumental fraction of the international agricultural commodity market, driven primarily by heavy industrial demand for livestock feed and cooking oil processing. On a global scale, China leads international imports by a substantial margin, followed closely by major trading blocks within the European Union, Southeast Asia, and South Asian nations, including India and Pakistan. However, these annual import volumes are subject to significant market variables, fluctuating continuously based on regional crop performance, international tariff adjustments, and localized domestic agricultural storage capabilities [16].

This volatility poses noticeable economic pressures, especially for developing economies where large-scale imports of soy meal can exert considerable stress on foreign exchange reserves, strongly reinforcing the critical need for enhanced domestic cultivation strategies. The global export ecosystem for soybean is led by large-scale agricultural economies in North and South America, with Brazil and the United States standing out as the largest international suppliers, as shown in figure 2.4.

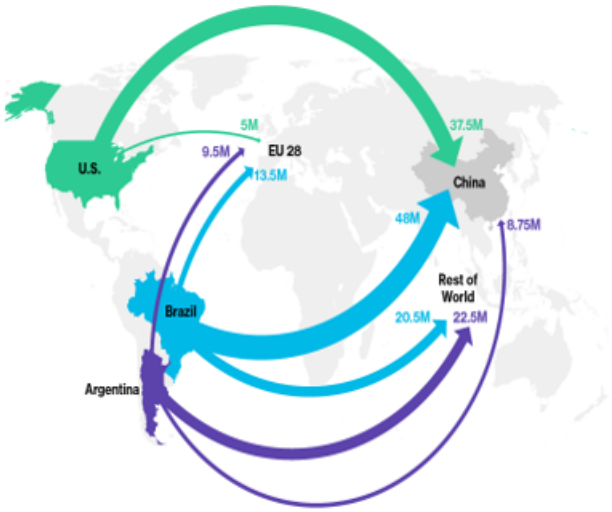


FIGURE 2.4: Global soybean trade. This image illustrates global trade routes of soybean, China emerges as the ultimate primary destination, pulling massive imports that include 48M from Brazil, 37.5M from the U.S., and 8.75M from Argentina [17].

These nations export massive quantities of whole dry beans, crushed oil, and dense animal feed meals to highly populated markets. The international trade balance

remains dynamic, influenced heavily by shifting seasonal yields, international maritime logistics, and the expanding global trend toward plant-based proteins and commercial livestock management [17].

2.5 Industrial Uses of Soybean

Beyond direct dietary applications, soybeans serve a multitude of heavy industrial functions, establishing themselves as a highly versatile raw material for sustainable manufacturing, as shown in figure 2.5. In the field of material science, soybean polymers and protein matrices are actively researched and developed as eco-friendly alternatives to traditional petroleum-based plastics, helping to reduce long-term environmental waste. Similarly, the textile and manufacturing sectors benefit from soybean protein isolates, which are industrially processed into specialized adhesives, binders, and synthetic fiber finishings that offer a lower chemical footprint than conventional synthetics. The energy sector also heavily utilizes the crop, as soybean oil represents a primary feedstock for international biodiesel production, offering a cleaner, renewable energy alternative that lowers carbon emissions compared to fossil fuels. Furthermore, soybeans play a vital ecological role directly in the field; within sustainable crop management systems, it serves as an excellent green manure and cover crop that significantly enriches organic soil matrices and improves soil structure upon being plowed back into the ground [18].

2.6 Nutrition Profile and Health - Promoting Benefits of Soybean

Soybeans exhibit profound physiological and medicinal advantages when integrated systematically into human nutrition, serving as a cornerstone for functional food research. It is uniquely packed with comprehensive amino acid profiles that fulfill human dietary requirements, alongside an array of bioactive phytochemicals and therapeutic proteins that work synergistically to support long-term metabolic

FIGURE 2.5: Industrial products of soybean. This picture features items made from soybeans, like goodyear tires, paint, plastics, household cleaners, personal care products, furniture cushions, and automotive fuel additives [19].

homeostasis. Beyond baseline nutrition, the high content of polyunsaturated fatty acids and soluble fibers within soybean actively promotes healthy vascular and gastrointestinal functions. Regular dietary intake is clinically correlated with reduced levels of low-density lipoprotein (LDL) cholesterol, lowered blood pressure profiles, and enhanced gastrointestinal mobility, which effectively soothes the stomach lining and mitigates chronic digestive irritation [20].

Featuring a low glycemic index and a high concentration of structural dietary fiber, soybean acts as an exceptional component in calorie-conscious regimens and therapeutic diets. Its ingestion triggers prolonged satiety signaling and stabilizes baseline blood glucose levels, aiding in long-term weight management by reducing overall caloric cravings and improving cellular insulin sensitivity.

Simultaneously, soybeans are rich in natural antioxidants, including tocopherols,

vitamin C components, and specialized phenolic compounds, that fortify the body's internal defense mechanisms as shown in figure 2.6. These bio-molecules neutralize systemic free radicals throughout the body, mitigating cellular damage induced by oxidative stress and effectively slowing down chronic cellular degeneration pathways [21].

The active biological compounds within soybean, particularly specialized low-molecular-weight peptides and phytoestrogenic isoflavones, demonstrate distinct and potent anti-inflammatory mechanisms. These molecular agents assist in down-regulating pro-inflammatory cytokines, making soybean derivatives potentially advantageous in managing chronic inflammatory conditions such as rheumatoid arthritis and vascular inflammation.

Furthermore, soybeans are uniquely rich in specific isoflavones like genistein and daidzein. Advanced clinical and molecular assessments indicate that these secondary metabolites play protective roles against hormone-dependent malignancies, support long-term bone density preservation, and assist in balancing general endocrine signaling pathways, thereby offering a comprehensive dietary shield against various chronic diseases [22].

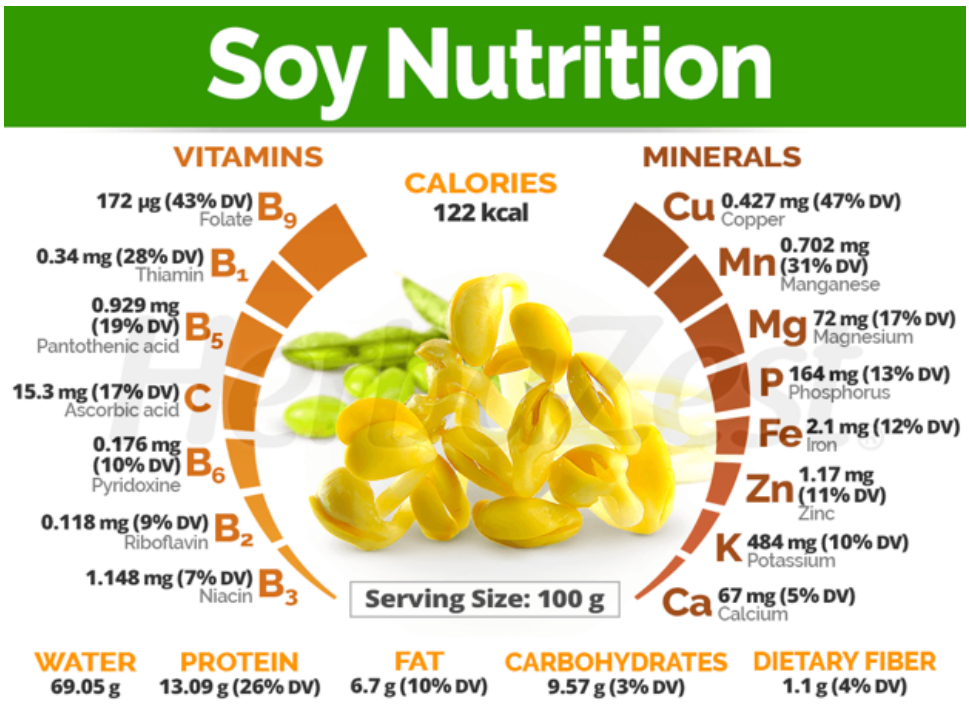


FIGURE 2.6: Soybean nutritional content. Soybeans consist of a variety of minerals and vitamins along with water, protein, fat, carbohydrates, and dietary fiber [20].

2.7 Soybean Cultivation Requirements

Soybeans thrive under warm, tropical to warm-temperate environmental parameters, requiring specific agronomic conditions to achieve their full genetic yield potential. The crop demands well-drained, fertile loamy soils with an optimal, near-neutral pH range to facilitate unimpeded root expansion and efficient nutrient uptake. Additionally, soybean cultivation relies heavily on precise photoperiod management because the plant is short-day by nature, specific maturity groups exhibit extreme sensitivity to day-length variations, which directly trigger the transition from vegetative growth to reproductive development, as shown in figure 2.7.

Successful seasonal management also demands balanced, well-timed inputs of moisture, particularly during the critical vegetative, flowering, and early reproductive phases. Maintaining adequate soil moisture during these windows ensures optimal floral induction, robust pod development, and maximum seed filling, which ultimately dictates the final harvest index. Abiotic stressors, most notably extreme seasonal temperatures, prolonged drought periods, and high soil salinity, pose severe risks to global soybean yields and overall seed quality. Among these environmental variables, water availability remains incredibly critical to structural development [23].

Severe water deficiencies occurring during the highly sensitive pre-flowering and pod-filling growth phases disrupt internal turgor pressure and vascular transport, causing substantial reductions in total seed weight, oil accumulation, and overall biological output. Cellular nitrogen management interacts continuously with these environmental constraints, acting as a buffer against seasonal volatility. Proper and efficient nitrogen assimilation bolsters the crop's structural and osmotic resilience, enabling it to better withstand temporary physiological stress. Conversely, when severe nutritional or nitrogen deficiencies occur, the plant's metabolic machinery stalls. This leads to the onset of classic visual symptoms, including standard leaf chlorosis due to chlorophyll degradation, stunted lateral root and shoot architecture, and significantly compromised seed vitality [24].

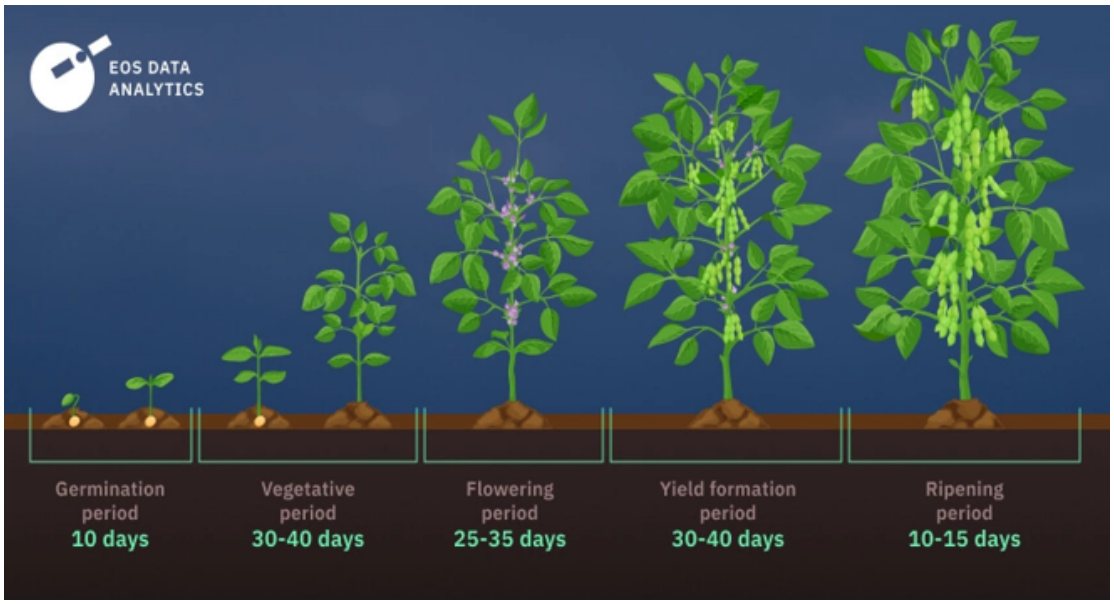


FIGURE 2.7: Growth stages of soybean plant. This image depicts the soybean plant's journey from a 10-day germination period through vegetative and flowering stages to reach the reproductive stage [23].

2.8 Nitrogen's Function in Legume Plants

Nitrogen get absorbed by plant from the soil matrix in both organic amino acids, urea and inorganic forms, with nitrate and ammonium representing the predominant sources. Legumes are particularly distinguished by their capacity to engage in symbiotic biological nitrogen fixation via rhizobia interactions within specialized root nodules. However, soil nitrogen concentrations remain highly variable due to leaching, microbial dynamics, and agricultural management practices. To counteract environmental flux, plants deploy an advanced regulatory network that senses external nitrogen availability, adjusting root system architecture (RSA) and uptake velocities to maintain metabolic equilibrium [25].

As a core macronutrient, nitrogen forms the foundational basis of leguminous plant biology, serving as a core constituent of amino acids, metabolic enzymes, and structural proteins. It acts as the indispensable element for building structural and functional proteins, which constitute the primary machinery of plant cells, including vital transport channels embedded within the plasma membrane that orchestrate the internal movement of micro-molecules. Beyond these protein

networks, the structural backbone of deoxyribonucleic acid and ribonucleic acid depends directly on nitrogenous bases. Adequate nitrogen availability ensures flawless genetic replication, regulated cell division, and healthy tissue differentiation during development, making it a limiting factor for overall genomic and cellular integrity throughout the life cycle of the legume [26].

Furthermore, nitrogen serves as a central constituent of the chlorophyll molecule, the light-absorbing pigment responsible for driving photosynthesis. An effective nitrogen supply elevates light energy capture, optimizing the production of organic carbohydrates and driving the energetic processes of the plant as shown in figure 2.8. Crucial plant enzymes involved in cellular respiration, biological nitrogen fixation, and hormonal synthesis also require specific nitrogenous configurations to maintain their structural integrity and catalytic activity. When nitrogen levels are optimal, this macronutrient directly shapes overall plant architecture by dictating leaf expansion rates, stem elongation dynamics, and healthy root branching profiles, ensuring that both underground and aboveground organs are fully developed to maximize resource acquisition [27].

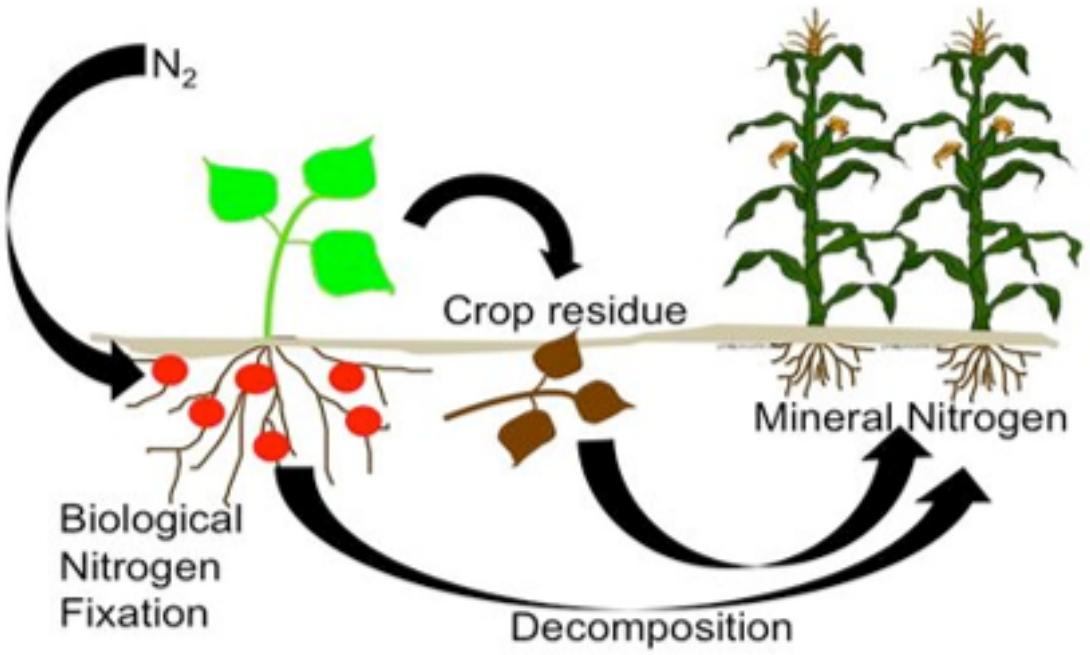


FIGURE 2.8: Legume cropping system. This figure illustrates the process of biological nitrogen fixation, where a legume plant converts atmospheric nitrogen into soil-bound nutrients via root nodules [27].

In addition to driving structural development, nitrogen plays a vital role in the internal regulation of complex metabolic pathways, including the synthesis of secondary metabolites that are essential for mounting defensive cellular responses against emerging pests and pathogens. This physiological buffering extends directly to the plant's survival mechanisms, a balanced nitrogen availability improves the legume's capacity to adjust osmotically under environmental challenges.

By optimizing solute concentrations and maintaining cellular turgor, adequate nitrogen levels facilitate survival during sudden abiotic shocks like severe drought, soil salinity, or localized chemical stress, making it an indispensable element for both baseline productivity and stress resilience [28].

