***Original Research Article***

**Unravelling the Evolutionary Blueprint of *Phytochrome A4* in Legumes: A Molecular Phylogenetic Approach**

**ABSTRACT**

**Background**: Phytochromes are the best characterized photoreceptors that perceive Red (R)/Far-Red (FR) signals and mediate key developmental responses in plants. It is well established that photoperiodic control of flowering is regulated by *PHY A* (*phytochrome A*) gene. So far, the members of *PHY A4* gene family remains unexplored in *Glycine max* and therefore, their functions are still not deciphered. *Phytochrome A4* (*PhyA4*) has involvement in adaptive responses under low-light and stress conditions. The present study is the first effort to identify any photoreceptor gene (*PHYA4*) in *Glycine max* and decipher its phylogeny with related legumes.

**Methods:** In present study, the nucleotide sequence of Glycine max *PhyA4* (*GmPhyA4*) was used to identify homologous sequences within the *Fabaceae* family through BLASTn analysis. As well as for protein identify homologous sequences. A total of 17 *Fabaceae* *PhyA4* homologs, along with Arabidopsis thaliana *PHYA* as an outgroup, were aligned using the CLUSTAL W algorithm in MEGA 11.

**Result:** The Tamura 3-parameter model with Gamma distribution (T92+G) was identified as the best-fit nucleotide substitution model, based on the lowest BIC and AIC values. For protein JTT (Jones–Taylor–Thornton) substitution model found to best model. Phylogenetic reconstruction was performed using the Maximum Likelihood (ML) method reliability assessment using 1,000 bootstrap replicates.

**Conclusion:** Phylogenetic analysis confirmed its close evolutionary relationship with other *Fabaceae* species. This phylogenetic insight underscores the evolutionary conservation and diversification of *PhyA4* within legumes, providing a foundation for understanding its functional adaptation in different species and potential applications in crop improvement under varying photoperiodic conditions.

**Keywords:** ***Phytochrome A4***, **BLASTn, BLASTp ClustalW, Substitution model and Phylogenetic analysis**

**INTRODUCTION**

Legume crops belonging to the *Fabaceae* (*Leguminosae*) family are a vital group of plants cultivated globally for food, fodder, oil and soil fertility enhancement. Their unique ability to fix atmospheric nitrogen through symbiotic associations with *Rhizobium* spp. in root nodules makes them indispensable for sustainable agriculture, enhancing soil fertility and reducing dependency on synthetic nitrogen fertilizers (Graham & Vance, 2003; Herridge *et al*., 2008). In India, legumes are widely grown across diverse agro-climatic zones, playing a critical role in food and nutritional security. With the stagnating yield trends of conventional staple crops and rising concerns about environmental sustainability, climate-resilient crops such as legumes have gained increasing attention (Foyer *et al*., 2016). Legumes exhibit significant genetic diversity and adaptive traits, making them attractive candidates for genomic and proteomic research targeting abiotic stress tolerance (Varshney *et al*., 2013). Phylogenetic analysis, a cornerstone of evolutionary biology, aids in elucidating the genetic relationships among species and in identifying evolutionarily conserved and stress-resilient genotypes (Wendel & Doyle, 2005). When combined with high-throughput genomics, this evolutionary insight supports molecular breeding programs aimed at crop improvement (Varshney *et al.*, 2005). Phylogenetic trees originally conceptualized by Darwin and now formalized using molecular data are powerful tools to trace evolutionary trajectories, infer gene flow and predict functional conservation (Gregory, 2008; Weyenberg & Yoshida, 2016).

The present study focuses on understanding the evolutionary relationships and structural characteristics of the *Phytochrome A4* (*PhyA4*) gene in legumes as shown by Rockwell *et al*. (2006) and Franklin and Quail (2010), PhyA4 plays an essential role in mediating photomorphogenic responses and circadian regulation. In agreement with these findings, our phylogenetic analysis revealed high conservation of PhyA4 among legumes, supporting its functional stability across species. Protein modeling also support the presence of conserved structural domains critical for its function. Further, through protein modeling, conserved functional domains and motifs were identified, contributing to a deeper understanding of *PhyA4* is role in plant signaling and aiding targeted breeding efforts.

