In silico characterization of *nif*H gene of *Rhizobium* sp. TN04 isolated from the rhizosphere of non-leguminous potato plants

Tahir Naqqash[1, 2*], Syed Aun Muhammad[1], Syed Bilal Hussain[1], Muhammad Kashif Hanif[2, 3], Muhammad Arshad[4]

[1]*Institute of Molecular Biology and Biotechnology, Bahauddin Zakariya University, 60800 Multan, Pakistan.*

[2]*National Institute for Biotechnology and Genetic Engineering (NIBGE), P.O. Box 577, Jhang Road, Faisalabad, Pakistan.*

[3]*Department of Biological Sciences, University of Lahore, Sargodha Campus,89/2-A Zafar Ullah Rd, Shamsheer Town, Sargodha, Punjab 40100, Pakistan.*

[4]*Department of Biotechnology, University of Okara, Okara, Pakistan.*

[*]**Corresponding author email:** tahirnaqqash@gmail.com, tahirnaqqash@bzu.edu.pk

Abstract

The reliance on nitrogen (N) fertilizers in global crop production has led to significant environmental concerns and economic burdens due to their excessive usage. In order to effectively address this problem, a comprehensive understanding of biological N-fixation (BNF), governed by the *Nif* genes is essential. In legumes, the role of *Rhizobium* in BNF is well-established. However, limited studies are available regarding the function and structure of *nif* genes in non-leguminous plants using *in silico* modeling. Therefore, the present study was conducted to predict the structural and functional analysis of *nif*H gene from a *Rhizobium* strain isolated from potato plant. Various bioinformatics tools (ExPasy ProtParam, PSIPRED, MEMSAT-SVM, CATH classification, COFACTOR, COACH and STRING) were used to predict the primary, secondary and 3D structure of nifH protein. Results showed that TN04 has stable structure and hydrophobic nature similar to *Rhizobium* sp. S1SS148 and *R. rosettiformans*. Amino acid composition showed presence of different resides with glycine being most prevalent. Secondary structure analysis proved its stability due to the presence of coils, helices, and sheets. The nifH protein model derived from TN04 using I-TASSER displayed excellent structural characteristics, as confirmed by ERRAT. Functional annotations highlighted enzyme similarities and specific ligand-binding sites associated with nitrogenase activity. CATH categorization revealed the presence of a P-loop NTPase domain known to bind

nucleotides, which can affect the activity of nitrogenase. In addition, the investigation of protein-protein interactions using STRING suggested potential interactions between nifH protein of TN04 and several nif proteins, hinting at its possible involvement in N-fixation. The results of these studies shed light on possible N-fixation mechanisms in *Rhizobium* sp. TN04 in non-legumes. Based on these predictions, the results suggest the possible pathways for implementation of sustainable agricultural methods. However, further studies are necessary to validate these findings and investigate the role of *Rhizobium* sp. TN04 in N-fixation in non-leguminous plants, thus, enhancing understanding in this domain.

Keywords: *Rhizobium*, Potato, *nif*H, 3D structure, Secondary structure

Article History: Received: 04th December 2023, **Revised:** 20th December 2023, **Accepted:** 24th December 2023, **Published:** 30th December 2023.

Introduction

Nitrogen (N) availability significantly limits agricultural crop productivity worldwide. The utilisation of N fertiliser on a global scale is increasing significantly, with around 40% of the world's population depending on N fertiliser for crop cultivation [1]. The excessive use of N fertiliser not only causes high costs but also pose environmental issues. Biological nitrogen fixation (BNF) is a fundamental microbiological process within soil and plant ecosystem, vital for providing N to crops [2]. The process of N-fixation is regulated by *Nod*, *Fix* and *Nif* genes [3]. Among them, *Nif* genes show maximum diversity and encode proteins crucial in regulating the process of N-fixation [4]. The *nif* genes only work in microaerophilic or anaerobic environments because they are highly vulnerable to the presence of oxygen.

The N-fixing machinery in diazotroph organisms consists of 19 *nif* genes [5], responsible for converting N from an unusable form to a useful form using the enzyme known as nitrogenases. The structural subunits of nitrogenase enzymes encoded by *nif*D, *nif*K, and *nif*H genes are responsible for nitrogenase activity. The

proteins identified in diazotrophs such as *Bradyrhizobium japonicum*, *Herbaspirillum seropedicae*, *Azotobacter vinelandii*, and *Pseudomonas stutzeri* have common function and structure along with similar sequences [6-8]. The availability of *nif*H protein structural model is essential for investigating biological N-fixation activities on the molecular scale. However, there is limited knowledge regarding the function and structure of *nif*H proteins in potato plants.

It is well established that leguminous plants obtain nitrogen by establishing endosymbiotic relationships with rhizobia [9]. These bacteria fix N by forming nodules on the roots of their host plants and play a beneficial role in promoting the growth of these plants. *Rhizobium*, characterized as non-sporulating, Gram-stain-negative aerobic rods, belongs to α-proteobacteria and β-proteobacteria [10, 11]. This bacterium is distributed among 18 genera within various families and is usually called as legume endosymbionts. Moreover, they have also been identified in association with the roots of non-leguminous plants, including certain cereals like rice, wheat, and maize [12-14]. However, the isolation and identification of *Rhizobium* as a free living diazotroph in potato (*Solanum tuberosum* L.) plants remains relatively less explored.