2.9 The Transporters Function in Nitrogen Absorption

The cellular transport of nitrogenous molecules across the plasma membrane of root cells is executed by highly specialized high-affinity and low-affinity transport systems, which frequently serve a dual purpose as both vehicles for influx and environmental sensors, commonly referred to as transceptors. These integrated molecular components allow the root system to continuously monitor external nutrient concentrations and transmit real-time signals directly to the plant's genomic machinery as shown in table 2.2.

To optimize this process, transporters work in close physiological coordination with complex, interconnected membrane-bound networks, including specific receptors and proton-pumping ATPases. By actively extruding protons, these ATPases generate the strong electrochemical gradients and localized proton motive forces required to drive the active transport and inward flux of essential nutrients against their concentration gradients [29].

The internalization of nitrate is mediated by distinct, multi-gene families embedded within the plasma membrane. Under conditions of low external nitrate

availability, high-affinity transport systems, primarily driven by NRT2 expression working alongside accessory proteins, are transcriptionally up-regulated to scavenge trace amounts of the nutrient from the soil.

Conversely, when nitrate levels in the rhizosphere are abundant, the plant dynamically shifts its reliance toward low-affinity transport pathways, primarily managed by NRT1/NPF members, which allow for rapid, high-volume influx without saturating the cellular machinery [30].

Parallel to this system, ammonium ions are directly absorbed from the soil matrix via localized ammonium transporter pathways, which are organized into distinct high-affinity and low-affinity sub-families. Because free intracellular ammonium can rapidly disrupt cellular pH gradients and induce severe mitochondrial and chloroplast toxicity, its uptake rate must be strictly regulated. Consequently, the influx velocity of ammonium through these channels is tightly linked to downstream metabolic pools and carbon skeleton availability, ensuring that internalized ammonium is immediately assimilated into non-toxic organic forms, such as glutamine and glutamate [31].

TABLE 2.2: Role of transporters in nitrogen uptake [30]

Transporter's role	**Working mechanism**
Grab nitrogen from the soil	Transporters bind to nitrate and ammonium in the soil.
Pull nitrogen into roots	They actively transport nitrogen across the root cell membrane into the plant.
Work at different levels	High-affinity transporters work when nitrogen is scarce; low-affinity transporters work when nitrogen is abundant.
Move nitrogen upward	After uptake, transporters move nitrogen from the roots to the shoots/stems throughout the plant.
Sense nitrogen levels	Some transporters also act as sensors, telling the plant how much nitrogen it has.

2.10 Effect of Nitrogen Use Efficiency

This is an complex physiological trait which is fundamentally divided into two major components: Nitrogen Uptake Efficiency (NUpE), which measures the capacity of the root system to harvest nitrogen from the soil, and Nitrogen Utilization Efficiency (NUtE), which reflects the plant's internal capacity to redirect and remobilize that absorbed nitrogen into vegetative biomass and seed yield. Enhancing soybean NUE is vital for mitigating the widespread environmental damages historically associated with excessive synthetic fertilizer applications, such as the leaching of nitrates into groundwater tables, harmful surface runoff into aquatic ecosystems, and the volatile release of potent greenhouse gases like nitrous oxide, as shown in figure 2.8. Because global agricultural systems rely heavily on nitrogen inputs, even minor incremental increases in leguminous NUE translate directly into reduced production costs for farmers by lowering seasonal overheads, while simultaneously minimizing the ecological footprint of large-scale oilseed cultivation [32].

To boost the nitrogen efficiency profiles of soybean crops, modern agricultural systems employ an array of sustainable agronomic, technical, and biological interventions. In the field, precision agriculture plays a vital role by utilizing data-driven technologies to apply highly targeted fertilizer inputs that directly match localized soil variations and real-time plant requirements, thereby reducing waste. This approach is heavily augmented by the deployment of controlled-release fertilizers, which chemically or physically synchronize nutrient release rates with the specific, peak-demand growth stages of the crop. Furthermore, incorporating sustainable crop rotation and intercropping systems featuring deep-rooting species helps optimize structural soil properties, breaking up compacted layers and capturing deep-set nitrogen before it leaches out of the rhizosphere [33].

On a biological level, optimization of microbial inoculation serves as a powerful lever for enhancing nutrient dynamics; by inoculating seeds with elite, hyper-efficient rhizobial strains, growers can maximize biological nitrogen fixation and

reduce dependency on synthetic chemical inputs. Finally, advanced genetic approaches offer long-term solutions by modifying key inner regulatory components, such as transcription factor networks and signaling cascades. Engineering these genetic pathways can significantly enhance overall root architecture for better nutrient capture, while simultaneously optimizing internal amino acid recycling and remobilization pathways during seed development [34].

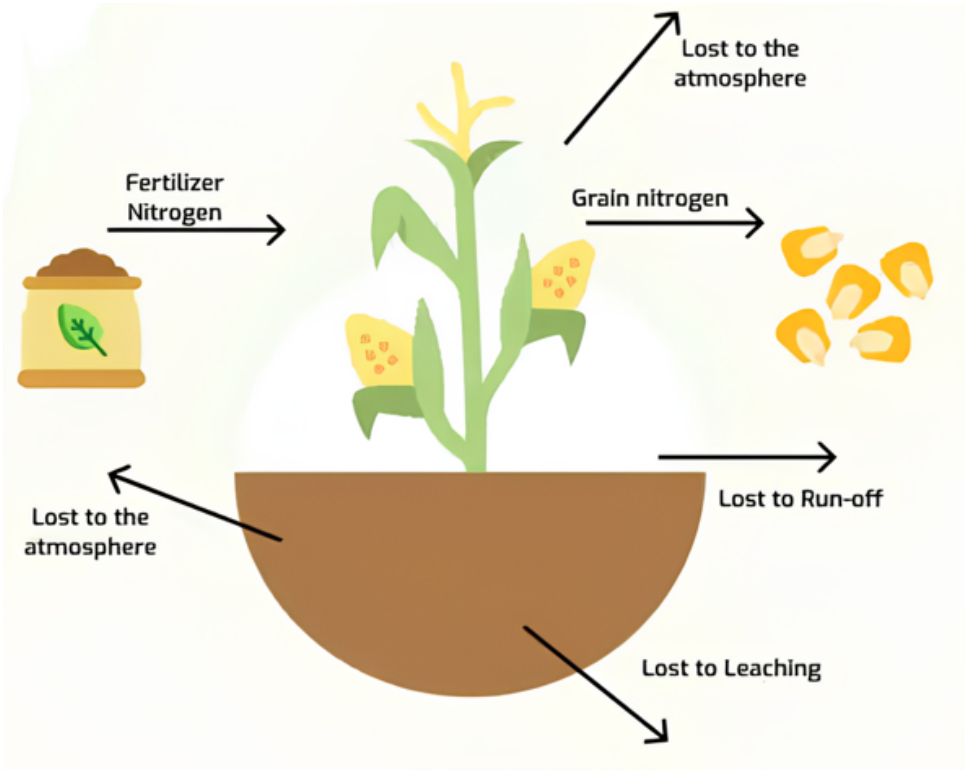


FIGURE 2.9: Nitrogen use efficiency of soybean. This picture examines how fertilizer nitrogen enters a crop system and its conversion into harvested grain nitrogen. It tells of significant efficiency losses during this process, notably nitrogen losses to the atmosphere, surface runoff, and deep soil leaching [34].

2.11 Regulatory Mechanisms of Transcription Factors and *NLPs*

nuclear localization signals, and transactivation regions. They act as fundamental molecular switches that control the rate of transcription by binding specifically to cis-regulatory elements located within the upstream promoter regions of target

genes, as shown in figure 2.9. These proteins coordinate the downstream expression of complex gene regulatory networks governing cellular differentiation, tissue specification, and multi-tiered stress adaptation. By establishing intricate protein-protein and protein-DNA interactions, Transcription factors allow the plant to fine-tune its physiological activities in response to shifting environmental cues, effectively coordinating tissue-specific growth and developmental transitions [35].

Among these regulatory networks, plant-specific transcription factors known as NIN-like proteins serve as crucial master regulators, bridging environmental nutritional signaling with direct genomic responses. The structural composition of *NLPs* distinguishes them within the broader landscape of plant transcription factors, as they possess a unique dual-domain architecture designed for both DNA binding and protein interaction. Domains facilitate high-affinity, sequence-specific binding to Nitrate Responsive Cis-elements (NREs) located within the promoters of primary nutrient transport and assimilation genes [36].

In addition to this DNA-binding domain, *NLPs* feature a characteristic Phox and Bem1 (PB1) domain at their C-terminus. This PB1 domain acts as a versatile protein-protein interaction platform, allowing *NLPs* to form homo or heterodimers with other regulatory proteins or to assemble multi-protein transcription complexes that modulate the activation or repression of downstream target genes. Functionally, *NLPs* act as the primary switches in the plant's quick response to nitrogen availability, transforming physical nutrient sensing into sweeping transcriptomic shifts. Under nitrogen-limiting conditions, *NLPs* are typically localized within the cytoplasm in an inactive state. Upon the sensing of external nitrate influx, intracellular signaling cascades, frequently involving calcium-dependent protein kinases (CDPKs), induce rapid post-translational phosphorylation of conserved residues within the NLP protein structure [37].

This structural modification triggers a conformational change that unmasks the nuclear localization signal, driving the swift translocation and retention of *NLPs* within the nucleus. Once inside the nuclear matrix, active *NLPs* bind directly to their target sites, rapidly accelerating the transcription of essential genes, including nitrate transporters, nitrite reductases, and metabolic enzymes. Through this

tightly controlled nucleocytoplasmic shuttling mechanism, *NLPs* orchestrate vital developmental processes ranging from vegetative biomass accumulation and lateral root modification to developmental transitions like seed germination, securing their position as foundational players in plant metabolic homeostasis and nutrient use efficiency [38].

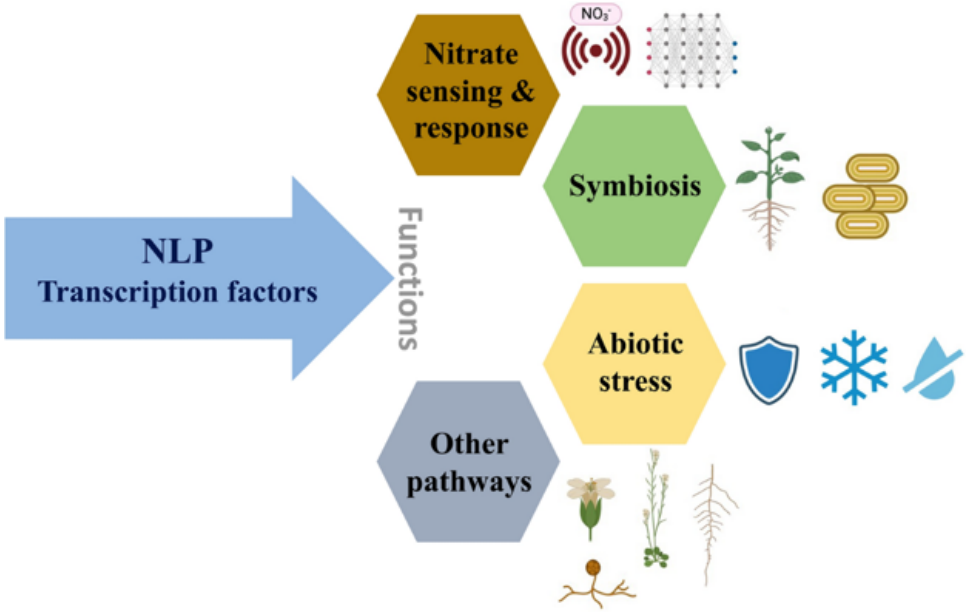


FIGURE 2.10: NLP transcription factors. This figure illustrates that NLP TFs play an important role in nitrate sensing, response, symbiosis, abiotic stress, and other pathways [35].

2.12 RWP-RK Transcription Factors Regulating Nitrogen Response

The RWP-RK gene family constitutes a highly specialized, plant-specific group of transcription factors that play foundational roles in regulating macronutrient systems, reproductive biology, and symbiotic root architecture. Specifically, these regulatory proteins serve as core components of primary nitrogen signaling pathways, cellular differentiation during gametogenesis, and the complex morphological transitions required for root nodule development in leguminous crops. Based on evolutionary divergence and structural configuration, this macromolecular family is divided into two distinct, functional sub-families: the RKD sub-family and the NIN-like Protein sub-family. The RKD sub-family is characterized structurally by

the presence of the conserved RWP-RK DNA-binding domain, which allows these proteins to recognize specific genomic motifs and regulate downstream gene targets primarily associated with egg cell differentiation and early embryonic patterning [39].

In contrast, the NLP sub-family is distinguished by a more complex, multi-domain architecture. While maintaining the core RWP-RK domain for sequence-specific DNA binding, *NLPs* feature a unique, highly conserved PB1 domain at their C-terminus. The addition of this terminal PB1 domain expands the functional capacity of the NLP sub-family, serving as a critical structural platform for mediated protein-protein interactions, as shown in figure 2.10. The PB1 domain facilitates these interactions through a distinct electrostatic mechanism, featuring a conserved arrangement of acidic or basic amino acid residues that create specific positive and negative surface charges. This structural design allows *NLPs* to undergo homo or heterodimerization with other NLP members or to bind with auxiliary regulatory proteins, corepressors, and coactivators within the cell [40].

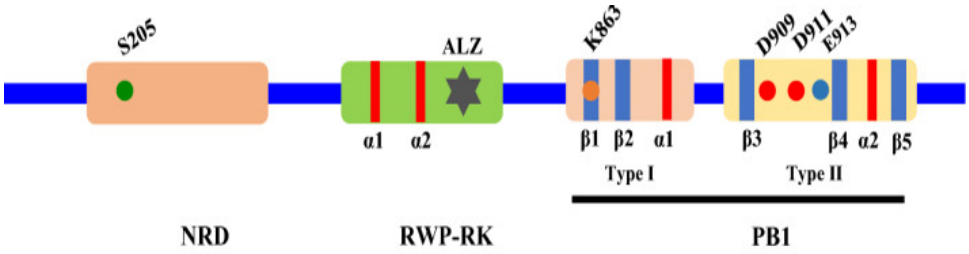


FIGURE 2.11: Plant *NLPs*' molecular structures. Nitrate-responsive domain (NRD), RWP-RK, and Phox and Bem1 (PB1) are the three domains found in *NLPs* [39].

2.13 NLP Gene Role in Nitrogen Signaling

NLPs serve as core elements of the primary nitrate response in higher plants, bridging the gap between external nutrient sensing and massive transcriptomic reprogramming. Under nitrogen-deficient conditions, *NLPs* are typically sequestered in the cytoplasm in an unphosphorylated, inactive state, which prevents unnecessary energy expenditure when substrate levels are low. However, upon the influx or sensing of nitrate signals within the rhizosphere, a rapid intracellular signaling

cascade is initiated. This process activates specific calcium-dependent protein kinases such as CPK10, CPK30, and CPK32, which perceive the transient influx of calcium ions triggered by nitrate entry. These specialized kinases directly phosphorylate specific, highly conserved serine or threonine residues within the regulatory N-terminal region of the NLP structure. This precise post-translational modification induces a conformational change that unmasks the nuclear localization signal, triggering the swift translocation of the transcription factor across the nuclear envelope and ensuring its prolonged retention inside the nucleus [41].

Once active within the nuclear matrix, by directly attaching to conserved Nitrate Responsive Cis-elements found in the promoter regions of essential nitrate transporters and important assimilatory enzymes, *NLPs* serve as primary molecular switches. These target genes include high-affinity and low-affinity nitrate transporters (NRT1.1, NRT2.1), as well as foundational downstream nitrogen-assimilating enzymes such as nitrate reductase and nitrite reductase, as shown in figure 2.12.

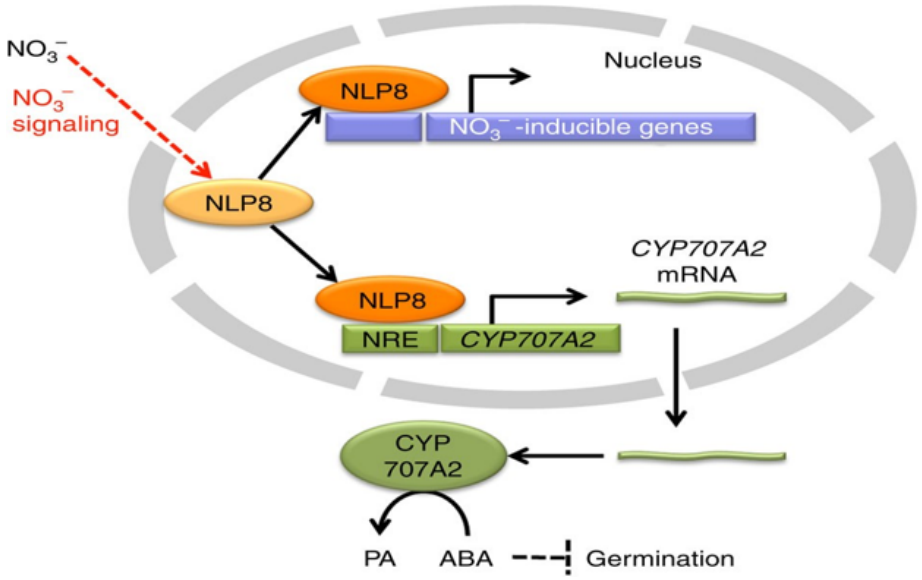


FIGURE 2.12: Role of NLP in nitrate-mediated signaling. This figure depicts how nitrate signaling activates the transcription factor NLP within the nucleus to drive the expression of nitrate-inducible genes [41].