Here, attached supportive reference for phylogeny analysis carried out.Wickland and Hanzawa (2015) studied the functional evolution and molecular mechanisms of  *FLOWERING LOCUS T/TERMINAL FLOWER 1* gene family. Phylogenetic analysis of the *FT/ TFL1* protein family in 50 plant species was carried out. The analysis indicated that *FT* and *TFL* homologs were clustered in two major distinct groups, in line with their antagonistic function as flowering inducer and repressor, respectively. Flores *et al.* (2018) studied comparative phylogenetic and expression analysis of small GTPases families in legume and non-legume plants. The analyses suggest that the number of family members and the primary sequence of small GTPases are well conserved between legume and non-legume plants. Krishna *et al.* (2022) studied the phylogenetic analysis of the *phytochrome A* gene from *Lablab purpureus* (L.) sweet. Phytochromes are the best-characterized photoreceptors that perceive Red (R)/Far-Red (FR) signals and mediate key developmental responses in plants. The analysis showed that this *phytochrome* gene evolved from a common ancestry root but diverged into different clades during evolution. The PHYA protein sequences from *Lablab purpureus, Vigna unguiculata* and *Glycine max* formed independent clade and were closest to *Vigna angularis* and *Cajanus cajan*.Abbas *et al*. (2023) assessed genetic variability and evolutionary relationships of *Rhizoctonia solani* isolates in legume crops. The phylogenetic analysis of *R. solani* isolates across different legumes indicated that the distinct clades or subclades formed by the isolates correspond to their specific anastomosis groups (AGs) and subgroups, rather than being determined by their host legume crop.

**MATERIALS AND METHODS**

**Data Acquisition**

Collecting accurate data is the first and most important step in any bioinformatics study. Selecting the correct sequences is a critical step, as it is essential to obtain complete sequence rather than partial sequence data. Sequence data for legume species were collected from the National Center for Biotechnology Information (NCBI) and retrieved in FASTA format (Anon., 2025). In the present study, flowering-related Phytochrome A4 gene were selected.

**Statistical Tools for Analysis**

 **Statistical analyses** were performed using validated computational tools, ensuring accurate interpretation of biological data and consistency in results.

**Analysis Process of Build Phylogenetic Tree**

The present study carried out phylogenetic analysis was conducted using Molecular Evolutionary Genetics Analysis version 11 (MEGA11) software (Hall, 2013). The following step-wise procedure was followed:

1. ***Data collection***

The first step in constructing the phylogenetic tree involved developing a high-quality dataset comprising DNA and protein sequences retrieved from GenBank (NCBI) as a primary source in FASTA format.

1. ***Homology search***

The retrieved sequences were used as queries in BLASTn (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) to identify homologous sequences within the NCBI GenBank database. Sequences from the *Fabaceae* family showing significant similarity were selected for further analysis.

1. ***Sequence curation***

Selected sequences were screened to ensure completeness and quality. Only full-length or near full-length sequences were retained.

1. ***Multiple Sequence Alignment (MSA)***

The curated sequences, including Arabidopsis thaliana as an outgroup, were aligned using the CLUSTAL W algorithm in MEGA11.

1. ***Model selection***

The best-fit nucleotide or protein substitution model was identified using MEGA11 go to **Models and select Find Best DNA/Protein Models.** Select the alignment file and run the model test. The model with the lowest Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) value was selected.

1. ***Phylogenetic tree construction***

The phylogenetic tree was constructed using the selected substitution model. Choose **Phylogeny in which select Construct/Test Neighbor-Joining (NJ), Maximum Likelihood (ML) or UPGMA Tree**,depending on the method selected. Load the aligned sequence file. Set **bootstrap replications** to 1000 for statistical support. Generate the tree and export it in image format for documentation (Felsenstein, 1985).

Phylogenetic tree construction methods were broadly categorized into two types: distance matrix methods (also known as clustering or algorithmic methods) and character-based methods (also referred to as discrete data methods). The most appropriate method among these was selected for constructing the phylogenetic tree.