Potato, a major vegetable crops cultivated in 79% of countries worldwide [15]. It ranks fourth global production, following wheat, and maize. It is known for its cost-effectiveness and rich nutritional profile (vital amino acids, proteins, minerals, antioxidants, vitamins, and carbohydrates) [16]. Various plant-growth promoting bacteria have been identified in potato plants, including *Bacillus*, *Aeromonas*, *Azospirillum*, *Morxella*, and *Pseudomonas*, among others [17-19]. However, limited studies are available that report the isolation of *Rhizobium* from non-leguminous potato plants. The study conducted by Naqqash et al., [20] reported the isolation and characterization of *Rhizobium* sp. TN04 from potato rhizosphere. from the results of their studyshowed increased levels of N under controlled and field conditions [20]. Thus, more detailed understanding of the pathway involved in *Rhizobium* N-fixation, particularly the role of the *nif*H gene responsible for its nitrogenase activity, is needed.

Therefore, the present study aimed to investigate the function of *nif*H gene in N-fixation by developing an *in silico* model. To gain a deeper understanding of N-

fixation, this study examined the protein structural variations and analysed the primary, secondary, and tertiary structures using *in silico* modelling techniques. However, the tertiary structures of numerous nitrogenase proteins from various diazotrophs, especially those of symbiotic organisms, have not been determined so far. Hence, constructing a model of the *nif*H tertiary structure is essential to gain a deeper insight into its activity.

Material and methods

The study conducted by Naqqash et al., [20] collected soil samples from rhizospheric region of potato plants located in Gujranwala, Pakistan. The isolated strain (*Rhizobium* sp. TN04) was fully characterized using morphological, biochemical and plant-growth promoting traits. TN04 was identified on the basis of its 16s rRNA gene sequence (Accession number: LN833444) and also exhibited the presence of the *nif*H gene (Accession No. LT596587). Through ARA activity, TN04 demonstrated its capacity for N-fixation (151.70 nmolmg/protein/h), suggesting its capability to convert atmospheric N into a usable form. Thus, this study aimed to investigate the N-fixing potential of TN04 using different *in silico* modelling techniques.

Sequence retrieval

A total of 21 *nif*H gene and protein sequences from various strains of Rhizobium, which exhibit nitrogen fixation activity, were obtained in FASTA file format from the NCBI database (http://www.ncbi.nlm.nih.gov/). All these sequences were utilized for subsequent *in silico* investigations [21].

Sequence alignment and phylogenetic analysis

The retrieve sequences were aligned using Clustal W, and two distinct phylogenetic trees were constructed using MEGA 11 software [22] analyzing both the gene and amino acid sequences. These trees were used to assess the evolutionary relationships among different strains of *Rhizobium*. The evolutionary history in both trees was deduced using the Neighbour-Joining approach [23]. In order to guarantee the accuracy of the tree structure, only bootstrap values exceeding 70% were considered [24].

Physicochemical characterization

After conducting phylogenetic analysis, the physiochemical properties of two strains (*Rhizobium* sp. S1SS148 and *Rhizobium*

rosettiformans W3), which showed maximum similarity with TN04 were assessed using the Expasy ProtParam tool (http://web.expasy.org/protparam/) as suggested by Gasteiger et al. [25]. These include molecular weight, number of amino acids (aa), aliphatic index (AI), isoelectric point (pI). The molecular weight was determined in ProtParam by adding the average isotopic masses of amino acids in protein with the average isotopic mass of H_2O molecule. The protein's pI was determined by utilising the pKa values of its aa based on their side-chain characteristics, which influence the protein properties under different pH conditions. The AI is a measure of the proportion of aliphatic side chains (specifically leucine, valine, alanine, and isoleucine) in relation to the overall volume. This index is considered a favourable factor in enhancing the thermostability of globular proteins. The grand average hydropathicity known as GRAVY was determined by adding the values of hydropathy of all aa and then dividing them with the number of residues. The positive GRAVY value suggests a higher level of hydrophobicity and also provide information about the composition of *nif*H protein and its instability index. A protein with an instability index below 40 is expected to be stable.

Amino acid composition

Amino acid composition of all three *Rhizobium* strains were assessed using the Expasy ProtParam (http://web.expasy.org/protparam/) [25].

Secondary structure prediction

For predicting the entire secondary structure, the PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) software was employed and all the secondary elements, namely coils, helices, and sheets were analyzed [26]. Moreover, MEMSAT-SVM was used to predict the transmembrane topology of protein.

3D structure modelling and verification

Among the three strains, *Rhizobium* sp. TN04 was selected for the prediction of its 3D structure. Iterative Threading Assembly Refinement (I-TASSER) is a modelling technique that predicts atomic models using primary protein sequences of amino acids. It utilises a multiple-threading strategy to identify templates for protein of interest and subsequently generates a 3-D model using data obtained from these templates [27]. The constructed 3D model was assessed using the SAVES server (http://services.mbi.ucla.edu/SAVES/) to verify its quality. ERRAT (http://services.mbi.ucla.edu/ERR

AT/) was used to differentiate accurately identified protein regions from those that might have been improperly registered. The non-random distribution of atoms within the protein may potentially become randomised throughout the process of protein modelling. ERRAT examines the query model representing the distribution of atoms and provides an overall quality factor assessing non-bonding atomic interactions. Higher scores indicate higher quality, with an accepted range typically greater than 50, as stated by Li and Wang in 2007 and Naveed et al. in 2016. COFACTOR and COACH were used to identify enzyme and ligand binding sites [27].