By recruiting necessary transcriptional co-activators and remodeling the local chromatin structure, bound *NLPs* dramatically accelerate the transcription and downstream expression of these metabolic networks.

This coordinated genomic activation optimizes the operational capacity, driving efficient nitrogen utilization, enhancing amino acid biosynthesis, and balancing carbon-to-nitrogen ratios within the plant cell. Through this highly sensitive nucleocytoplasmic shuttling mechanism, *NLPs* allow the plant to dynamically adjust its vegetative growth rate, lateral root branching, and photosynthetic capacity in response to fluctuating environmental nitrogen supplies [42].

2.14 *A. thaliana* NLPs

AtNLP1 to *AtNLP9* genes are present in *A. thalliana*, which are highly conserved structural proteins that function as key genomic switches, as shown in figure 2.12. Within this model plant network, *AtNLP7* is known to be a master regulator of localized signal transduction and nitrate-induced gene expression.

These proteins rely heavily on post-translational modifications, particularly calcium - dependent phosphorylation cascades that unmask nuclear localization signals and initiate vital nuclear retention mechanisms under changing nutrient conditions. Once retained within the nuclear matrix, *AtNLP7* directly coordinates the transcription of primary nitrogen assimilation pathways, managing vegetative growth dynamics and altering root system layout.

Genetic studies using knock-out mutants emphasize that a disruption in *AtNLP7* results in severe nitrogen-starved phenotypes, significantly reducing rosette biomass. Collectively, these transcription factors perform both unique and redundant roles to coordinate baseline vegetative biomass production and nitrate-dependent seed germination under varying nutrient supplies [43].

2.15 *Brassica napus* NLPs

In allotetraploid oilseed *Brassica napus*, thirty-one *BnNLP* genes have been identified and mapped across the complex genomes. The structural distribution of

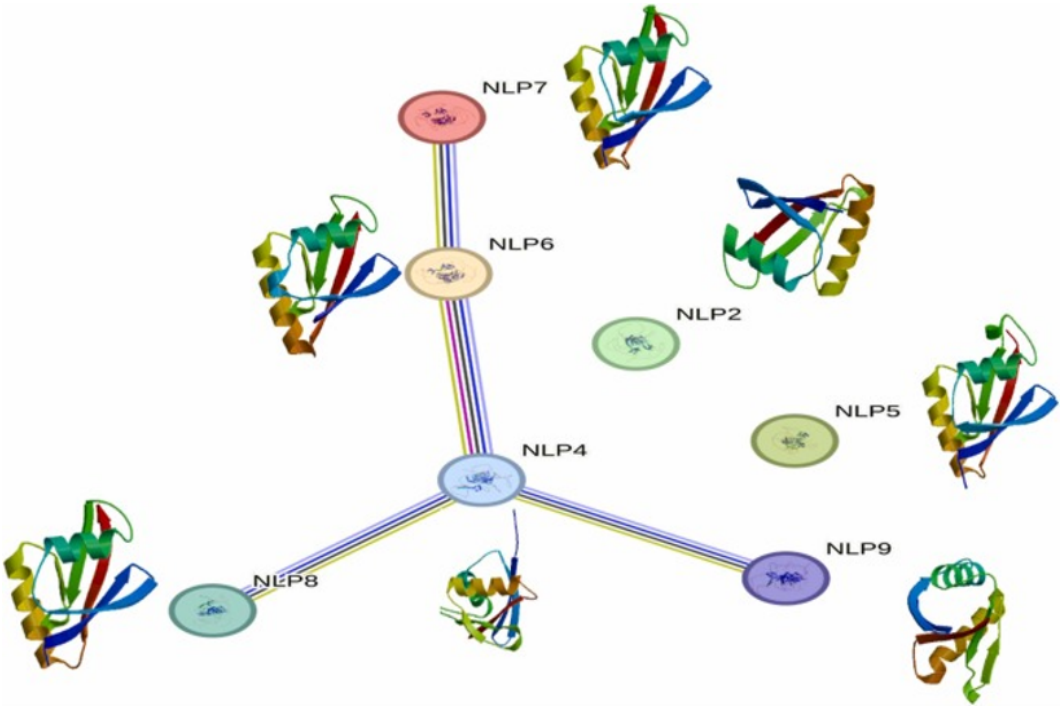


FIGURE 2.13: *Arabidopsis thaliana NLPs*. There are nine different *NLPs* in *Arabidopsis thaliana* that perform different functions [43].

this expanded gene family is primarily the result of ancestral whole-genome duplication and subsequent segmental duplication events common to the evolution of the Brassicaceae lineage. These genes exhibit distinct, organ-specific expression patterns under nitrogen-starved environments, serving unique evolutionary and functional roles in managing low-nutrient stress. Transcriptomic profiling indicates that specific *BnNLP* paralogs are preferentially upregulated in either root or shoot tissues when external nitrogen sources are heavily depleted. This refined division of labor allows the plant to alter its physiological activities and maximize nutrient scavenging without stalling overall metabolic production. Understanding the functional diversity of these genes offers an agronomic path toward improving nitrogen assimilation efficiency and stabilizing seed yield in commercial oil crops [44].

2.16 *Oryza sativa NLPs*

It includes key regulators such as *OsNLP1*, which is heavily induced under nitrogen-deprived conditions. Unlike typical dicot counterparts that respond primarily to

nitrate signals, *OsNLP1* directly targets both nitrate and ammonium assimilation genes, highlighting an evolutionary adaptation to the anaerobic, ammonium-rich soils typical of paddy fields.

Chromatin immunoprecipitation and molecular assays confirm that *OsNLP1* physically binds to the promoters of core transport and assimilatory pathways to drive their activation. Consequently, its overexpression has been shown to substantially increase grain yield, tiller numbers, and overall plant performance under nitrogen-limiting conditions. This broad regulatory capacity makes *OsNLP1* an ideal target for engineering high-yielding, nitrogen-efficient monocot varieties [45].

2.17 *Zea mays NLPs*

ZmNLPs exhibit constitutive yet distinct expression patterns across multiple embryonic and vegetative tissues. These genes act as central regulatory nodes within the plant's primary nitrogen response mechanism, adjusting transcriptomic output based on real-time soil analysis. Key members show significant upregulation immediately following nitrogen application, confirming their foundational value in driving the primary nitrogen response mechanism within monocot crop species.

Upon exposure to nitrate flushes, these factors coordinate the expression of high-affinity transport networks and amino acid biosynthetic enzymes. This swift response maximizes nutrient capture, supports early seedling development, and establishes the base necessary for optimal development and kernel fill [46].

2.18 Importance of Bioinformatics Tools in Soybean Genomic Research

The integration of advanced bioinformatics tools is essential for executing a high-throughput, identification of the NLP gene family throughout the soybean genome and structural analysis. By using specialized computational resources such as

BLASTp searches against the Phytozome database, InterPro for domain validation, and tools like ExPASy ProtParam, MEME, and GSDS, researchers can systematically decode the physical and structural architecture of these transcription factors. Bioinformatics assists in the precise mapping of chromosomal distributions, protein interactions, and evolutionary phylogenetic relationships [47].

This research emphasizes the importance of the NLP gene family in managing nitrogen signaling, cellular equilibrium and plant resistance to stress. Despite the fact that certain cereals and species like *Arabidopsis thaliana* have had these genes thoroughly investigated, a comprehensive analysis of the NLP gene family in soybean remains insufficient. This research aims to fill this gap by creating a computational framework for the structure, evolution, and function of soybean NLP genes. By identifying these genes and their regulatory elements, the study will help to develop resilient soybean varieties that yield high protein and oil in low-nitrogen environments, contributing to sustainable agriculture and food security.

Chapter 3

Research Methodology

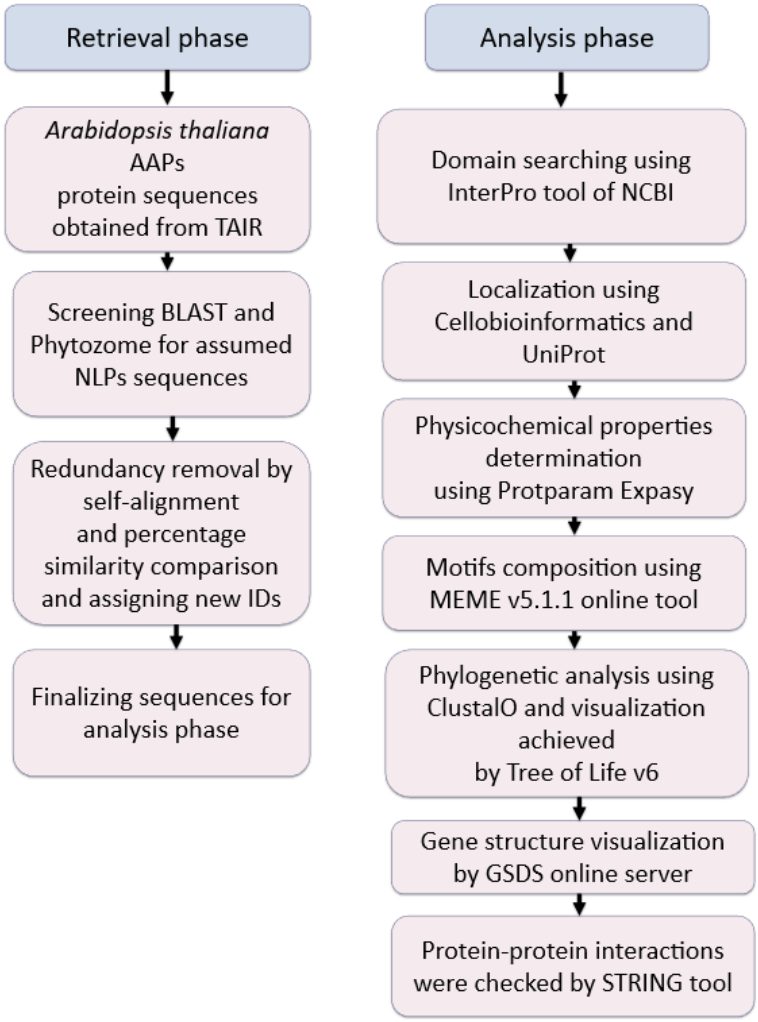


FIGURE 3.1: Methodology flowchart

3.1 Examining Factors of Transcription and Genome Databases

For this study, the *Arabidopsis thaliana* genome was used as the reference point, and the TAIR database (http://arabidopsis.org) was employed in order to acquire the full sequence of genes, amino acids, and codes of all members of the *Arabidopsis thaliana NLP* gene family.

Identification of potential *AtNLPs* was carried out using the NCBI database (https://www.ncbi.nlm.nih.gov/) and TAIR (https://www.arabidopsis.org/). The total number of *NLP* proteins in this family was nine. The sequences of these nine *NLPs* were collected in FASTA format and saved in a notepad file [48].

3.2 Protein Sequences Retrieval of *Glycine max* NLPs

The sequences of *AtNLP* proteins were used to find *NLP* protein sequences of *Glycine max NLPs* through query sequences in the BLAST part of the NCBI database. The results collected from there were saved in a CSV file, and then an Excel sheet was created.

All of the results were aligned, and the accessions that repeated each other were removed to avoid any kind of redundancy. Sequences were sorted, and matching was done to find out the identical sequences.

This is done to make sure that no duplicate sequence exists and that the samples are exact and accurate. After proper filtration, twelve hits were chosen for analysis. Additionally, selected *NLP* protein sequences were acquired from Phytozome (https://phytozome-next.jgi.doe.gov/) using accession numbers. The sequences were retrieved in FASTA format [49].

3.3 Sequence Similarity establishment and Conserved Domain Analysis of *NLPs*

After redundancy removal, the selected twelve sequences were allocated with unique IDs named *GmNLP1* to *GmNLP12*. Sequences with similarity below 80% would not be considered. To confirm the *NLP* status of selected genes, they were assessed for conserved domains linked with *NLP* function, i.e., RWP-RK and PB1 domains. Interpro (https://www.ebi.ac.uk/interpro/) was used to check the presence of these domains, and the results were downloaded in a JSON file to check E-value, bit score, and accession numbers of selected *GmNLPs* [50].

3.4 Determination of Physicochemical Properties and Subcellular Locations

After confirmation of potential *GmNLPs* based on the presence of conserved domains, these *NLPs* were assessed for physicochemical analysis. Physical and chemical properties of *AtNLPs* and some *GmNLPs* were analyzed using Protparam Expasy (https://web.expasy.org/protparam/). The parameters analyzed include molecular weight (MW), hydropathy (GRAVY), and theoretical isoelectric point (pl) [51]. Subcellular localization of some *NLPs* was predicted by using CELLO [52].

3.5 Gene Structure Determination of *GmNLP* Gene Family

To identify the gene structure elements, the gene sequences and coding sequence (CDS) of the *GmNLPs* were subjected to gene structural analysis using the online server GSDS, i.e., Gene Structure Display Server (https://gsds.gao-lab.org/). The CDS and genes of all *NLPs* were downloaded in the FASTA format from the

Phytozome database and compiled into a single file for use as input at the GSDS homepage. In addition, chromosome number, position and lengths of proteins were also established. It may be helpful to compare gene length between two gene families [53].

3.6 The Interactions between Proteins of *GmNLPs*

Protein interactions with the *GmNLPs* and *AtNLPs* were performed using an online, interactive platform for network analysis named STRING (https://string-db.org/). Interacting proteins along with their corresponding figures were downloaded from the website for future references and comparisons.

The interaction of proteins with *GmNLPs* was compared with that of *AtNLPs* to find any functional similarity.

Colorful lines depict the interactions of proteins, whether predicted or known; the colors pink and blue symbolize the known interactions, while the green, dark green, and red symbolize the predicted interactions. The colored balls indicate the nodes, while the colored lines connecting them indicate the edges, which is the protein sequence. In PPI, proteins interacting with *NLP* are linked to nitrogen assimilation, transport, etc. [54].

3.7 Analysis of Motifs in *GmNLPs*

Whether the motifs showed conserved regions was assessed via an online software program known as Multiple Em for Motif Elicitation (MEME v5.1.1) (https://meme-suite.org/meme/tools/meme). It is the most precise online software program for motif elicitation. All parameters were set to the defaults in order to maintain precision. The presence of consensus motifs served as proof of similarities in local regions. [55].

3.8 Phylogeny Evaluation of *GmNLPs*

Phylogenetic tree shows the evolution relationship between the species. For creating the phylogenetic tree of *GmNLPs*, the alignment was done of various species such as *Arabidopsis thaliana*, *Glycine max*, *Brassica napus*, *Zea mays* and *Oryza sativa* using the online web-based software Clustal Omega (https://www.ebi.ac.uk /jdispatcher/msa/clustalo/). Protein FASTA sequences of all these organisms were used as an input and the results were generated in Newick format file, which was further uploaded to another software called iTOL v6 (https://itol.embl.de/) to generate the phylogenetic tree of *Glycine max* with the chosen plants. [56].

Chapter 4

Results

4.1 Analysis of Genes of *NLP* and Conserved Domains for *Glycine max*

For the present study, the complete sequences of the coding region of the *NLP* genes of the *NLP* gene family from *Arabidopsis thaliana* were extracted from the TAIR database. Nine different variations of the *NLP* genes were found, and for each variant, the longest one was chosen. After that, these sequences underwent protein BLAST analysis using NCBI. The output was saved in CSV form and was tabulated in Excel. Self-alignment was done for these sequences to obtain a similarity percentage table.

After a detailed filtration analysis, twelve candidate *NLP* genes from *Glycine max* were selected, and their sequences were downloaded in FASTA format from Phytozome. Conserved domain analysis was performed simultaneously for validation by using InterPro. *NLP* genes in the *Glycine max* genome were identified using their protein sequences. Domains were searched along their pfam accession numbers. All twelve selected *Glycine max* genes have both domains, confirming their validity as NIN-like proteins. These genes were assigned names as *GmNLP1* to *GmNLP12*. The names of these genes along with their position, E-value, bitscore and pfam accession numbers are given in table 4.1 below.