 **RESULT AND DISCUSSION**

Retrived dataset from NCBI Genebank

Homology search using BLASTn for selected nucleotide

Multiple Sequence Alignment for assess sequence similarity

**Substitution model selection** based on BIC and AIC

**Phylogenetic Tree Construction using various methods**

**Fig. 1: Flow chart of Phylogeny tree construction**

**Sequence retrieval, alignment and model selection for phylogenetic analysis**

The processed *Glycine max Phytochrome A4* (GmPhyA4) nucleotide sequence was used as a query in BLASTn (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) to identify homologous sequences in the NCBI nucleotide database. Homologous sequences from species within the *Fabaceae* family were retrieved. Multiple sequence alignment of 17 *Fabaceae PhyA4* species, along with Arabidopsis thaliana, was performed using the CLUSTAL W algorithm in MEGA 11 (Molecular Evolutionary Genetics Analysis). The optimal substitution model was determined using the "Find Best DNA/Protein Models" function in MEGA 11. Nucleotide substitutions were analyzed using sequences with complete deletion of gaps or missing data.

**Table 1: Nucleotide substitution model of *PhyA4***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **BIC** | **AIC** | **Gamma** | **Freq A** | **Freq T** | **Freq C** | **Freq G** |
| **T92+G** | 66116.14 | 65783.66 | 0.49 | 0.29 | 0.29 | 0.21 | 0.21 |
| **T92+G+I** | 66127.38 | 65785.67 | 0.49 | 0.29 | 0.29 | 0.21 | 0.21 |
| **HKY+G** | 66137.26 | 65786.31 | 0.49 | 0.27 | 0.30 | 0.21 | 0.21 |
| **TN93+G** | 66147.46 | 65787.28 | 0.49 | 0.27 | 0.30 | 0.21 | 0.21 |
| **GTR+G** | 66147.97 | 65760.08 | 0.50 | 0.27 | 0.30 | 0.21 | 0.21 |
| **HKY+G+I** | 66148.50 | 65788.32 | 0.49 | 0.27 | 0.30 | 0.21 | 0.21 |
| **TN93+G+I** | 66158.70 | 65789.28 | 0.49 | 0.27 | 0.30 | 0.21 | 0.21 |
| **GTR+G+I** | 66159.21 | 65762.08 | 0.50 | 0.27 | 0.30 | 0.21 | 0.21 |
| **K2+G** | 66496.86 | 66173.62 | 0.51 | 0.25 | 0.25 | 0.25 | 0.25 |

The results of nucleotide substitution model selection for the PhyA4 gene, evaluated based on Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) values presented in Table 1. These criteria help identify the most appropriate evolutionary model by balancing model complexity with goodness-of-fit. Lower BIC and AIC scores indicate a better-fitting model.

Among the tested models, T92+G (Tamura 3-parameter model with Gamma distribution) exhibited the lowest BIC (66116.14) and AIC (65783.66) values, indicating it as the best-fitting model for the PhyA4 sequence data. This model accounts for unequal nucleotide frequencies and variation in substitution rates among sites, making it suitable for capturing the evolutionary dynamics of the gene. Gamma (Γ) values, which represent the shape parameter of the Gamma distribution used to model rate heterogeneity among sites. Most models yielded a Gamma value of approximately 0.49–0.51, suggesting moderate rate variation across sites in the PhyA4 sequence.

Additionally, nucleotide frequencies (A, T, C and G) were calculated for each model. The observed base composition was relatively balanced in most models, with adenine (A) and thymine (T) frequencies generally around 0.27–0.30 and cytosine (C) and guanine (G) frequencies around 0.21. Notably, the K2+G model shows a more balanced nucleotide distribution (0.25 for each base), but it performed poorly in terms of BIC and AIC, indicating a less accurate fit to the data. Overall, the T92+G model was selected for downstream phylogenetic analysis due to its optimal balance of statistical support and biological relevance.