CATH classification

CATH (http://www.cathdb.info/search/by_structure) is a system that organises protein domain structures into a hierarchical classification. The acronym CATH is formed by the first letters of the highest four levels of the classification system. The "Class" level represents the general composition of the domain's secondary structure, indicating whether it mainly consists of alpha-helices, beta-sheets, a mixture of both, or a small number of secondary structures. "Architecture" denotes significant structural resemblance, but there is no indication of homology, such as the presence of an alpha/beta sandwich. "Topology" refers to the arrangement of elements on an enormous level with specific structural characteristics. The Homologous Superfamily provides a verified evolutionary connection [28]. This classification was performed on the nifH protein of *Rhizobium* sp. TN04 to elucidate the structural characteristics of this protein.

Protein–protein association network

The STRING (http://stringdb.org/newstring_cgi) database was used to find genes that have functional relationships with *Rhizobium* sp. TN04. To achieve this, STRING executes searches within the clusters where the query gene has been identified repetitively. The concept of STRING has been based upon earlier studies indicating that genes which frequently appear near each other in genomes have a tendency to produce proteins that are functionally related and participate in similar metabolic pathways [29, 30].

Results and discussion

This study uses bioinformatic tools to predict the structure and function of *nif*H gene of *Rhizobium* sp. TN04 isolated from non-leguminous potato plants. The

prediction is based on phylogenetic analysis, sequence alignment, assessment of secondary structure, CATH classification and functional activity analysis. Since *nif*H genes are agriculturally important and highly conserved [31], therefore it was expected that there would be maximum structural and functional similarity between *nif*H genes of non-leguminous *Rhizobium* sp. TN04 to the strains isolated from

Table 1: Identification of *nif*H gene and protein of TN04 strain isolated from non-leguminous potato plant

Identification	***nif*H Gene**	**nifH Protein**
Closest GenBank match	*Rhizobium sp.* strain S1SS148 (100%)	*Rhizobium sp.* strain S1SS148 (100%)
GenBank accession number	LT596587	SBV08691

environmental sample and leguminous plants.The gene and amino acid sequences of *nif*H from 21 strains of *Rhizobium* spp., including 11 species were obtained from the NCBI and UniProt database. UniProt is a well-known database that researchers can use to acquire detailed information about the richness, accuracy, and quality of particular proteins. It offers a wide range of querying interfaces and freely accessible cross-references [32]. These strains showed varying lengths of amino acid residue and include partial CDS sequence of *nif*H, excluding entire genome sequences. The *nif*H gene is mainly used as a genetic marker to identify diazotrophic bacteria and has been conserved throughout evolutionary history [33]. BLAST analysis showed that the *nif*H gene and protein of TN04 isolate exhibited significant similarity (99-100%) with other strains of *Rhizobium*. TN04 (Gene and protein accession number LT596587 and SBV08691, respectively) showed a 99% similarity with the *nif*H gene and protein of *Rhizobium* sp. S1SS148 (Table 1). The *nif*H gene and protein sequences of TN04 were subjected to phylogenetic analysis using the neighbour-joining method, which demonstrated that TN04 clustered closely with *Rhizobium* sp. S1SS148 and *Rhizobium rosettiformans* W3, with a bootstrap value greater than 90 (Figure 1 and Figure 2). Celador et al., [34] conducted phylogenetic analysis of *nif*H genes from *Rhizobium* sp. isolated from

maize plant, and similarly observed that two distinct strains of *Rhizobium* clustered together.