TABLE 4.1: Conserved domains of *NLP* gene families of *Arabidopsis thaliana* and *Glycine max*

Organism	**Query**	**Hit type**	**Position**		**E-value**	**Bit Score**	**Accession Number**	**Short Name**
			From	**To**				
Arabidopsis thaliana	*AtNLP1*	specific	812	893	6.54E-41	112.802	cd06407	PB1-NLP
			608	656	6.62E-24	54.0518	pfam02042	RWP-RK
	AtNLP2	specific	896	944	1.26E-41	112.802	cd06407	PB1-NLP
			648	696	4.74E-24	54.0518	pfam02042	RWP-RK
	AtNLP3	specific	674	758	1.56E-40	112.802	cd06407	PB1-NLP
			498	546	1.69E-24	54.0518	pfam02042	RWP-RK
	AtNLP4	specific	745	826	7.15E-43	112.802	cd06407	PB1-NLP
			558	606	1.67E-24	54.0518	pfam02042	RWP-RK
	AtNLP5	specific	711	787	3.72E-36	112.802	cd06407	PB1-NLP
			549	597	5.58E-25	54.0518	pfam02042	RWP-RK
	AtNLP6	specific	742	822	3.01E-34	112.802	cd06407	PB1-NLP
			556	604	8.48E-25	54.0518	pfam02042	RWP-RK
	AtNLP7	specific	866	944	4.32E-34	112.802	cd06407	PB1-NLP
			591	639	4.66E-25	54.0518	pfam02042	RWP-RK
	AtNLP8	specific	848	928	6.25E-39	112.802	cd06407	PB1-NLP
			590	651	1.25E-20	54.0518	pfam02042	RWP-RK
	AtNLP9	specific	793	874	3.37E-34	112.802	cd06407	PB1-NLP
			535	584	5.08E-25	54.0518	pfam02042	RWP-RK
Glycine max	*GmNLP1*	specific	815	896	8.18E-44	151.322	cd06407	PB1-NLP
			588	660	3.70E-22	89.6	pfam02042	RWP-RK
	GmNLP2	specific	688	769	7.49E-40	139.76	cd06407	PB1-NLP
			511	584	5.10E-21	85.9	pfam02042	RWP-RK
	GmNLP3	specific	817	898	3.00E-15	67.3	cd06407	PB1-NLP
			590	662	5.70E-22	89	pfam02042	RWP-RK
	GmNLP4	specific	891	971	1.55E-37	133.602	cd06407	PB1-NLP
			571	644	1.20E-22	91.1	pfam02042	RWP-RK
	GmNLP5	specific	897	977	4.05E-39	137.84	cd06407	PB1-NLP
			580	652	1.70E-23	93.9	pfam02042	RWP-RK
	GmNLP6	specific	866	946	7.75E-31	114.342	cd06407	PB1-NLP
			595	665	4.30E-22	89.4	pfam02042	RWP-RK
	GmNLP7	specific	874	954	1.13E-31	116.654	cd06407	PB1-NLP
			597	669	9.70E-23	91.4	pfam02042	RWP-RK
	GmNLP8	specific	888	968	5.12E-35	126.284	cd06407	PB1-NLP
			570	640	4.50E-21	86.1	pfam02042	RWP-RK
	GmNLP9	specific	816	897	3.00E-15	67.3	cd06407	PB1-NLP
			589	661	5.70E-22	89	pfam02042	RWP-RK
	GmNLP10	specific	872	952	2.06E-30	113.187	cd06407	PB1-NLP
			595	665	2.20E-22	90.3	pfam02042	RWP-RK
	GmNLP11	specific	897	977	6.50E-38	134.373	cd06407	PB1-NLP
			579	651	1.7E-23	93.9	pfam02042	RWP-RK

Table 4.1 continued from previous page

Organism	Query	Hit type	Position		E-value	Bit Score	Accession Number	Short Name
			From	To				
	*GmNLP12*specific		814	895	8.32E-44	151.322	cd06407	PB1-NLP
			587	659	3.70E-22	89.6	pfam02042	RWP-RK

4.2 Determination of Physicochemical Properties

The physicochemical properties of *NLP* proteins from *A. thaliana* and *Glycine* max were analyzed using the ExPASy ProtParam tool. Protein molecular weight, Grand Average of Hydropathicity and theoretical isoelectric point were evaluated as given in table **??**. It shows that the molecular weights, pI values, and GRAVY values of both plant species are closer. All proteins showed negative GRAVY values, indicating a hydrophilic nature. Moreover, the chromosome number, position, gene and protein length were also checked for all sequences. Subcellular locations were also analyzed for all selected *NLP* sequences using the CELLO tool. Data depict the average length for both species, and subcellular localization analysis suggests that *NLP* proteins of both species are localized in the nucleus.

Table 4.2

Plant	Gene name	Chromosome	Position	Gene length (bp)	Protein length (aa)	Molecular weight	Isoelectric point	GRAVY	Localization
Arabidopsis thaliana	*AtNLP1*	1	22324437-22327359	2933	909	100886.25	5.09	-0.443	Nuclear
	AtNLP2	4	16777264-16782054	4791	963	107278.54	5.65	-0.476	Nuclear
	AtNLP3	4	17954710-17958063	3354	767	85066.66	8.35	-0.271	Nuclear
	AtNLP4	4	7154410-7158287	3878	844	94231.93	5.45	-0.472	Nuclear
	AtNLP5	5	17367827-17369510	1684	808	90684.13	5.8	-0.467	Nuclear
	AtNLP6	4	950546-953690	3145	841	93863.46	5.95	-0.356	Nuclear
	AtNLP7	5	1698270-1702659	4390	959	105742.04	5.55	-0.42	Nuclear
	AtNLP8	2	18061716-18066708	4993	947	104883.79	5.47	-0.43	Nuclear

Table 4.2 continued from previous page

Plant	Gene name	Chromosome	Position	Gene length (bp)	Protein length (aa)	Molecular weight	Isoelectric point	GRAVY	Localization
	AtNLP9	3	22009008-22012804	3797	894	98712.97	5.43	-0.383	Nuclear
Glycine max	*GmNLP1*	4	1352713-1357474	4800	908	100821.32	5.71	-0.439	Nuclear
	GmNLP2	4	72327-75769	6200	784	86862.93	5.51	-0.31	Nuclear
	GmNLP3	6	1338316-1343018	5702	910	101404.11	5.6	-0.467	Nuclear
	GmNLP4	9	33895570-33900191	4872	930	109527.15	5.94	-0.376	Nuclear
	GmNLP5	10	46304592-46310226	3405	931	109783.52	5.53	-0.412	Nuclear
	GmNLP6	11	9530121-9536109	4732	965	106778.91	5.64	-0.464	Nuclear
	GmNLP7	15	2228854-2237078	4600	972	107384.25	5.85	-0.393	Nuclear
	GmNLP8	16	34323617-34328331	5687	931	109305.98	6.09	-0.365	Nuclear
	GmNLP9	6	1338316-1343018	5990	909	101275.97	5.6	-0.464	Nuclear
	GmNLP10	12	3589379-3595580	8000	971	107666.98	6	-0.475	Nuclear
	GmNLP11	20	39860746-39866224	4895	931	109646.3	5.64	-0.402	Nuclear
	GmNLP12	4	1352713-1357474	4897	907	100693.19	5.71	-0.436	Nuclear

4.3 Identification of Structure of Gene

Structural aspects of *AtNLPs* and *GmNLPs* were analyzed using the information on their gene and coding sequences, including exons, introns and untranslated regions. Gene structure analysis was carried out using an online tool Gene Structure Display Server (GSDS). Gene sequences and CDS sequences in FASTA format were taken from Phytozome database and used as input.

The output provided graphical representations of gene organization and exon-intron distribution among *NLP* family members, as shown in figure 4.1. Yellow boxes represent coding sequences, i.e., CDS/exons.

Black lines represent introns, i.e., non-coding regions. Blue boxes represent upstream and downstream untranslated regions. The scale at the bottom indicates gene length in kilobases.

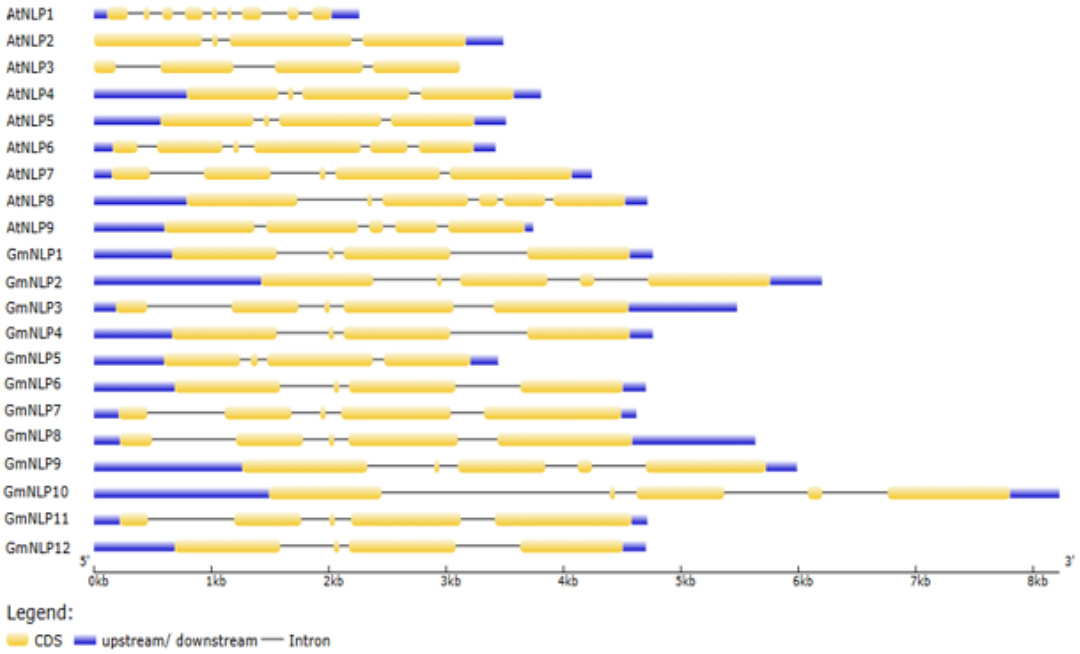


FIGURE 4.1: Gene structure determination of *A. thaliana* and *Glycine max*

The gene structure analysis revealed that *NLP* genes of both *A. thaliana* and *G. max* possess multiple exons and introns, indicating a conserved gene organization within the *NLP* family. Most genes showed a similar exon-intron arrangement, suggesting that these genes have been highly conserved during evolution. The larger gene sizes observed in some soybean *NLP* genes indicate species-specific diversification after gene duplication events.

4.4 The Interactions Between Proteins of *GmNLPs*

The proteins cannot work independently, they need to form a complex structure. A network has been predicted online using the STRING database (https://www.expasy.org/resources/string). All of the *GmNLP* proteins have been shown to interact with several N-interacted genes. The protein FASTA sequence was given as input, and the results were downloaded in form of interaction image. The proteins interacting with *NLP* are those responsible for nitrogen uptake, assimilation &

transport, hence predicting their involvement in NUE as shown in figures and tables below.

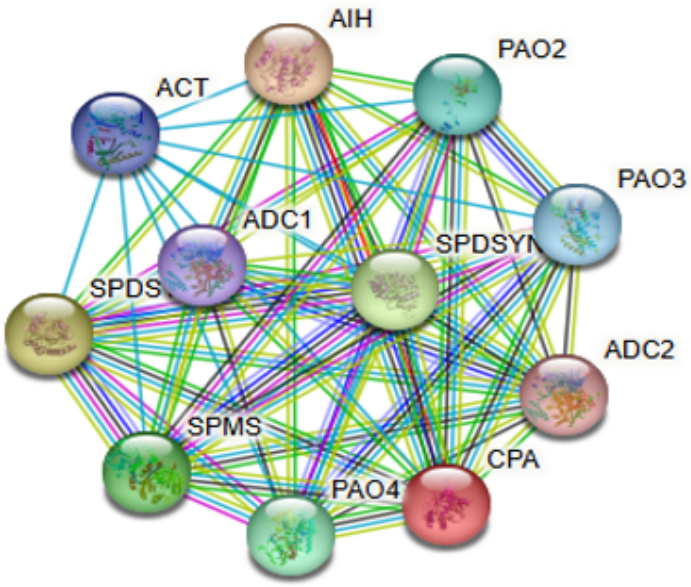


FIGURE 4.2: *AtNLP1* results

TABLE 4.3: *AtNLP1* analysis data

Patterns of Function	
PAO3	The flavoenzyme involved in the back-conversion of polyamines is polyamine oxidase 3.
PAO4	Flavoenzyme involved in polyamine back-conversion; likely polyamine oxidase 4.
PAO2	The flavoenzyme involved in the back-conversion of polyamines is polyamine oxidase 2.
ACT	The production of hydroxycinnamic acid amides involves the enzyme agmatine coumaroyltransferase.
CPA	N-carbamoylputrescine amidase: Participates in the production of polyamines.
AIH	Agmatine deiminase: facilitates the conversion of agmatine to N-carbamoylputrescine.
SPMS	synthase for spermine.
SPDSYN1	S-adenosylmethionine decarboxylase 1 alpha chain: necessary for spermidine and spermine production.
SPDSYN2	S-adenosylmethionine decarboxylase 2 alpha chain: necessary for spermidine and spermine production.

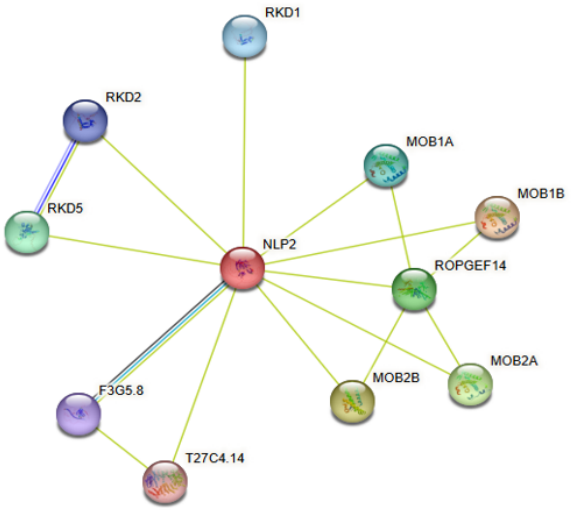


FIGURE 4.3: *AtNLP2* results

TABLE 4.4: *AtNLP2* analysis data

Patterns of Function	
RKD1	potential transcriptional factor.
MOB1A	Cell division and growth are regulated by MOB kinase activator-like 1A.
MOB1B	MOB kinase activator-like 1B is a member of the family MOB1/phocein.
RKD2	potential transcriptional factor.
RKD5	Potential transcription factor.

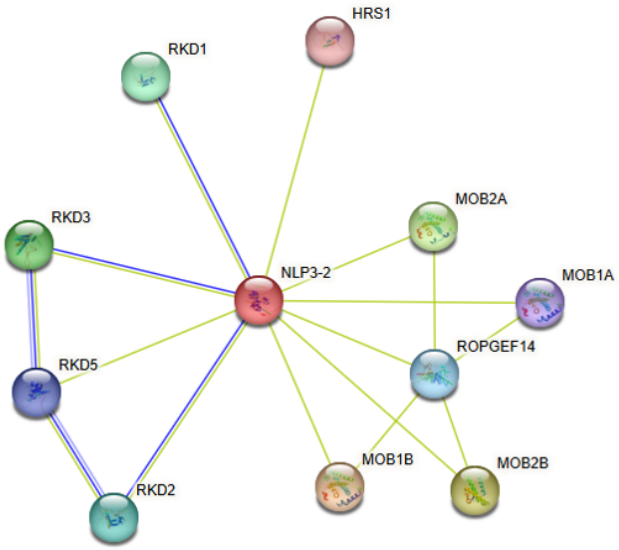


FIGURE 4.4: *AtNLP3* results

TABLE 4.5: *AtNLP3* analysis data

Patterns of Function	
HRS1	Nitrate and phosphate signaling in roots is mediated by the transcription factor HRS1.
RKD1	Potential transcription factor.
RKD3	Potential transcription factor.
MOB1B	MOB kinase activator-like 1B is a member of the MOB1/phocein family.
MOB2B	2B is a putative MOB kinase activator.

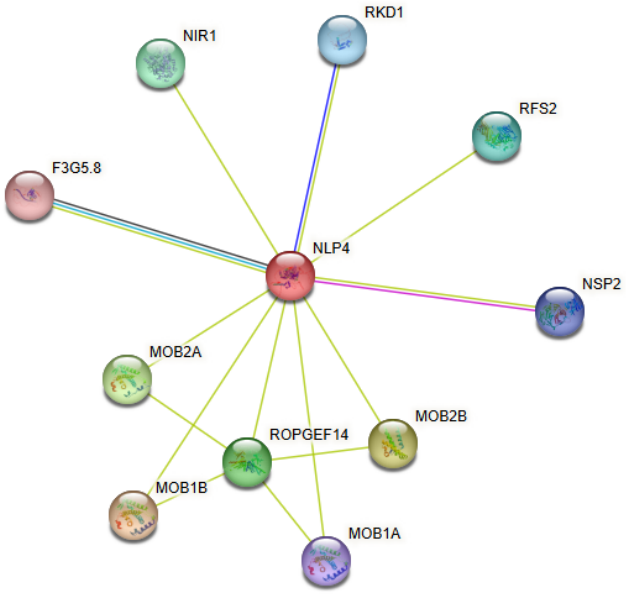


FIGURE 4.5: *AtNLP4* results

TABLE 4.6: *AtNLP4* analysis data

Predicted functional patterns	
RSF2	Probable galactinol–sucrose galactosyltransferase 2; Transglycosidase operating by a ping-pong reaction mechanism.
NIR1	The six-electron reduction of nitrite to ammonium is catalyzed by ferredoxin-nitrite reductase, a chloroplastic.
NSP2	Nitrile-specifier protein 2: Encourages the production of simple nitriles but not thiocyanates or epithionitriles.

Table 4.6 continued from previous page

Predicted functional patterns	
F3G5.8	The gyp1p superfamily protein's Ypt/Rab-GAP domain.
MOB2A	MOB kinase activator-like 2A.