**Table 2: Substitution matrix using Tamura 3 parameter model of *PhyA4***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **From\To** | **A** | **T** | **C** | **G** |
| **A** | - | 0.0586 | 0.0432 | 0.1258 |
| **T** | 0.0586 | - | 0.1258 | 0.0432 |
| **C** | 0.0586 | 0.1707 | - | 0.0432 |
| **G** | 0.1707 | 0.0586 | 0.0432 | - |

The substitution matrix shows the rates at which one nucleotide changes to another nucleotide represented in Table 2. Tamura 3-parameter model is a specific evolutionary model used in genetics to describe the rate of nucleotide substitutions. It takes into account differences in rates between transitions (purine to purine or pyrimidine to pyrimidine) and transversions (purine to pyrimidine or vice versa). The numbers represent the probability of one nucleotide changing to another within a given time period. Higher numbers indicate a higher rate of substitution**.**

In the substitution rate matrix derived from the Tamura 3-parameter model, the value 0.0586 reflects a relatively low rate of nucleotide substitution, corresponding to transitions such as A→T, T→A, C→A, and G→T. An even lower substitution rate of 0.0432 was observed for A→C, T→G, C→G, and G→C. In contrast, the rate of 0.1258 indicates a higher frequency of change, particularly for A→G and T→C. The highest substitution rate, 0.1707, occurs for C→T and G→A. These values represent the relative frequencies at which one nucleotide is expected to be replaced by another over evolutionary time, under the assumptions of the Tamura 3-parameter model.

The phylogenetic tree was constructed using the Maximum Likelihood (ML) method based on the Tamura 3-Parameter model. An initial tree was generated using the Neighbor-Joining (NJ) method with default parameters, followed by heuristic refinement using the Nearest-Neighbor Interchange (NNI) algorithm. Assess the reliability of the tree topology, bootstrap analysis with 1,000 replicates was performed using the Poisson correction model. The best-scoring ML tree was selected to represent the evolutionary relationships among the 18 analyzed specimens.

A phylogenetic tree visually represents evolutionary relationships among organisms. In present study, GmPhyA4 was compared with PhyA4 gene sequences of various plant species, primarily within the *Fabaceae* family. The resulting tree showed distinct clades, indicating divergence from a common ancestral gene through evolutionary changes. Closely related species clustered together, reflecting shared evolutionary history.

**Fig. 2: Phylogenetic tree of the *PhyA4*  nucleotide sequences of 17 different taxa (The tree was constructed using the maximum likelihood method with 1000 bootstrap replications)**

Fig. 2 *Glycine max PhyA4* and *Glycine soja PhyA4* were very closely related, with strong bootstrap support (100). This indicated a high degree of confidence in their recent common ancestry based on the *PhyA4* gene. *Arachis duranensis PhyA4* and *Arachis hypogaea PhyA4* were also very closely related with strong support (100). *Phaseolus vulgaris PhyA4* (common bean) and *Vigna radiata PhyA4* (mung bean) form a strongly supported group (100). *Medicago truncatula PhyA4* and *Trifolium pratense PhyA4* were grouped together with good support (99). *Abrus precatorius PhyA4* and *Cajanus cajan PhyA4* form a group with moderate support (93). *Gastrolobium bilobum PhyA4* and *Arabidopsis thaliana PHYA* were the most distantly related to the other species shown in this tree based on this analysis of the *PhyA*4 gene.

**Phylogeny Analysis of Phytochrome A4 (PhyA4) protein in Legumes**

 In present study, the tree of phylogeny analysis focuses on the Phytochrome A4 (PhyA4) protein. The processed *Glycine max* Phytochrome A4(GmPhyA4) protein sequence was used as a query in BLASTp (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) to identify homologous sequences in the NCBI protein database. Homologous sequences from species within the *Fabaceae* family were retrieved. Multiple sequence alignment of 19 *Fabaceae PhyA4* species, was performed using the CLUSTAL W algorithm in MEGA 11 (Molecular Evolutionary Genetics Analysis). The optimal substitution model was determined using the "Find Best DNA/Protein Models" function in MEGA 11. Nucleotide substitutions were analyzed using sequences with complete deletion of gaps or missing data.