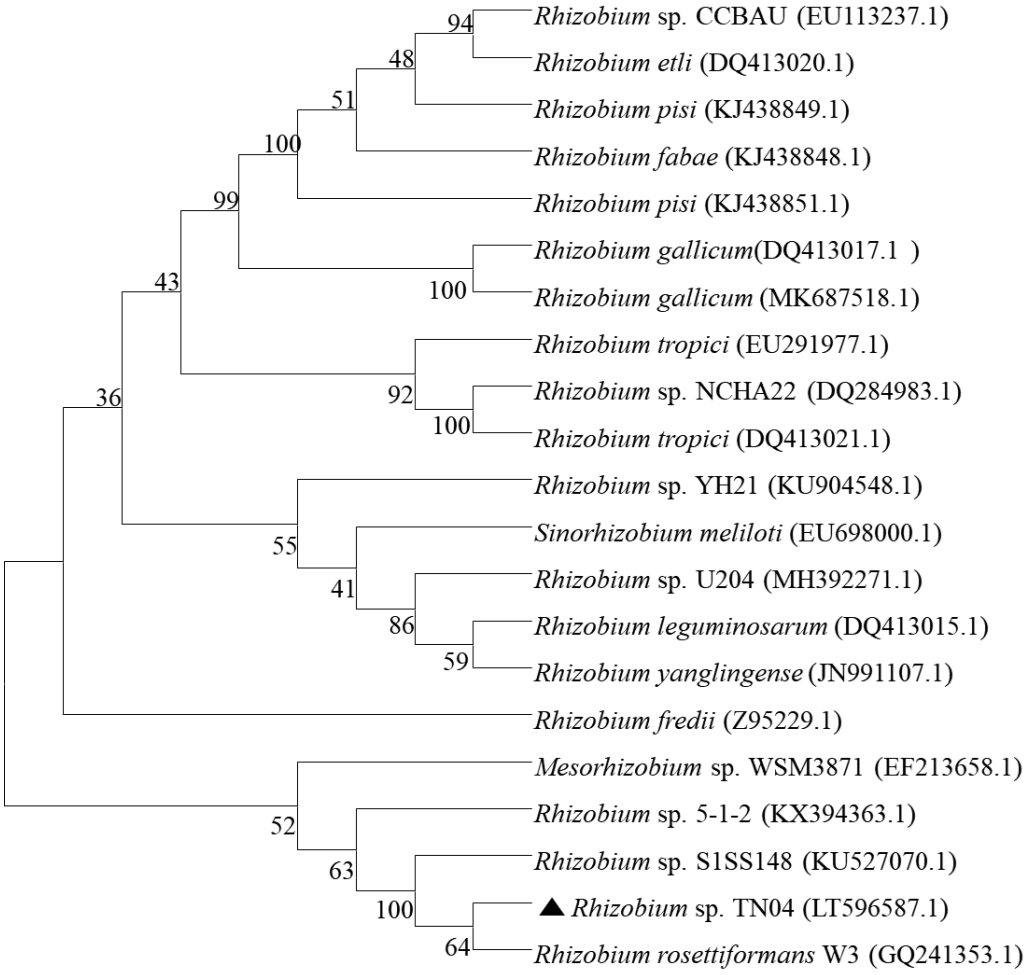

Figure 1: Phylogenetic tree constructed using *nif*H gene of *Rhizobium* sp. isolated from potato rhizosphere (▲) in comparison with previously published sequences. The numbers indicated at the branch points represent bootstrap values greater than 70%.