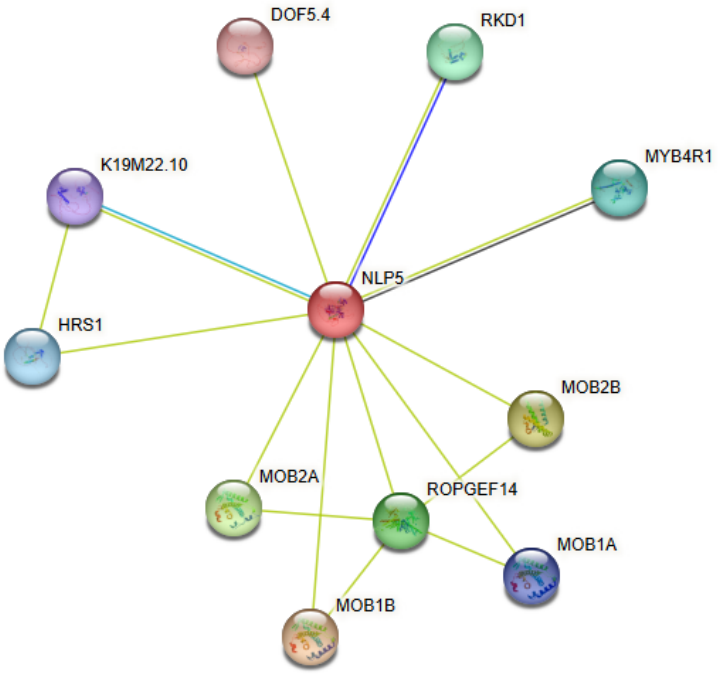


FIGURE 4.6: *AtNLP5* results

TABLE 4.7: *AtNLP5* analysis data

Patterns of Function	
HRS1	Nitrate and phosphate signaling in roots is mediated by the transcription factor HRS1.
DOF5.4	The transcription factor Dof zinc finger protein DOF5.4 binds exclusively to a 5′-AA[AG]G-3′ consensus core sequence.
PKD1	potential transcription factor.
ROPGEF14	Rop guanine nucleotide exchange factor 14 is a guanine-nucleotide exchange factor (GEF) that activates Rop.
MOB1B	MOB kinase activator-like 1B is a member of the MOB1/phocein family.

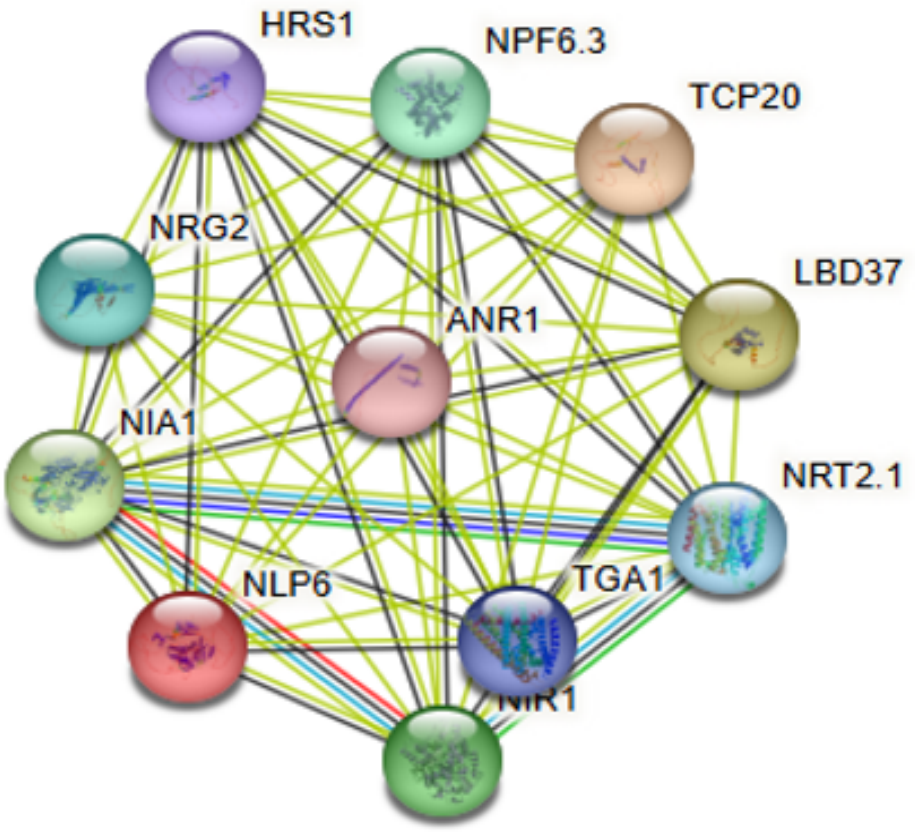


FIGURE 4.7: *AtNLP6* results

TABLE 4.8: *AtNLP6* analysis data

Patterns of Function	
ANR1	ANR1, a probable transcription factor for MADS-box transcription.
NPF6.3	Dual affinity nitrate transporter, protein NRT1/PTR FAMILY 6.3.
NRG2	Nitrate signaling requires the nitrate regulating gene 2 protein.
LBD37	protein 37 that has the LOB domain.
TCP20	TCP20 is a transcription factor that attaches itself to ribosomal protein genes and the site II motif.
TGA1	TGA1 is a transcriptional activator that binds exclusively to the DNA sequence 5′-TGACG-3′.

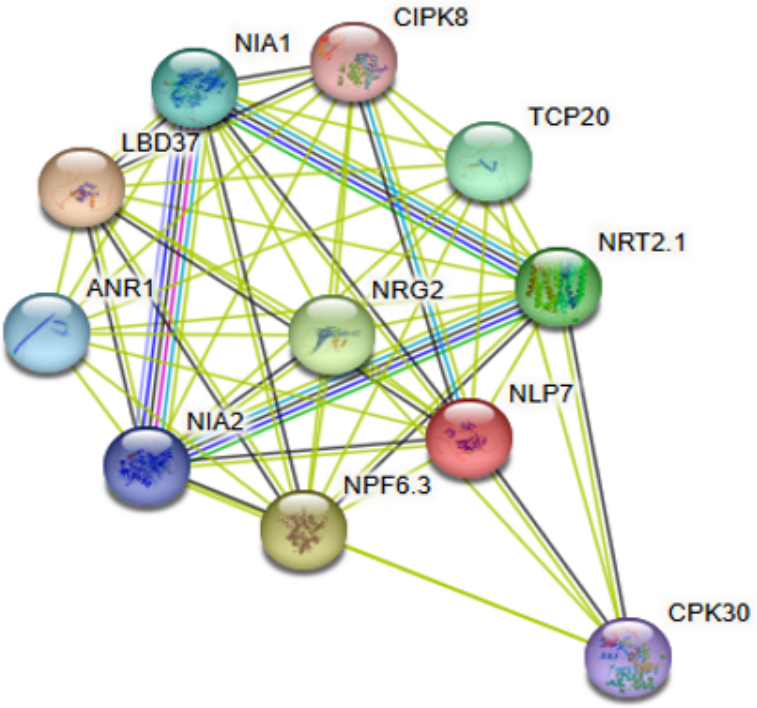


FIGURE 4.8: *AtNLP7* results

TABLE 4.9: *AtNLP7* analysis data

Patterns of Function	
NIA1	One important enzyme involved in the initial stage of nitrate absorption in plants is nitrate reductase [NADH] 1.
LBD37	protein 37 that has the LOB domain.
CIPK8	CIPK serine-threonine protein kinases interact with CBL proteins, as does CBL-interacting serine/threonine-protein kinase 8.
NIA2	A crucial enzyme in the initial stage of nitrate absorption in plants is nitrate reductase [NADH] 2.
NRT2.1	Although it is involved in nitrate transport, high-affinity nitrate transporter 2.1 does not appear to be able to mediate transport on its own.
NRG2	Nitrate signaling requires the nitrate regulating gene 2 protein.

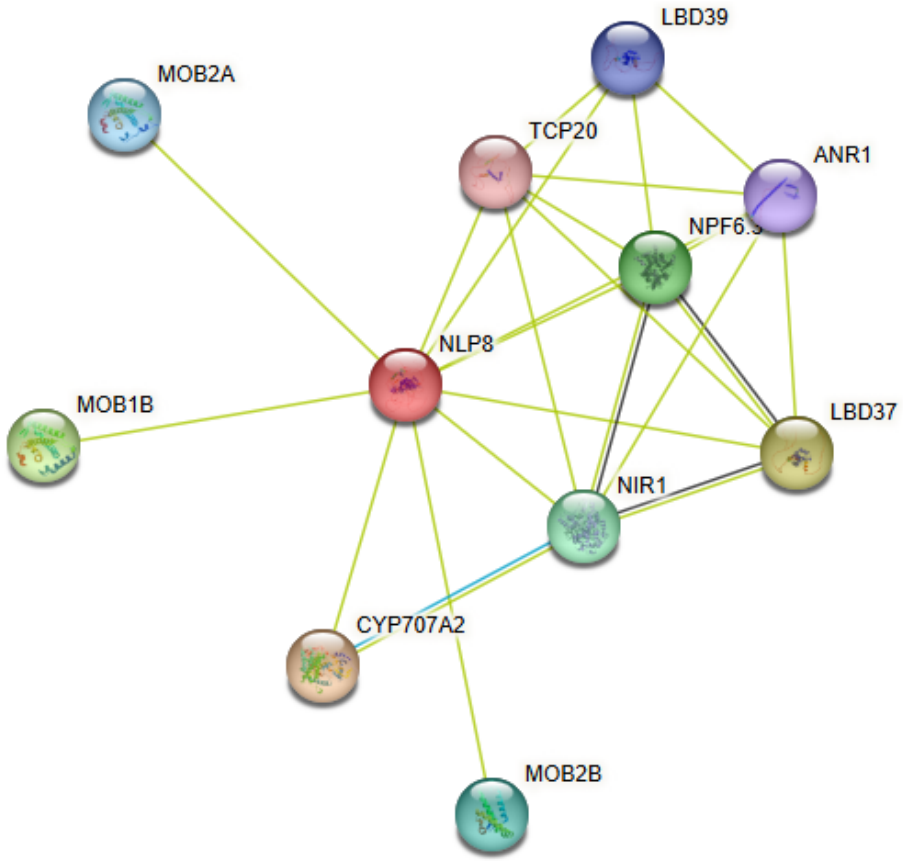


FIGURE 4.9: *AtNLP8* results

TABLE 4.10: *AtNLP8* analysis data

Predicted functional patterns	
CYP707A2	The oxidative breakdown of abscisic acid is facilitated by abscisic acid 8′-hydroxylase 2.
TCP20	TCP20 is a transcription factor that attaches itself to ribosomal protein genes and the site II motif.
LBD37	protein 37 that has the LOB domain.
NIR1	The six-electron reduction of nitrite to ammonium is catalyzed by ferredoxin-nitrite reductase, a chloroplastic.
MOB1B	MOB kinase activator-like 1B is a member of the MOB1/phocein family.

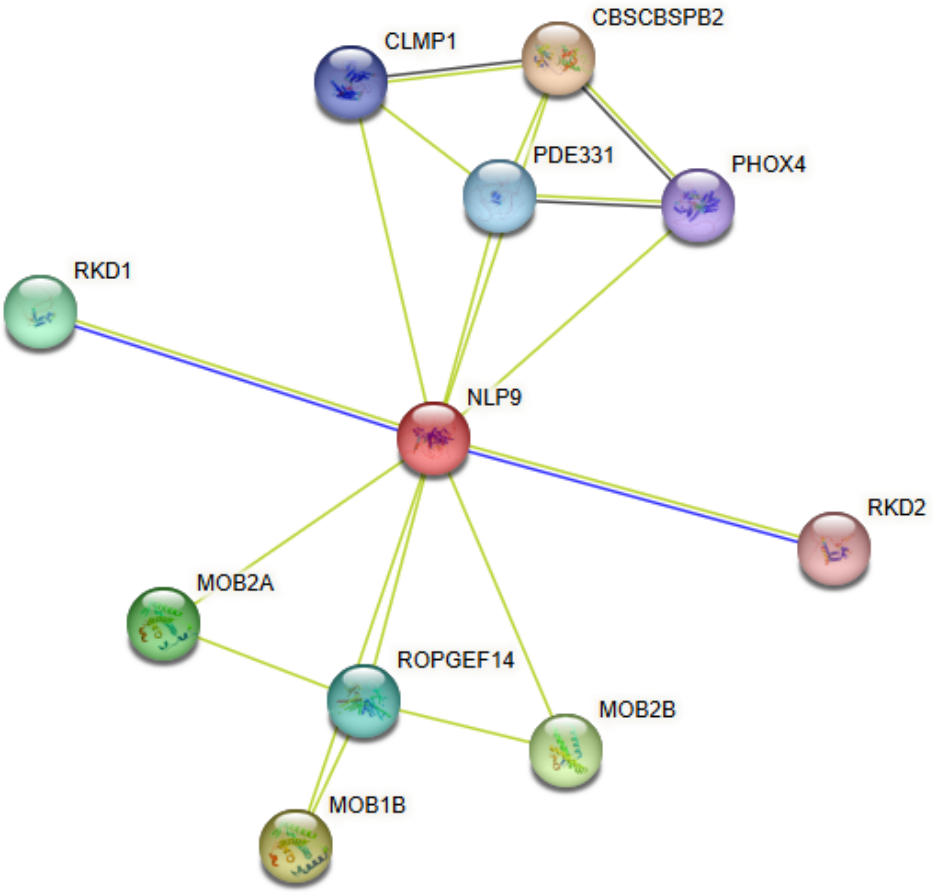


FIGURE 4.10: *AtNLP9* results

TABLE 4.11: *AtNLP9* analysis results

Patterns of Function	
ROPGEF14	Guanine-nucleotide exchange factor (GEF) that activates Rop is known as Rop guanine nucleotide exchange factor 14.
PDE331	Phox/Bem1p family protein/octicosapeptide.
CLMP1	The protein CLMP1 is necessary for the partitioning and separation of plastids during cell division.
PHOX4	The carboxylate clamp-type tetratricopeptide repeat protein PHOX4 has the potential to function as a chaperone.
CBSCBSPB2	CBSCBSPB2, a protein with a CBS domain.

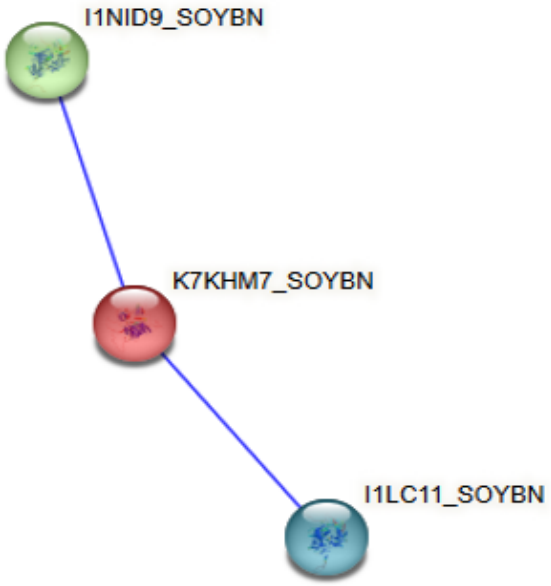


FIGURE 4.11: *GmNLP1* results

TABLE 4.12: *GmNLP1* analysis of data

Patterns of function	
1NID9_SOYBN	protein with a PHD domain.
K7KHM7_SOYBN	protein that has not been identified.
1LC11_SOYBN	protein with a PHD domain.

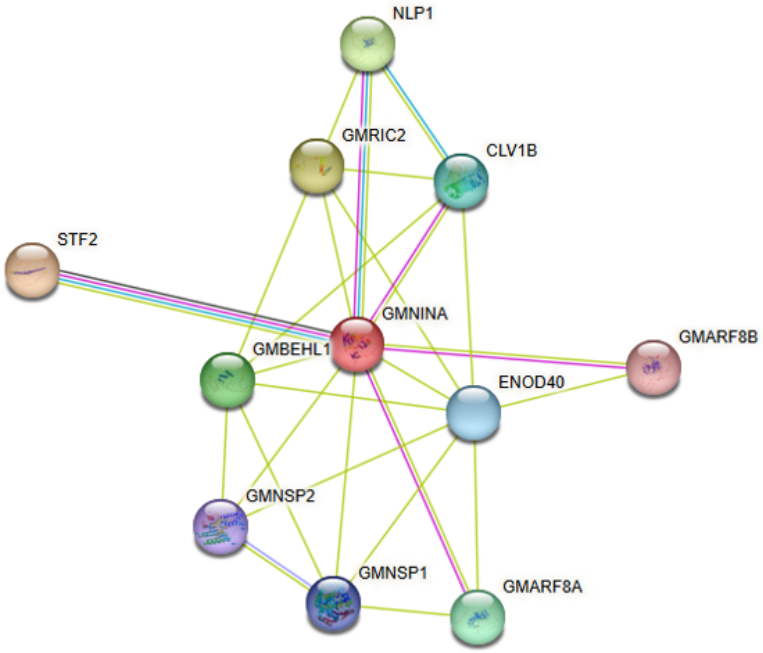


FIGURE 4.12: *GmNLP1* results

TABLE 4.13: *GmNLP2* analysis

Patterns of function	
GMARF8B	Auxin response factor; Auxin response factors (ARFs) are transcriptional factors that bind specifically to the DNA sequence 5′-TGTCTC-3′.
GMNSP1	protein that has a GRAS domain and is a member of the GRAS family.
GMBEHL1	protein with the BES1_N domain.
STF2	Uncharacterized protein.
ENOD40	Early nodulin-40; Modulates the action of auxin, and may function as plant growth regulator.
GMNINA	Uncharacterized.