**Table 3: Substitution model of *PhyA4* protein**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **BIC** | **AIC** | **Gamma** |
| JTT+G | 15208.44 | 14919.97 | 0.27 |
| JTT+G+I | 15218.46 | 14921.97 | 0.27 |
| JTT+G+F | 15334.93 | 14894.30 | 0.27 |
| JTT+I | 15335.68 | 15047.20 | n/a |
| JTT+G+I+F | 15344.95 | 14896.31 | 0.27 |
| WAG+G | 15418.34 | 15129.86 | 0.27 |
| cpREV+G | 15428.04 | 15139.56 | 0.27 |
| WAG+G+I | 15428.35 | 15131.87 | 0.27 |
| cpREV+G+I | 15438.12 | 15141.63 | 0.27 |

The results of substitution model selection for the PhyA4 gene, evaluated based on Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) values presented in Table 3. These criteria help identify the most appropriate evolutionary model by balancing model complexity with goodness-of-fit. Lower BIC and AIC scores indicate a better-fitting model.

Among the tested models, JTT+G (Jones-Taylor-Thornton model with Gamma distribution) exhibited the lowest BIC (15208.44) and AIC (14919.97) values, indicating it as the best-fitting model for the PhyA4 sequence data. The model accounts for amino acid substitution probabilities based on empirical data, providing an accurate representation of evolutionary relationships. Bootstrap analysis with 1000 replicates was conducted to assess the reliability of the inferred phylogeny

The phylogenetic tree was constructed using the Maximum Likelihood (ML) method based on the Jones-Taylor-Thornton model. The best-scoring ML tree was selected to represent the evolutionary relationships among the 19 analyzed specimens.



**Fig. 3: Phylogenetic tree of the PhyA4 protein sequences of 17 different taxa (The tree was constructed using the maximum likelihood method with 1000 bootstrap replications)**

The phylogenetic analysis of Phytochrome A4 (PhyA4) protein sequences across 22 leguminous species reveals distinct evolutionary relationships and conserved clustering patterns. The tree was constructed using the Maximum Likelihood method based on the JTT (Jones–Taylor–Thornton) substitution model, with bootstrap support values indicated at each node. The species were grouped into well-supported clades, suggesting a high degree of sequence conservation within certain genera. Notably, *Vigna umbellata*, *Vigna angularis*, *Vigna radiate* and *Vigna unguiculata* formed a robust clade with high bootstrap support (≥93), reflecting close evolutionary proximity and possibly conserved functional roles of PhyA4 in these species. Similarly, *Glycine max* and *Glycine soja* clustered together with 100 per cent bootstrap support, indicating strong genetic similarity and shared ancestry.

A moderately supported clade comprising *Cajanus cajan*, *Abrus precatorius* and *Lupinus angustifolius* suggests an intermediate divergence pattern within the group. The Arachis species *A. duranensis, A. ipaensis*, *A. stenosperma* and *A. hypogaea* also formed a distinct cluster, albeit with varying bootstrap support, indicating possible sub-genomic divergence or evolutionary adaptation within the genus. In contrast, *Gastroloibium bilobum* appeared basal to the Arachis group, suggesting an earlier divergence event. Additionally, species from the Galegoid clade, such as *Pisum sativum*, *Medicago truncatula*, *Cicer arietinum* and *Lotus japonicus*, formed a well-supported terminal clade, with bootstrap values of 100 per cent, reinforcing the conserved nature of PhyA4 within this evolutionary lineage.

Overall, the tree topology highlights both conserved and divergent evolutionary trajectories of PhyA4 among legume species. The clustering patterns corroborate known taxonomic relationships and provide insights into the functional conservation of PhyA4, particularly in regulating light-mediated developmental processes across diverse leguminous taxa.