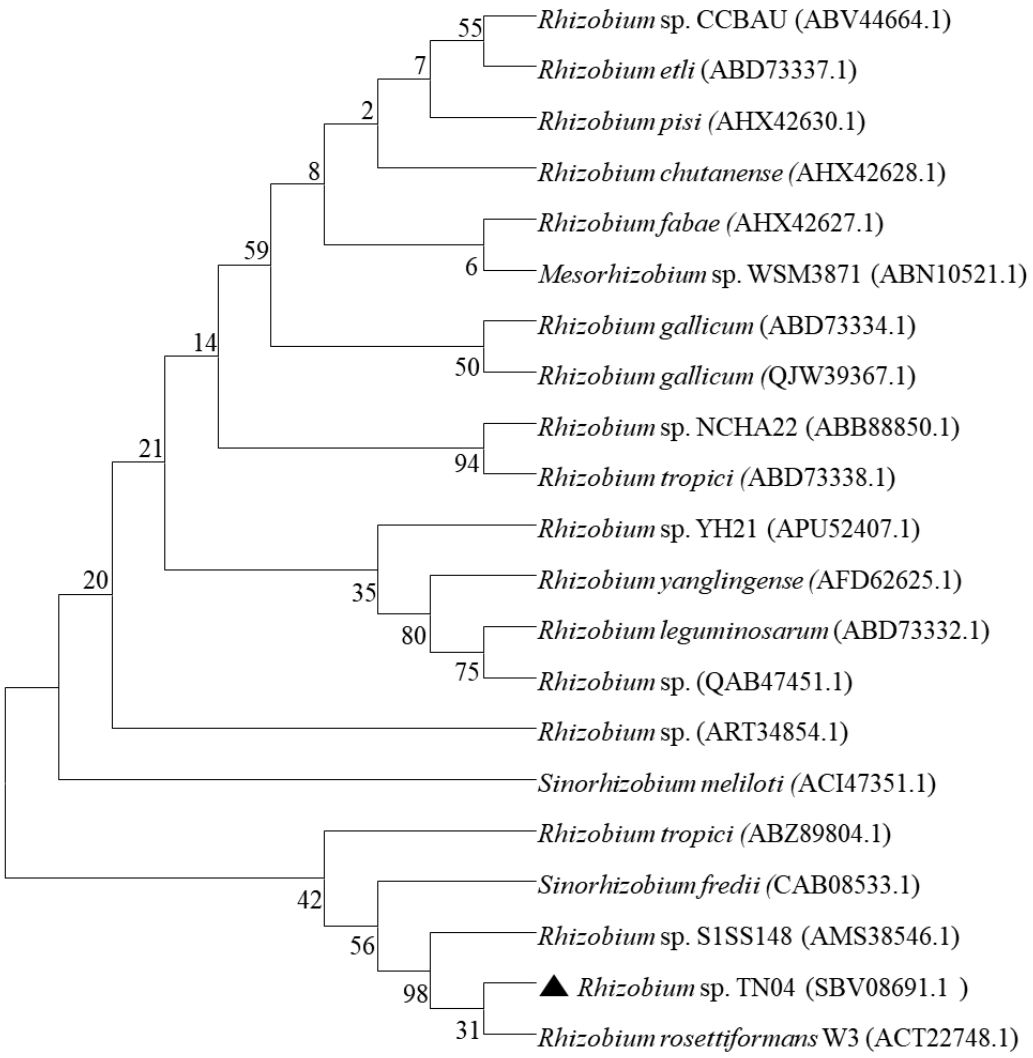

Figure 2: Phylogenetic tree constructed using nifH protein of *Rhizobium* sp. isolated from potato rhizosphere (▲) in comparison with various protein sequences. The numbers indicated at the branch points represent bootstrap values greater than 70%.

The strains showing maximum similarity with *Rhizobium* sp. TN04 were further characterized on the basis of their physiochemical properties including molecular weight, number of amino acid residues, theoretical pI, instability indices, half-life, aliphatic indices, GRAVY, positively and negatively charged residues, stability and extinction coefficients. The proteins have a molecular weight ranging from 12221.79 to 10220.43 Da while their amino acid residues were determined based on the main sequence, which ranges between from 96 to 115. The isoelectric point is defined as the pH value at which a protein surface becomes charged

Table 2: Physiochemical properties of nifH protein of Rhizobium strains

Organism name	***Rhizobium* sp. TN04**	***Rhizobium* sp. S1SS148**	***Rhizobium rosettiformans* W3**
Molecular Weight	10220.43	11587.07	12221.79
No. of amino acids	96	109	115
pI	3.99	4.29	4.19
Half Life (hrs)	0.8	1.9	30
Instability index (II)	27.45	31.17	31.9
GRAVY	0.07	0.092	0.071
AI	96.35	97.43	93.22
-R	18	18	20
+R	6	8	8
Stability	Stable	Stable	Stable
EC (all pairs of Cys residues form cystine)	7,575	7,575	7,575
EC (assuming all Cys residues are reduced)	7,450	7,450	7,450

whereas its net charge is zero, indicating the stable and compact state of the protein. The isoelectric values of all the three proteins fall within the range of 3.99 to 4.19 indicating their acidic nature consistent with previous studies [35, 36]. While designing buffer systems to purify recombinant proteins using the isoelectric targeting approach, theoretically estimated pI is helpful [37].

The term "half-life" represents the duration required for half of the protein molecules to undergo degradation [38]. *Rhizobium* sp. TN04 exhibits the shortest half-life (0.8 hr), indicating that it undergoes degradation at a considerably faster rate compared with the other two strains of *Rhizobium*. The instability index is an important for determining the stability of protein. A protein is considered stable if its instability index is less than 40 [39]. Among three strains, *Rhizobium* sp. TN04 exhibits the lowest vale of instability index (27.45), suggesting relatively higher stability. The GRAVY index determines the hydrophilic or hydrophobic nature of proteins [40]. A low GRAVY value indicates enhanced affinity towards water, suggesting the hydrophobic nature of protein [41]. In this study, all three strains showed slight positive GRAVY values. Both *Rhizobium* sp. TN04 and *R. rosettiformans* W3 showed similar GRAVY vales (0.07), slightly less than the *Rhizobium* sp. S1SS148 (0.092), suggesting their hydrophobic nature.

Aliphatic index determines the protein stability, and there exists a positive relationship between the higher AI values and increased thermal stability of globular proteins [42]. Among all strains, *Rhizobium* sp. S1SS148 exhibits the highest AI value (97.43), whereas *Rhizobium* sp. TN04 showed an AI value of 96.35. Regarding the negatively (Glu + Asp) and positively (Arg + Lys) charged residues, *Rhizobium* sp. TN04 displayed 18 and 8, respectively, *Rhizobium* sp. S1SS148 exhibited the same counts of 18 and 8, while *R. rosettiformans* showed 20 and 8, respectively. The total positively charged residues were less compared to negatively charged one, thus it suggests the extracellular nature of the nifH protein [43]. The extinction coefficient measures the light absorption by proteins at a specific wavelength. Computed extinction coefficients and protein concentration enable quantitative analysis of protein-ligand and protein-protein interactions in solution [44]. In this study, EC at a wavelength of 280 nm in water was estimated, assuming all cysteine residues are either in their reduced state or not. The EC value of nifH protein of all three strains was similar.

The polypeptide chains of proteins are structured with 20 amino acid residues, each possessing distinct characteristics crucial for particular functions within a protein. The percentages of charge, polarity, aromatic, and aliphatic characteristics of proteins change depending on their function and location [45]. Phosphorylation plays a pivotal role in enabling signaling pathways to function. Among the primary amino acid

residues, Threonine and Tyrosine are commonly phosphorylated due to their side chains containing hydroxyl groups that facilitate phosphate group binding [46]. In this study, ProtParam tool was used to estimate all 20 amino acids. Among them, Glycine had the highest percentage, with values of 13.5, 12.80, and 13 in *Rhizobium* sp. TN04, *Rhizobium* sp. S1SS148 and *R. rosettiformans*, respectively. While no percentage of Typtophan was observed in any of the three strains (Table 3).