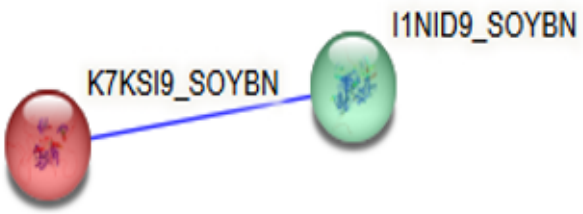


FIGURE 4.13: *GmNLP3* results

TABLE 4.14: *GmNLP3* analysis

Patterns of function	
K7KSI9_SOYBN	Uncharacterized protein.
1NID9_SOYBN	Protein with PHD part.

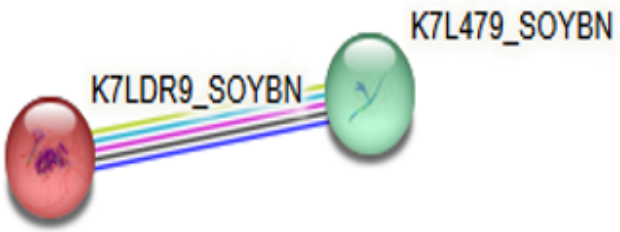


FIGURE 4.14: *GmNLP4* results

TABLE 4.15: *GmNLP4* analysis

Patterns of function	
K7LDR9_SOYBN	Uncharacterized protein.
K7L479_SOYBN	Protein with J portion.

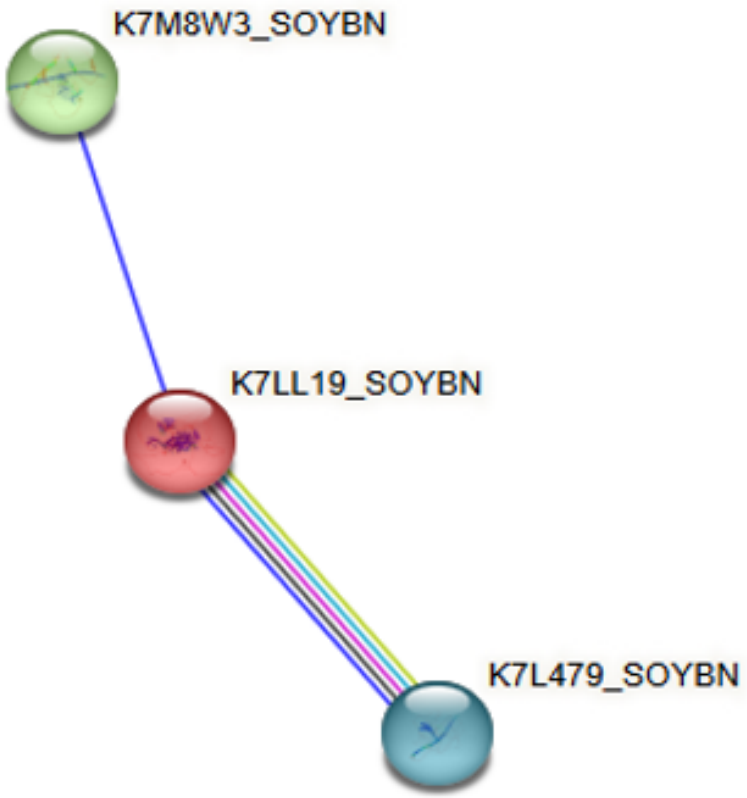


FIGURE 4.15: *GmNLP5* results

TABLE 4.16: *GmNLP5* analysis

Patterns of function	
K7M8W3_SOYBN	Uncharacterized protein.
K7LL19_SOYBN	Uncharacterized protein.
K7L479_SOYBN	Protein with J portion.

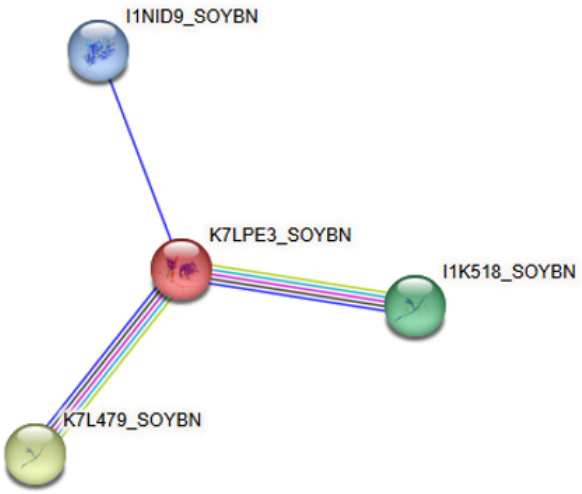


FIGURE 4.16: *GmNLP6* results

TABLE 4.17: *GmNLP6* analysis

Patterns of function	
1NID9_SOYBN	PHD domain-containing protein.
K7LPE3_SOYBN	Uncharacterized protein.
1K518_SOYBN	Protein with J portion.
K7L479_SOYBN	Protein with J portion.

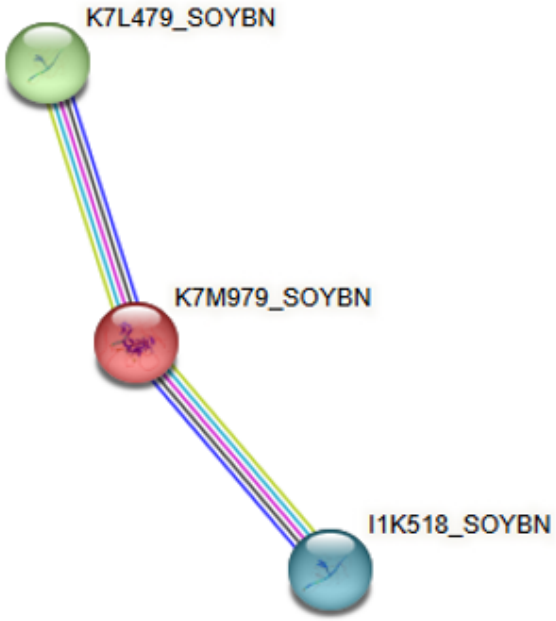


FIGURE 4.17: *GmNLP7* results

TABLE 4.18: *GmNLP7* analysis

Patterns of function	
K7L479_SOYBN	Protein with J portion
K7M979_SOYBN	Uncharacterized protein.
1K518_SOYBN	Protein with J portion

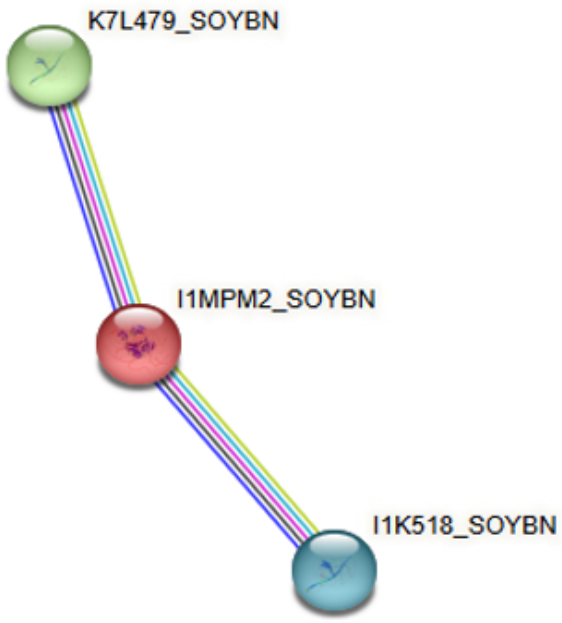


FIGURE 4.18: *GmNLP8* results

TABLE 4.19: *GmNLP8* analysis

Patterns of function	
K7L479_SOYBN	Protein with J domain.
1MPM2_SOYBN	Uncharacterized protein.
1K518_SOYBN	Protein with J domain.

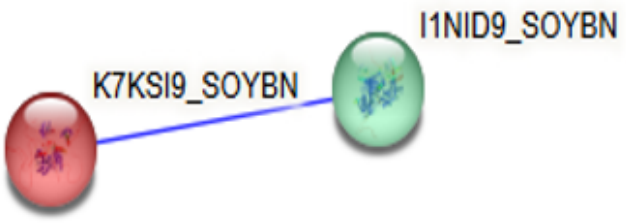


FIGURE 4.19: *GmNLP9* results

TABLE 4.20: *GmNLP9* analysis

Patterns of function	
1NID9_SOYBN	PHD domain-containing protein.
K7KSI9_SOYBN	Uncharacterized

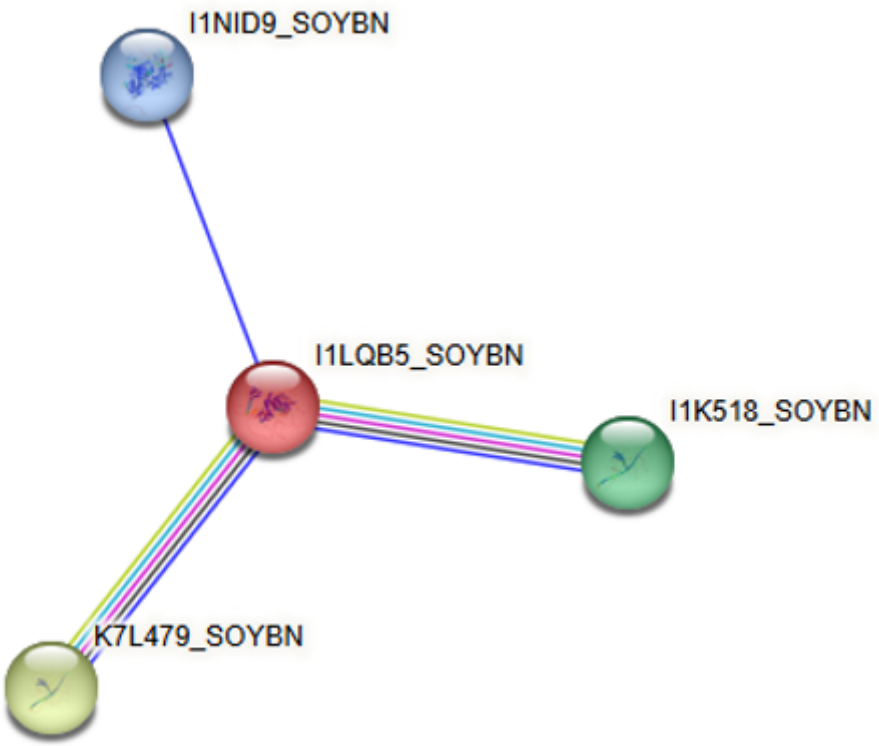


FIGURE 4.20: *GmNLP10* results

TABLE 4.21: *GmNLP10* analysis

Patterns of function	
1NID9_SOYBN	PHD domain-containing protein.
1LQB5_SOYBN	Uncharacterized protein.
1K518_SOYBN	Protein with J portion.
K7L479_SOYBN	Protein with J portion.

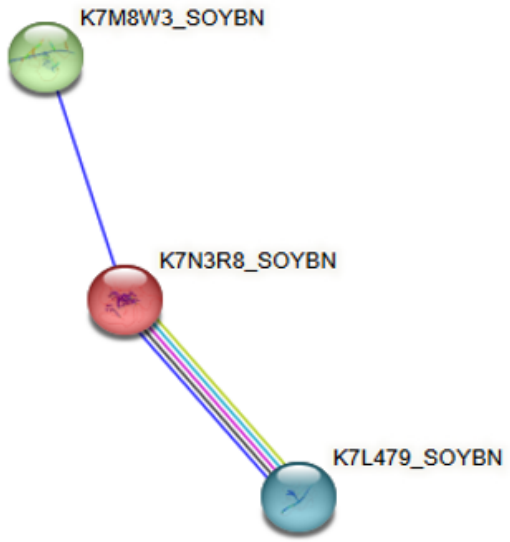


FIGURE 4.21: *GmNLP11* results

TABLE 4.22: *GmNLP11* analysis

Patterns of function	
K7L479_SOYBN	J domain-containing protein.
K7N3R8_SOYBN	Uncharacterized
K7M8W3_SOYBN	Uncharacterized

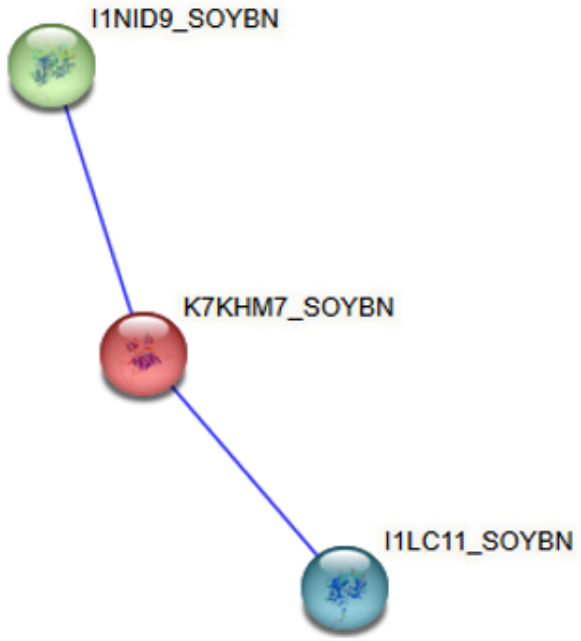


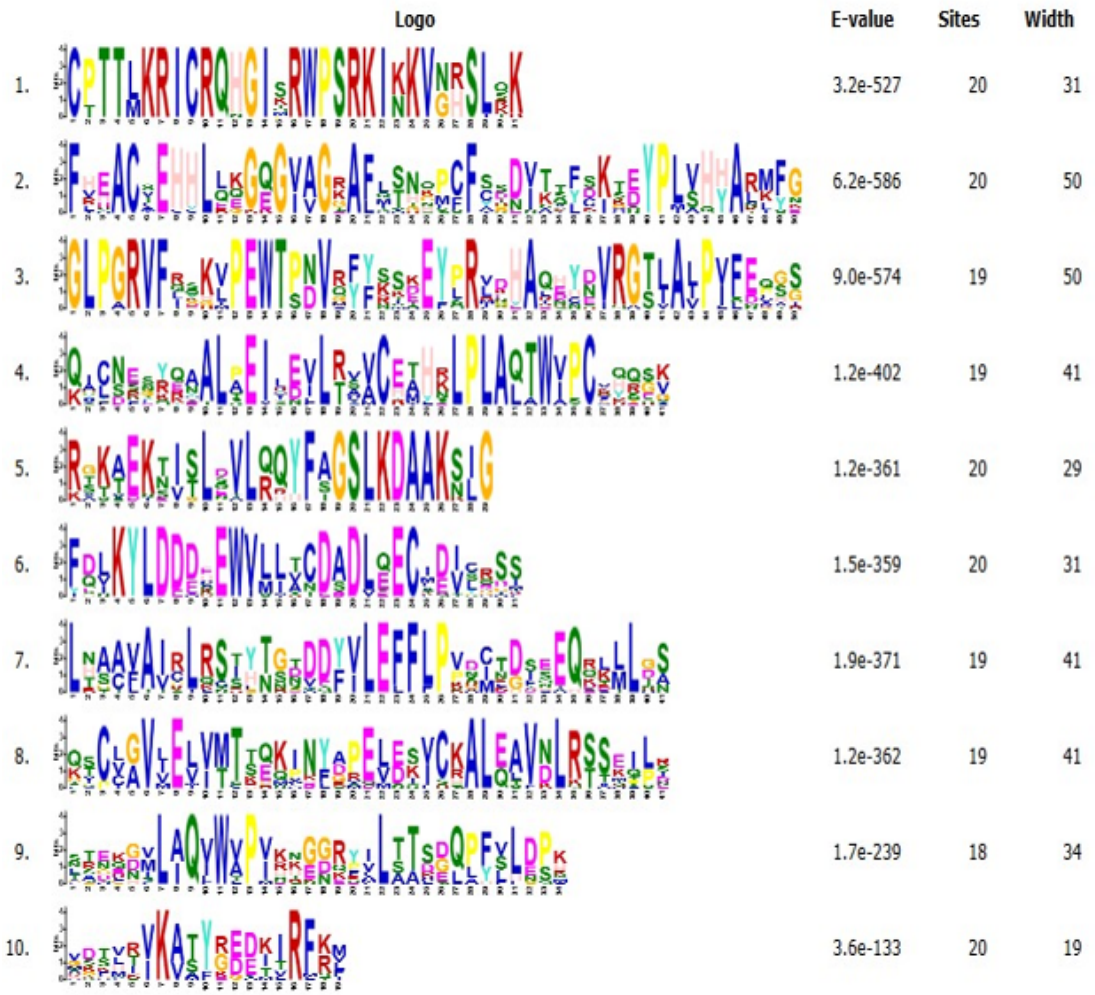
FIGURE 4.22: *GmNLP12* results

TABLE 4.23: *GmNLP12* analysis

Patterns of function	
1LC11_SOYBN	PHD domain-containing protein.
K7KHM7_SOYBN	Uncharacterized protein.
1NID9_SOYBN	PHD domain-containing protein.