**CONCLUSION**

The phylogenetic analyses of the *PhyA4* gene at both the protein and nucleotide levels revealed consistent and biologically meaningful clustering among leguminous species, highlighting both the evolutionary conservation and divergence of this important photoreceptor. The nucleotide -based phylogeny, constructed using the Maximum Likelihood method with the Tamura 3 parameter substitution model, demonstrated the formation of robust clades with high bootstrap support. Closely related species such as *Glycine max* and *Glycine soja*, along with various *Vigna* species, clustered together, suggesting strong functional conservation of the *PhyA4* gene within these groups. Additionally*, Phaseolus vulgaris* and *Vigna radiata* formed a strongly supported clade, indicating a high degree of evolutionary relatedness in their PhyA4 sequences. The protein -based phylogenetic analysis constructed using the Maximum Likelihood method with the JTT substitution model supported these findings, further confirming the recent common ancestry of *Glycine max* and *Glycine soja* (100% bootstrap support), as well as the close relationship between *Arachis duranensis* and *Arachis hypogaea*. Collectively, these results provide compelling evidence of lineage-specific diversification and conserved functional evolution of the PhyA4 gene across legumes, offering valuable insights into the mechanisms of phytochrome-mediated light signaling and adaptation in plant evolution.

**REFERENCES**

Abbas, A.; Ali, A.; Hussain, A.; Ali, A.; Alrefaei, A. F.; Naqvi, S. A. H. and Baloch, F. S. (2023). Assessment of Genetic Variability and Evolutionary Relationships of Rhizoctonia solani Inherent in Legume Crops. *Plants,* **12**(13): 2515.

Annonymous, (2025). NCBI (National Center for Biotechnology Information). Retrieved from https://www.ncbi.nlm.nih.gov/ [Accessed on 18 January, 2023].

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, **125**(1): 1-15.

Flores, A. C.; Via, V. D.; Savy, V.; Villagra, U. M.; Zanetti, M. E. and Blanco, F. (2018). Comparative phylogenetic and expression analysis of small GTPases families in legume and non-legume plants. *Plant signaling and behavior*, **13**(2): e1432956.

Foyer, C.H.; Lam, H.M.; Nguyen, H.T.; Siddique, K.H.; Varshney, R.K.; Colmer, T.D.; Cowling, W.; Bramley, H.; Mori, T.A.; Hodgson, J.M. and Cooper, J.W., 2016. Neglecting legumes has compromised human health and sustainable food production. *Nature plants*, ***2***(8):10.

Franklin, K. A. and Quail, P. H. (2010). Phytochrome functions in Arabidopsis development. *Journal of experimental botany*, **61**(1): 11-24.

Graham, P. H. and Vance, C. P. (2003). Legumes: Importance and constraints to greater use. *Plant Physiology*, **131**(3): 872–877.

Gregory, T. R. (2008). Understanding evolutionary trees. *Evolution: Education and Outreach*, **1**: 121-137.

Hall, B. G. (2013). Building phylogenetic trees from molecular data with MEGA*. Molecular biology and evolution*, **30**(5): 1229-1235.

Herridge, D. F.; Peoples, M. B. and Boddey, R. M. (2008). Global inputs of biological nitrogen fixation in agricultural systems*. Plant and Soil*, 311: 1–18.

Krishna, S.; Modha, K.; Parekh, V.; Patel, R. and Chauhan, D. (2022). Phylogenetic analysis of phytochrome A gene from Lablab purpureus (L.) Sweet. *Journal of Genetic Engineering and Biotechnology*, **20**(1): 9.

Rockwell, N. C., Su, Y. S., & Lagarias, J. C. (2006). Phytochrome structure and signaling mechanisms. *Annu. Rev. Plant Biol.*, **57**(1): 837-858.

Varshney, R. K., Roorkiwal, M. and Nguyen, T. (2013). Legume genomics: from genomic resources to molecular breeding. *The Plant Genome*, **6**(3): 1-7.

Varshney, R. K.; Graner, A. and Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends in plant science*, **10**(12): 621-630.

Wendel, J. F. and Doyle, J. J. (1998). Phylogenetic incongruence: window into genome history and molecular evolution. In *Molecular systematics of plants II: DNA sequencing* (pp. 265-296). Boston, MA: Springer US.

Weyenberg, G. and Yoshida, R. (2016). Phylogenetic Tree Distances. In: *“Encyclopedia of Evolutionary Biology”,* (R. M. Kaliman,Ed.)., *Academic Press*, 285-290.

Wickland, D. P. and Hanzawa, Y. (2015). The FLOWERING LOCUS T/TERMINAL FLOWER 1 gene family: functional evolution and molecular mechanisms. *Molecular plant*, **8**(7): 983-997.