Table 3: Amino acid composition of nifH protein of *Rhizobium* strains

Sr. No.	**Amino acids**	***Rhizobium* sp. TN04**	***Rhizobium* sp. S1SS148**	***Rhizobium rosettiformans* W3**
1	Alanine	7.30%	8.30%	8.70%
2	Arginine	4.20%	4.60%	4.30%
3	Asparagine	3.10%	2.80%	2.60%
4	Aspartic acid	8.30%	7.30%	7.80%
5	Cysteine	3.10%	2.80%	2.60%
6	Glutamine	2.10%	1.80%	1.70%
7	Glutamic acid	10.40%	9.20%	9.60%
8	Glycine	13.50%	12.80%	13%
9	Histidine	0%	0.90%	0.90%
10	Isoleucine	8.30%	8.30%	7.80%
11	Leucine	5.20%	6.40%	6.10%
12	Lysine	2.10%	2.80%	2.60%
13	Methionine	2.10%	2.80%	4.30%
14	Phenylalanine	2.10%	1.80%	1.70%
15	Proline	3.10%	2.80%	2.60%
16	Serine	5.20%	6.40%	6.10%
17	Threonine	2.10%	2.80%	2.60%

18	Tryptophan	0%	0%	0%
19	Tyrosine	5.20%	4.60%	4.30%
20	Valine	12.50%	11%	10.40%

The PSIPRED tool was used to predict the secondary structure [26] of nifH protein in all three *Rhizobium* strains. Results showed that nifH protein from both non-leguminous and leguminous strain of *Rhizobium* consist of the three main secondary conformations: sheets, coils, and helices (Figure 3 and 4). Coils are flexible segments within a protein that do not possess well-defined secondary structures. A higher percentage of coils may suggest increased surface accessibility and flexibility [47]. These areas are frequently involved in substrate

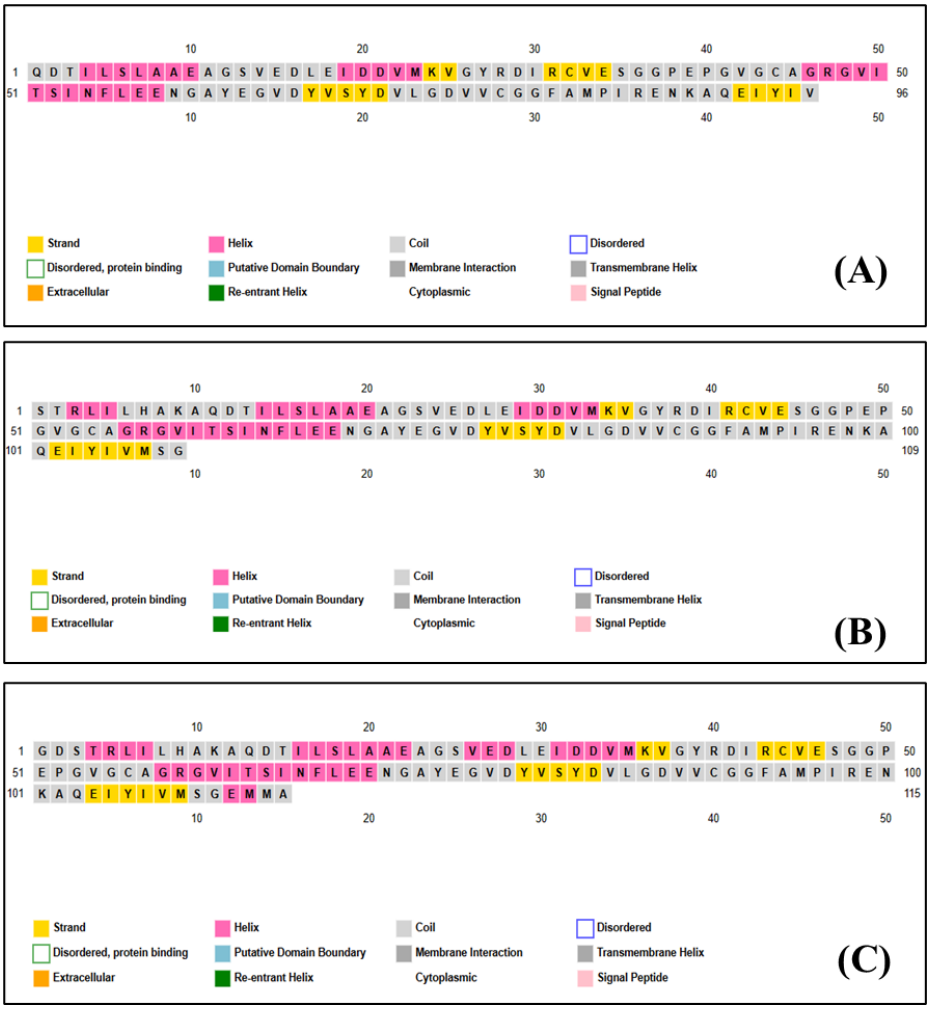

Figure 3: Secondary structure map of nifH protein of (A) *Rhizobium* sp. TN04; (B) *Rhizobium* sp. S1SS148; and (C) *Rhizobium rosettiformans*

binding, enzymatic activity, and protein-protein interactions due to their flexibility, which enables them to adjust with various molecular substrates [48].