4.5 Motif Makeup in *AtNLPs* and *GmNLPs*

The sequence information for all 21 proteins in the *AtNLPs* and *GmNLPs* was utilized for the discovery of consensus motifs. The results obtained in the form of a colored logo along with E-value sites and widths are shown in table 4.24 below.



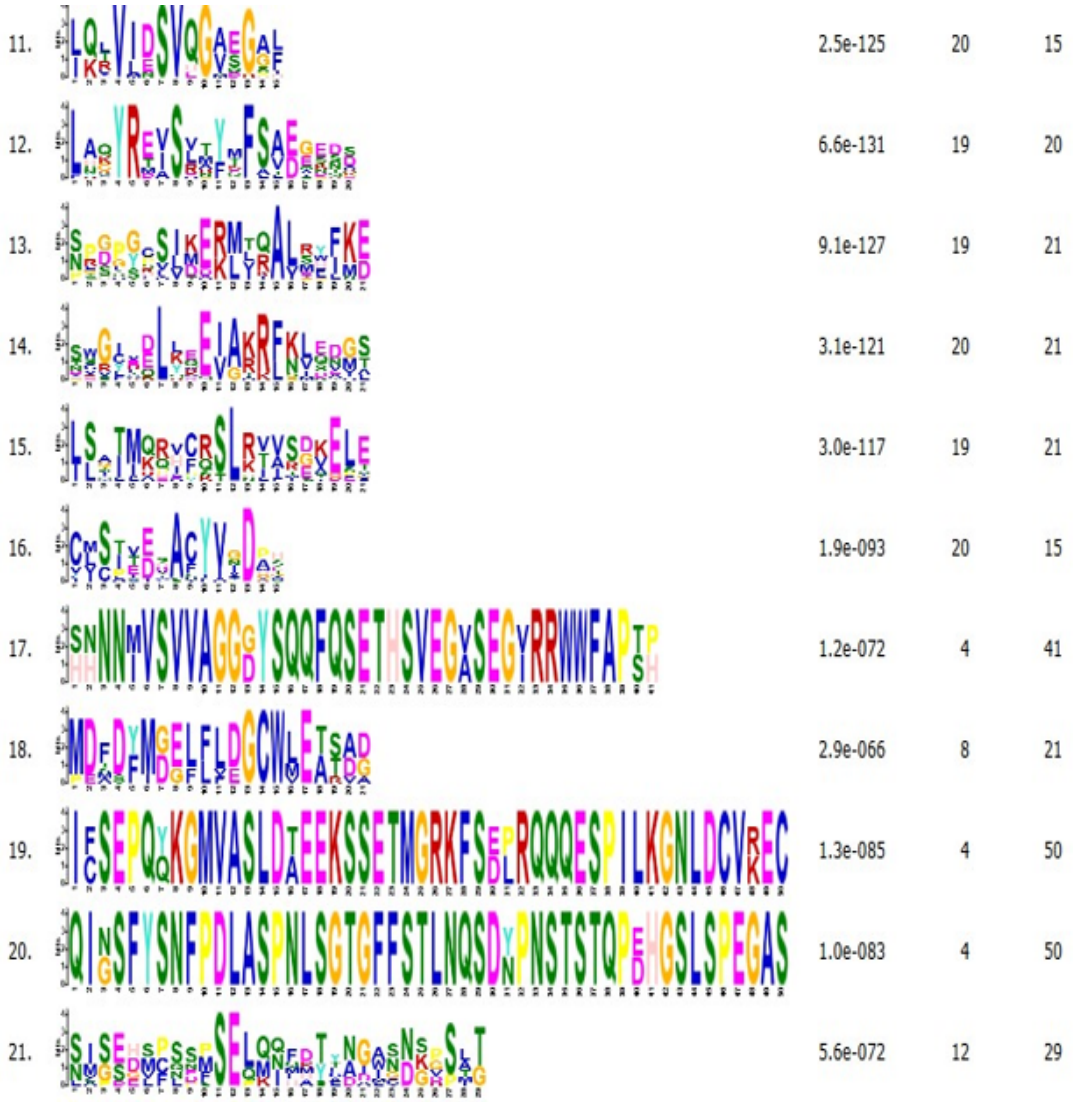


FIGURE 4.23: Logos of identified motifs

In MEME analysis, 21 conserved motifs were found in the *NLP* proteins of *Arabidopsis thaliana* and *Glycine max*. Motifs 1-9 in *A. thaliana* and motifs 10-16 in G. max were found in almost every member of the family and had very low E-values, indicating significant evolutionary conservation.

The major functional areas of *NLP* proteins are probably represented by these motifs. On the other hand, only a portion of soybean *NLP* proteins contained patterns 17-21, indicating the possibility of species-specific diversification after gene duplication events. In comparison to *A. thaliana*, soybean *NLP* proteins generally showed more motifs and higher motif variability, suggesting greater structural complexity and possible functional specialization within the *Glycine max NLP* gene family.

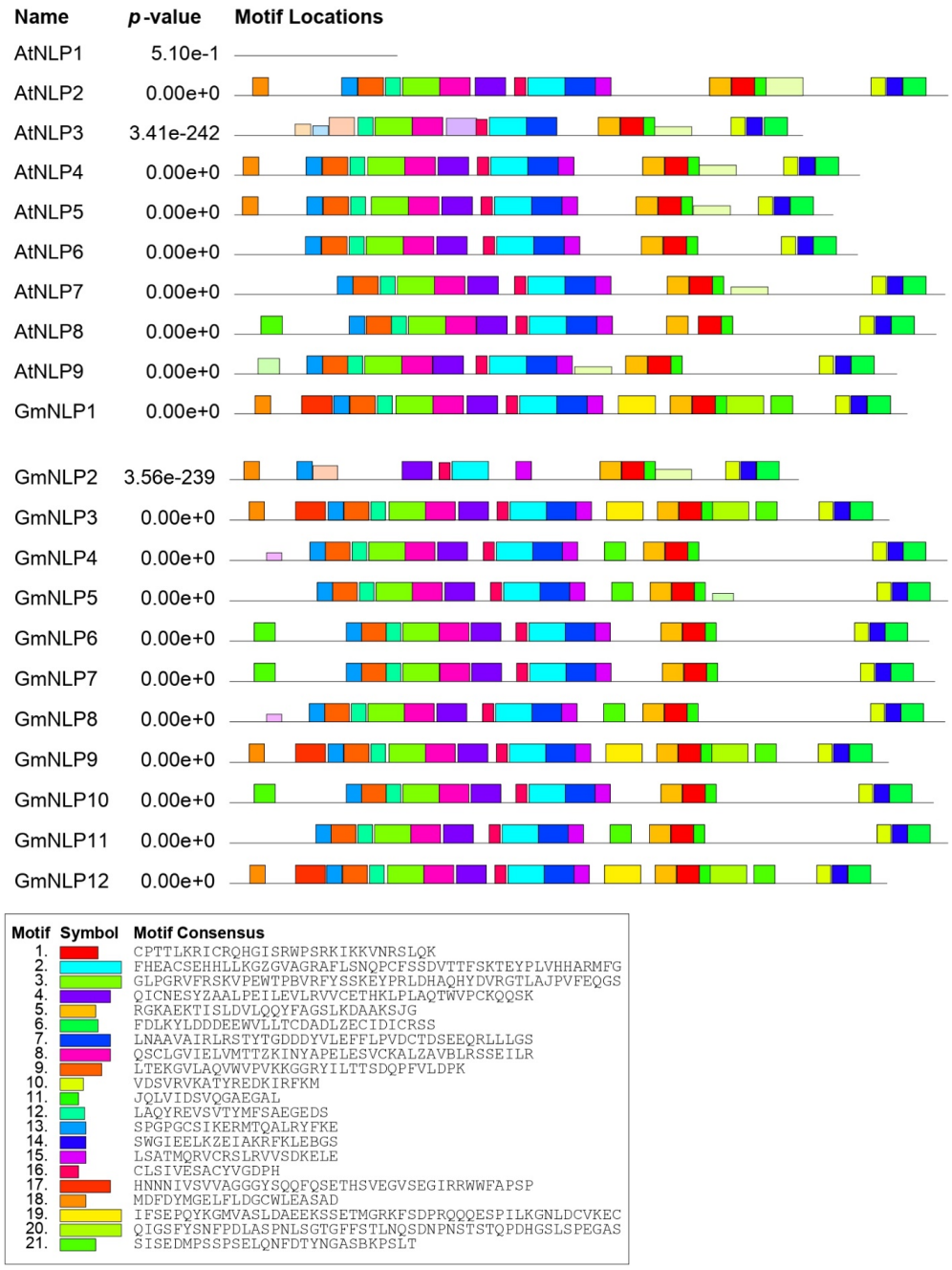


FIGURE 4.24: *AtNLPs* and *GmNLPs* details of motif

The above image number 4.23 illustrates the conserved protein motifs within the *NLP* transcription factor family in *Arabidopsis thaliana* and *Glycine max*. Each colored block on the horizontal lines signifies a specific sequence of amino acids that has been evolutionarily preserved, as indicated in the legend.

The p-value column confirms that the matching patterns are statistically significant and not due to random chance. Most proteins display a nearly identical

arrangement of these blocks, suggesting they are evolutionarily similar, with likely similar roles in nitrogen regulation within cells.

4.6 The Analysis of Phylogeny

The protein sequences of *NLP* genes in different strains were retrieved from different plant species like *Arabidopsis thaliana, Glycine max, Zea mays, Oryza sativa,* and *Brassica napus*. The sequences were downloaded in FASTA format and aligned in one file. Phylogenetic tree is designed with the help of neighbour-joining approach. CLUSTAL Omega was used to construct the tree and later on, the tree was modified with iTol. The interaction of *GmNLPs* with various species which have *NLP* genes was studied with the help of a phylogenetic tree (figure 4.1).

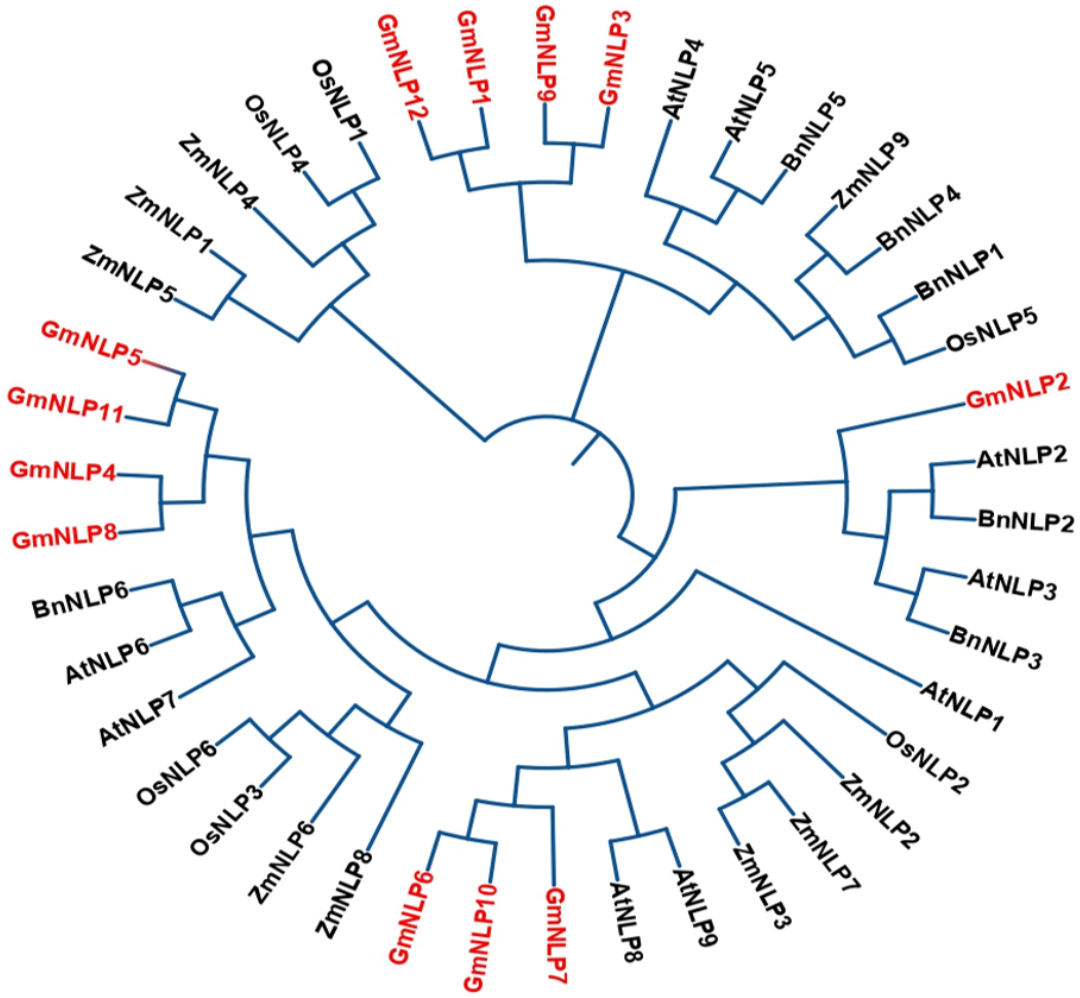


FIGURE 4.25: Phylogenetic analysis of *GmNLPs*

The phylogenetic tree illustrates the evolution of the *NLP* protein family across five plant species, with a focus on *Glycine max*, where its *GmNLP* proteins are highlighted in red. Moving outward along blue lines from the center reveals how a singular ancient ancestor gene diverged over millions of years. Groupings of red proteins indicate close genetic relationships, suggesting shared functions within the *GmNLPs*. Noteworthy findings include the grouping of *GmNLP4*, 5, 8, and 11, which relate to known nitrogen-regulating genes in *AtNLP6* and *AtNLP7*. Another grouping connects *Glycine max* genes to rice and corn, signifying a significant evolutionary relationship among different plant classifications. Contrarily, *GmNLP2* stands isolated on the right side of the tree, showing close association with *AtNLP2* and *AtNLP3*, indicating its early divergence and specialized function within the *Glycine max* species.

Chapter 5

Discussion

The growth and development of the plants are greatly affected by transcription factors under abiotic and biotic stress conditions. One family of plant-specific transcription factors is the NODULE-INCEPTION-Like Proteins. Prior research has demonstrated that *NLPs* play a significant role in nitrogen acquisition, assimilation, and transportation, regulated by the availability of nitrogen [57]. It has already been determined that nitrogen availability does not result in the activation of *NLPs*, but *NLPs* are responsible for activating nitrogen-responsive genes by *NLP* localization via a nuclear-retention mechanism in the presence of nitrogen. As a result of higher levels of *NLP* proteins due to an increased level of nitrogen in the environment, plants can take up more nitrogen as the expression of nitrogen-responsive genes becomes higher. However, *NLPs* have not been studied in horticultural plants, although many efforts have been made regarding the characterization of *NLPs* in various vascular plants [58].

Glycine max, also known as soybean, is a vascular plant that works on the same phenomena. The chance to analyze *GmNLPs* in this research study would have been possible due to the complete *Glycine max* genome being annotated in 2018. Genome-wide studies have proven to be quite useful in mining the genomic database in order to identify the structure and function of a gene along with its preliminary features, despite the fact that it does not prove any molecular mechanism

taking place within the cells. Past genome-wide studies have proven to be valuable due to future studies carried out on those genes [59].

The study on the *Glycine max* genome identified twelve distinct NIN-like protein transcription factors, an increase from the nine found in *Arabidopsis thaliana*. This variation illustrates the complex genomic evolution in plants and indicates whole-genome duplication events that drive functional specialization in economic crops. Validation shows that all twelve *GmNLPs* possess RWP-RK and PB1 domains, confirming their role as master regulators of the primary nitrate response. The RWP-RK motif facilitates specific interactions with nitrate-responsive elements, while the PB1 domain enables protein-protein interactions essential for transcription complex formation and gene activation. The *GmNLP* protein family exhibits hydrophilicity with negative GRAVY values and is primarily acidic, indicating electrostatic conservation for nuclear interactions. Variations in gene length underscore differences in intron retention and non-coding structures, while conserved exon-intron boundaries suggest evolutionary continuity [60].

Glycine max genes are notably larger than those in *Arabidopsis thaliana*, reflecting post-duplication diversification. MEME motif profiling reveals shared motifs among species, with unique motifs in certain soybean variants, indicating increased structural complexity in *Glycine max* for nutrient assimilation and signaling. Phylogenetic analysis revealed evolutionary insights into regulatory sequences across different plant lineages, particularly highlighting the proximity of *GmNLP4*, 5, 8, and 11 to *AtNLP6* and *AtNLP7*, indicating conserved roles in nitrogen regulation. *AtNLP7* in *Arabidopsis thaliana* serves as a master regulator for the nitrate response through a calcium-dependent mechanism. The evolutionary link suggests similar functions for soybean genes in nitrogen assimilation. In contrast, *GmNLP2*'s separate phylogenetic position near *AtNLP2* and *AtNLP3* suggests distinct roles, possibly in carbon-nitrogen balance or nodule symbiosis. This diversity in the *GmNLP* family indicates its potential as genetic targets for improving soybean nitrogen use efficiency and promoting sustainable agriculture [61].