The presence of secondary configurations including α and β helices suggests that the nifH proteins in all strains are not in an unfolded state, indicating their stable nature. Thermophiles have been found to have a higher proportion of their amino acid residues in α-helical configuration in order to tolerate high temperatures [49]. In this study, the presence of α-helices in nifH protein suggests its thermally stable nature. Moreover, Roy et al., [50] demonstrated that it is important to detect structural modifications in protein of interest, especially regarding its function and stability, as it undergoes changes under different conditions.

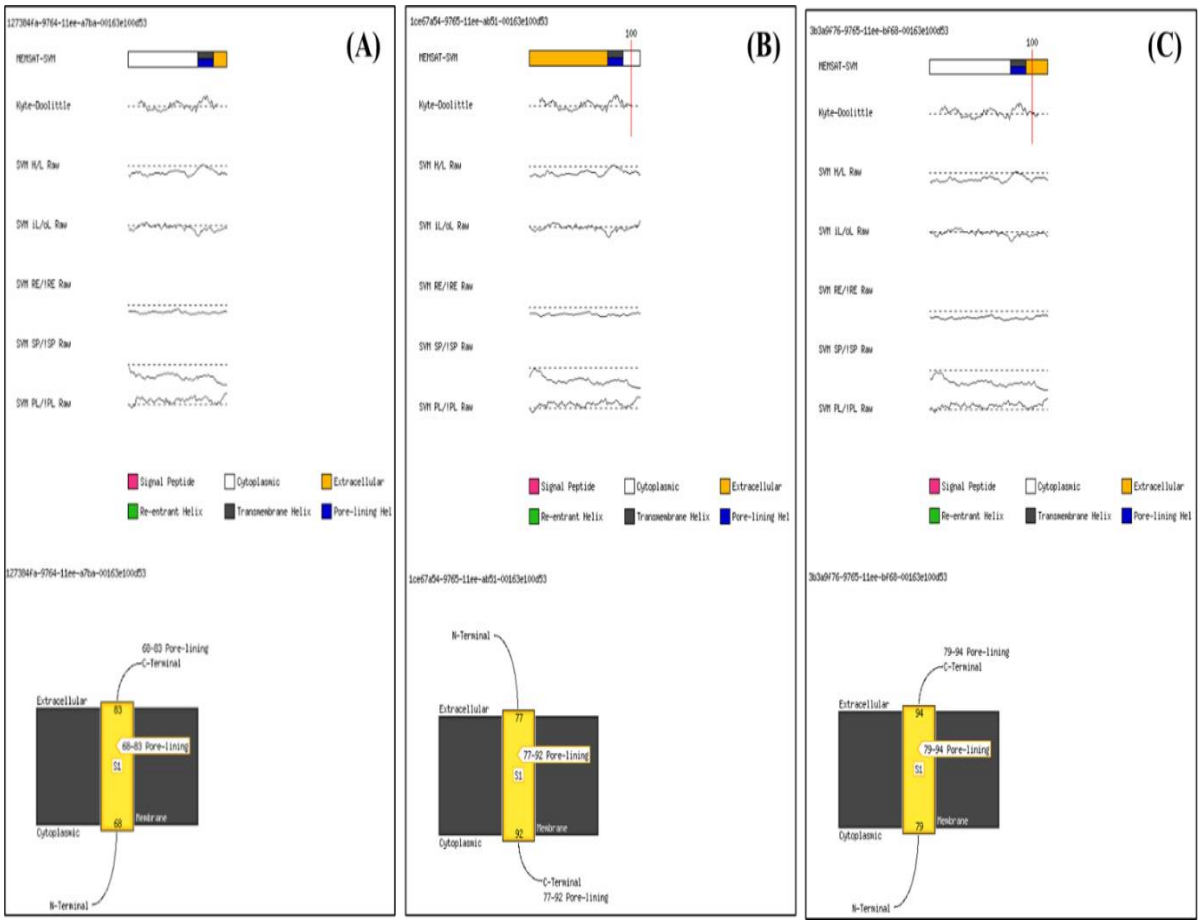

Figure 5: Transmembrane topology of nifH protein of (A) *Rhizobium* sp. TN04; (B) *Rhizobium* sp. S1SS148; and (C) *Rhizobium rosettiformans*