An evolutionary relationship can be identified from the characteristics of the *NLP* genes and their protein sequences. It is evident from this study that there is an

evolutionary divergence between *NLP* gene families as indicated by the consensus protein motif. The proteins of *Arabidopsis thaliana* and *Glycine max* have protein motifs involved in the growth and development of plants. From this study, it has been concluded that there is more phylogenetic divergence and ancestral linkages in different plants' *NLPs* due to conserved domains RWP-RK and PB1. Plant growth and development are significantly influenced by the *Glycine max NLP* gene, particularly influenced by nitrogen availability [62]. *In silico* analyses suggest that while all *GmNLPs* are primarily involved in these mechanisms, their responses to stress and phytohormones may be secondary. Moreover, the predicted interacting proteins may have essential roles in nitrogen uptake, transport, and assimilation. Previous studies in rice indicate that *NLP* genes are responsive to nitrogen and enhance nitrogen use efficiency. Further research using advanced molecular techniques is necessary to confirm these findings.

Chapter 6

Conclusion and Future Recommendations

In conclusion, this study offers an initial, genome-wide identification and classification of the *NLP* family of transcription factors in *Glycine max*, with the identification of 12 transcription factors that are unique to this family. The analysis showed that the identified members possessed the specific RWP-RK and PB1 motifs, thereby confirming that they belong to this group of nitrate-responsive transcription factors. The physicochemical analysis of these transcription factors showed that the predicted protein structures were hydrophilic, stable, and nuclear-localized. Protein-protein interaction and gene structure helped to analyze the structure and function of proposed family. Finally, the phylogenetic analysis of these transcription factors showed that they are evolutionarily related to known transcription factors in model plants, which confirms that the mechanism for regulating nitrogen responses is well conserved across these plant groups.

6.1 Future Recommendations

This research has provided an exhaustive genome-wide identification and characterization of the *NLP* transcription factor family in *Glycine max*, comprising 12 members unique to this functional category. Indeed, from the structural and

motif analyses, it was evident that all the identified members contain the distinctive RWP-RK and PB1 motifs that confirm their function as nitrate-responsive transcription factors. The physicochemical properties of the identified proteins were found to be stable, hydrophilic, and nuclear proteins. Additionally, the phylogenetic analysis of the *NLP* transcription factors has shown that the *Glycine max NLP* transcription factors are evolutionarily conserved in relation to known transcription factors in model plants. Moving forward, the following functional validation studies can be considered as the next steps to capitalize on the computational advances. Firstly, quantitative real-time PCR analysis of expression is required to determine the exact expression patterns of specific *GmNLP* genes in various plant tissues, such as roots, leaves, and nodules, under the conditions of changing nitrogen levels and stress. Apart from the gene expression studies, the application of advanced genome editing techniques, such as CRISPR/Cas9, will be necessary for the direct identification of the gene functions. Generation of knockout and activation mutations will make it possible to identify the effects of candidate *NLPs* on nitrogen uptake and assimilation. In conclusion, this information can be used in breeding programs to obtain elite nitrogen-efficient soybean lines.

Bibliography

[1] M. L. Sámano, K. Nanjareddy, and M. K. Arthikala, "Nin-like proteins (nlps) as crucial nitrate sensors: an overview of their roles in nitrogen signaling, symbiosis, abiotic stress, and beyond," *Physiol. Mol. Biol. Plants*, vol. 30, no. 7, pp. 1209–1223, 2024.

[2] A. Fathi, "Role of nitrogen (n) in plant growth, photosynthesis pigments, and n use efficiency: A," *Agrisost*, vol. 28, pp. 1–8, 2022.

[3] J. Y. Ye, W. H. Tian, and C. W. Jin, "Nitrogen in plants: from nutrition to the modulation of abiotic stress adaptation," *Stress Biol.*, vol. 2, no. 1, p. 4, 2022.

[4] D. Yan and E. Nambara, "Conserved and unique functions of nin-like proteins in nitrate sensing and signaling," *Plant Sci.*, vol. 336, p. 111842, 2023.

[5] G. Duville, "Studies of nlp transcription factors: Roles and mechanisms of nlp-dependent nitrate signaling pathway," Ph.D. dissertation, Université Paris-Saclay, n.d.

[6] S. Das, B. Kinjarapu Palaka, R. Kuiry, and S. R. Choudhury, "Insights into the interactions of rwp-rk and their targets: Role of serine and its conservation across species," *Biochem. Biophys. Res. Commun.*, vol. 763, p. 151750, 2025.

[7] K. G. Agyenim-Boateng *et al.*, "The nutritional composition of the vegetable soybean (maodou) and its potential in combatting malnutrition," *Front. Nutr.*, vol. 9, p. 1034115, 2023.

[8] M. A. Ibáñez, C. De Blas, L. Cámara, and M. G. Estimado, "Chemical composition, protein quality and nutritive value of commercial soybean meals produced from beans from different countries: A meta-analytical study," *Anim. Feed Sci. Technol.*, vol. 267, p. 114531, 2020.

[9] U. Singh, A. K. Choudhary, and S. Sharma, "Agricultural practices modulate the bacterial communities, and nitrogen cycling bacterial guild in rhizosphere: Field experiment with soybean," *J. Sci. Food Agric.*, vol. 101, no. 7, pp. 2687–2695, 2021.

[10] S. Bagale, "Nutrient management for soybean crops," *Int. J. Agron.*, vol. 2021, no. 1, p. 3304634, 2021.

[11] S. Thapa *et al.*, "Managing micronutrients for improving soil fertility, health, and soybean yield," *Sustainability*, vol. 13, no. 21, p. 11766, 2021.

[12] N. Shilpashree *et al.*, "Morphological characterization, variability and diversity among vegetable soybean (glycine max l.) genotypes," *Plants*, vol. 10, no. 4, p. 671, 2021.

[13] R. Mishra, M. K. Tripathi, R. S. Sikarwar, Y. Singh, and N. Tripathi, "Soybean (glycine max l. merrill): A multipurpose legume shaping our world," *Plant Cell Biotechnol. Mol. Biol.*, vol. 25, no. 3-4, pp. 17–37, 2024.

[14] S. A. Asad, M. A. Wahid, S. Farina, R. Ali, and F. Muhammad, "Soybean production in pakistan: experiences, challenges and prospects," *Int. J. Agric. Biol.*, vol. 24, no. 4, pp. 995–1005, 2020.

[15] M. A. Hussain, A. Wahab, U. Ahmad, and Z. Arshad, "Current scenario of oilseeds in pakistan," *Trends Biotech. Plant Sci.*, vol. 1, no. 1, pp. 29–36, 2023.

[16] V. Voora, C. Larrea, and S. Bermudez, "Global market report: soybeans," International Institute for Sustainable Development (IISD), Tech. Rep., 2020.

[17] J. G. dos Reis, E. Aktas, and S. T. Machado, "Soybean supply chains, markets, and global trade," in *Soybean Production Technology: Physiology, Production and Processing*. Springer Nature Singapore, 2025, pp. 429–446.

[18] Z. Shea, W. M. Singer, and B. Zhang, "Soybean production, versatility, and improvement," in *Legume Crops—Prospects, Production and Uses*. IntechOpen, 2020, p. 10.

[19] J. Y. Chua and S. Q. Liu, "Soy whey: More than just wastewater from tofu and soy protein isolate industry," *Trends Food Sci. Technol.*, vol. 91, pp. 24–32, 2019.

[20] G. Dukariya, S. Shah, G. Singh, and A. Kumar, "Soybean and its products: Nutritional and health benefits," *J. Nutr. Sci. Healthy Diet*, vol. 1, no. 2, pp. 22–29, 2020.

[21] I. S. Kim, C. H. Kim, and W. S. Yang, "Physiologically active molecules and functional properties of soybeans in human health—a current perspective," *Int. J. Mol. Sci.*, vol. 22, no. 8, p. 4054, 2021.

[22] P. Qin, T. Wang, and Y. Luo, "A review on plant-based proteins from soybean: Health benefits and soy product development," *J. Agric. Food Res.*, vol. 7, p. 100265, 2022.

[23] M. Staniak, E. Szpunar-Krok, and A. Kocira, "Responses of soybean to selected abiotic stresses—photoperiod, temperature and water," *Agriculture*, vol. 13, no. 1, p. 146, 2023.

[24] E. Wójcik-Gront, D. Gozdowski, A. Derejko, and R. Pudełko, "Analysis of the impact of environmental and agronomic variables on agronomic parameters in soybean cultivation based on long-term data," *Plants*, vol. 11, no. 21, p. 2922, 2022.

[25] Q. Wang, S. Li, J. Li, and D. Huang, "The utilization and roles of nitrogen in plants," *Forests*, vol. 15, no. 7, p. 1191, 2024.

[26] S. K. Kakraliya *et al.*, "Nitrogen and legumes: a meta-analysis," in *Legumes for Soil Health and Sustainable Management*. Springer Singapore, 2018, pp. 277–314.

[27] A. Raza *et al.*, "Nitrogen fixation of legumes: Biology and physiology," in *The Plant Family Fabaceae: Biology and Physiological Responses to Environmental Stresses*. Springer Singapore, 2020, pp. 43–74.

[28] R. S. Meena, A. Das, G. S. Yadav, and R. Lal, Eds., *Legumes for Soil Health and Sustainable Management*. Springer, 2018.

[29] M. Tegeder and C. Masclaux-Daubresse, "Source and sink mechanisms of nitrogen transport and use," *New Phytol.*, vol. 217, no. 1, pp. 35–53, 2018.

[30] Y. Y. Wang, Y. H. Cheng, K. E. Chen, and Y. F. Tsay, "Nitrate transport, signaling, and use efficiency," *Annu. Rev. Plant Biol.*, vol. 69, pp. 85–122, 2018.

[31] H. Wang *et al.*, "Phylogeny and gene expression of the complete nitrate transporter 1/peptide transporter family in triticum aestivum," *J. Exp. Bot.*, vol. 71, no. 15, pp. 4531–4546, 2020.

[32] B. Qiang *et al.*, "Effects of nitrogen application levels on soybean photosynthetic performance and yield: Insights from canopy nitrogen allocation studies," *Field Crops Res.*, vol. 326, p. 109871, 2025.

[33] N. C. La Menza *et al.*, "Nitrogen limitation in high-yield soybean: Seed yield, n accumulation, and n-use efficiency," *Field Crops Res.*, vol. 237, pp. 74–81, 2019.

[34] L. Wang *et al.*, "Assessing the impact of biochar and nitrogen application on yield, water-nitrogen use efficiency and quality of intercropped maize and soybean," *Front. Plant Sci.*, vol. 14, p. 1171547, 2023.

[35] M. L. Sámano, K. Nanjareddy, and M. K. Arthikala, "Nin-like proteins (nlps) as crucial nitrate sensors: an overview of their roles in nitrogen signaling, symbiosis, abiotic stress, and beyond," *Physiol. Mol. Biol. Plants*, vol. 30, no. 7, pp. 1209–1223, 2024.

[36] Y. Ueda and S. Yanagisawa, "Transcription factor module nlp–nigt1 fine-tunes nitrate transporter2.1 expression," *Plant Physiol.*, vol. 193, no. 4, pp. 2865–2879, 2023.

[37] F. Liu *et al.*, "The long noncoding rna t5120 regulates nitrate response and assimilation in arabidopsis," *New Phytol.*, vol. 224, no. 1, pp. 117–131, 2019.

[38] M. Konishi and S. Yanagisawa, "The role of protein-protein interactions mediated by the pb1 domain of nlp transcription factors in nitrate-inducible gene expression," *BMC Plant Biol.*, vol. 19, no. 1, p. 90, 2019.

[39] X. Mu and J. Luo, "Evolutionary analyses of nin-like proteins in plants and their roles in nitrate signaling," *Cell. Mol. Life Sci.*, vol. 76, no. 19, pp. 3753–3764, 2019.

[40] Z. Wu *et al.*, "Genome-wide identification, characterization, and regulation of rwp-rk gene family in the nitrogen-fixing clade," *Plants*, vol. 9, no. 9, p. 1178, 2020.

[41] S. Mishra, G. Singroha, R. Tiwari, and P. Sharma, "Comprehensive genome-wide expression analysis of nlp transcription factors elucidates their crucial role in enhancing nitrogen response in wheat (triticum aestivum l.)," *Curr. Plant Biol.*, vol. 42, p. 100463, 2025.

[42] S. Bollam *et al.*, "Genome-wide association study and rna sequencing identify candidate genes regulating nitrogen use efficiency and associated traits in sorghum (sorghum bicolor (l.) moench)," *Food Energy Secur.*, vol. 15, no. 1, p. e70174, 2026.

[43] A. L. Van de Weyer, "The pan-nlr'ome of arabidopsis thaliana," Ph.D. dissertation, Universität Tübingen, 2019.

[44] M. Liu *et al.*, "Genome-wide identification and characterization of nodule-inception-like protein (nlp) family genes in brassica napus," *Int. J. Mol. Sci.*, vol. 19, no. 8, p. 2270, 2018.

[45] J. Pan *et al.*, "Prediction of protein–protein interactions in arabidopsis, maize, and rice by combining deep neural network with discrete hilbert transform," *Front. Genet.*, vol. 12, p. 745228, 2021.

[46] M. Ge *et al.*, "Genome-wide analysis of maize nlp transcription factor family revealed the roles in nitrogen response," *Plant Growth Regul.*, vol. 84, no. 1, pp. 95–105, 2018.

[47] A. Manta *et al.*, "Soybean databases: A means to explore soybean information," in *Soybeans 2026*. Apple Academic Press, 2026, pp. 87–113.

[48] L. Reiser, S. Subramaniam, P. Zhang, and T. Berardini, "Using the arabidopsis information resource (tair) to find information about arabidopsis genes," *Current Protocols*, vol. 2, no. 10, p. e574, 2022.

[49] B. Xue and S. Y. Rhee, "Status of genome function annotation in model organisms and crops," *Plant Direct*, vol. 7, no. 7, p. e499, 2023.

[50] T. Paysan-Lafosse *et al.*, "The pfam protein families database: embracing ai/ml," *Nucleic Acids Res.*, vol. 53, no. D1, pp. D523–D534, 2025.

[51] D. C. Mishra *et al.*, "Enhancing the classification of biosynthetic gene clusters through comprehensive nlp-based approach," *Bioinformatics Insights*, 2024.

[52] D. Kashyap and A. Khan, "In-silico subcellular localization and functional analysis of computationally predicted virulent proteins in x. oryzae pv. oryzae strain pxo99a causal organism of bacterial leaf blight (blb) of rice (o. sativa l.)," *Mol. Biol.: Open Access*, vol. 12, no. 3, p. 1000354, 2023.

[53] G. H. Ding, G. H. Xiao, and L. P. Zhu, "Genome-wide identification and expression analysis of nlp (nin-like protein) transcription factor gene family in cotton," *J. Plant Biol. Res.*, vol. 15, no. 2, pp. 112–126, 2024.

[54] D. Szklarczyk *et al.*, "The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D638–D646, 2023.

[55] S. I. Elhajjajy and Z. Weng, "A novel nlp-based method and algorithm to discover rna-binding protein (rbp) motifs, contexts, binding preferences, and interactions," *RNA*, vol. 32, no. 6, pp. 885–914, 2026.

[56] D. Pal, P. Ghosh, S. Dey, and S. Das, "Frequency rank vector based phylogeny: A new alignment-free protein sequence comparison," in *Proc. 2025 AI-Driven Smart Healthcare for Society 5.0*, 2025, pp. 60–65.

[57] N. Kumar and M. S. Mukhtar, "An nlp-based method to mine gene and function relationships from published articles," *Sci. Rep.*, vol. 15, no. 1, p. 7503, 2025.

[58] G. Yuan *et al.*, "Genome-wide identification and expression analysis of nin-like protein (nlp) genes reveals their potential roles in the response to nitrate signaling in watermelon," *Hortic. Plant J.*, vol. 8, no. 5, pp. 602–614, 2022.

[59] R. Mishra, M. K. Tripathi, R. S. Sikarwar, Y. Singh, and N. Tripathi, "Soybean (glycine max l. merrill): A multipurpose legume shaping our world," *Plant Cell Biotechnol. Mol. Biol.*, vol. 25, no. 3-4, pp. 17–37, 2024.

[60] H. Wang *et al.*, "Genome-wide analysis of nlp genes in peanut reveals significant roles of ahnina and ahninb in root nodule development," *Agronomy*, vol. 15, no. 10, p. 2377, 2025.

[61] S. Maheen *et al.*, "Genome-wide identification and characterization of nodule inception-like protein (nlp) gene family in mungbean (vigna radiata l.)," *SABRAO J. Breed. Genet.*, vol. 57, no. 5, pp. 1789–1798, 2025.

[62] A. B. Sultan *et al.*, "Divergent evolution of nlr genes in the genus glycine: impacts of annuals and perennials' life history strategies," *Front. Plant Sci.*, vol. 15, p. 1383135, 2024.