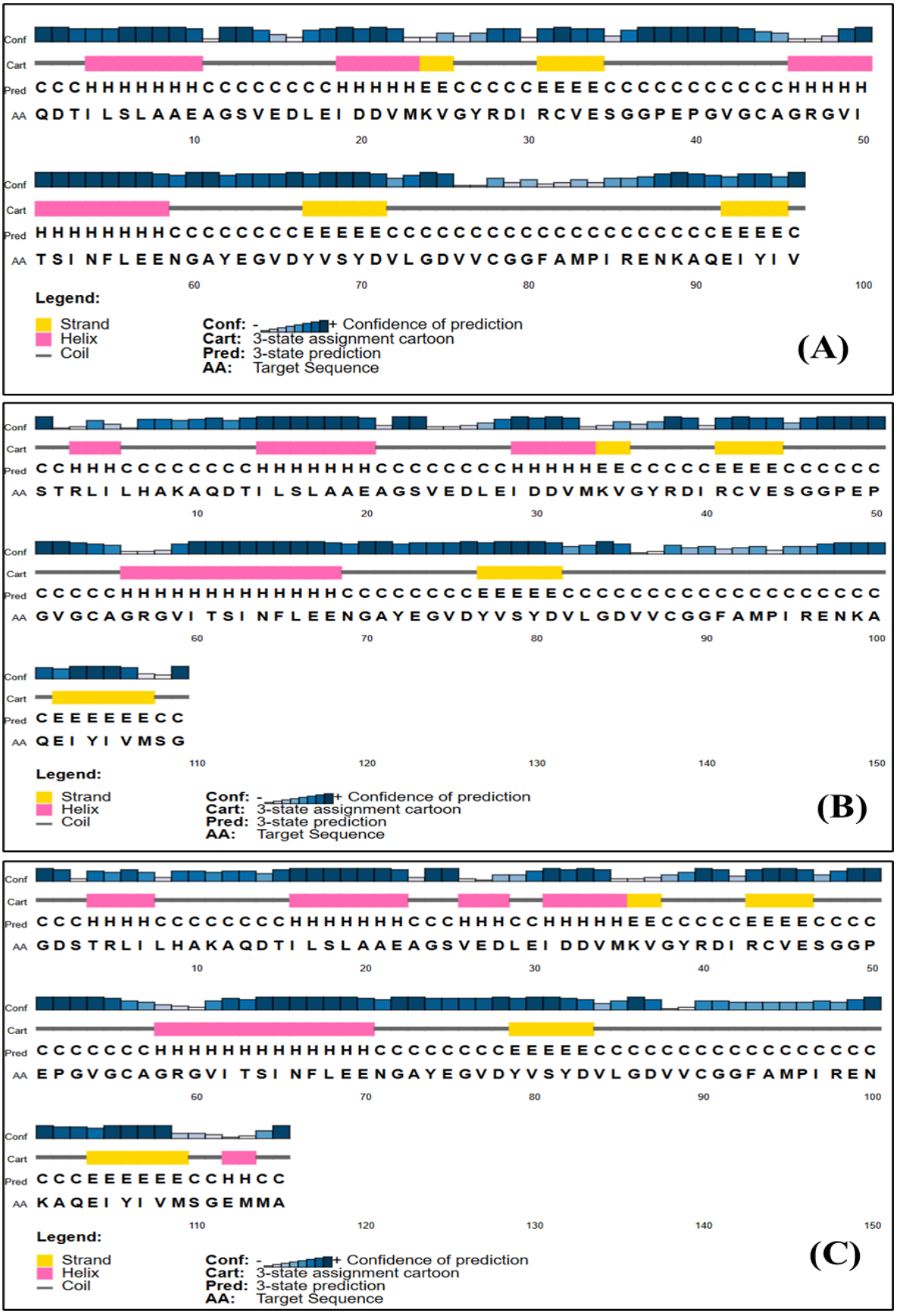
Conf
Cart
Pred C C C H H H H H H H C C C C C C C C C H H H H H E E C C C C C E E E E C C C C C C C C C C C C H H H H H
AA Q D T I L S L A A E A G S V E D L E I D D V M K V G Y R D I R C V E S G G P E P G V G C A G R G V I
10 20 30 40 50
Conf
Cart
Pred H H H H H H H H C C C C C C C C C E E E E E C C C C C C C C C C C C C C C C C C C C C E E E E C
AA T S I N F L E E N G A Y E G V D Y V S Y D V L G D V V C G G F A M P I R E N K A Q E I Y I V
60 70 80 90 100
Legend:
Strand
Helix
Coil
Conf: - + Confidence of prediction
Cart: 3-state assignment cartoon
Pred: 3-state prediction
AA: Target Sequence
(A)
Conf
Cart
Pred C C H H H C C C C C C C C C H H H H H H H C C C C C C C C H H H H H E E C C C C C E E E E C C C C C C
AA S T R L I L H A K A Q D T I L S L A A E A G S V E D L E I D D V M K V G Y R D I R C V E S G G P E P
10 20 30 40 50
Conf
Cart
Pred C C C C C H H H H H H H H H H H H C C C C C C C C C E E E E E C C C C C C C C C C C C C C C C C C C C
AA G V G C A G R G V I T S I N F L E E N G A Y E G V D Y V S Y D V L G D V V C G G F A M P I R E N K A
60 70 80 90 100
Conf
Cart
Pred C E E E E E E C C
AA Q E I Y I V M S G
110 120 130 140 150
Legend:
Strand
Helix
Coil
Conf: - + Confidence of prediction
Cart: 3-state assignment cartoon
Pred: 3-state prediction
AA: Target Sequence
(B)
Conf
Cart
Pred C C C H H H C C C C C C C C C H H H H H H H C C C H H H C C H H H H H E E C C C C C E E E E C C C C
AA G D S T R L I L H A K A Q D T I L S L A A E A G S V E D L E I D D V M K V G Y R D I R C V E S G G P
10 20 30 40 50
Conf
Cart
Pred C C C C C C C H H H H H H H H H H H H C C C C C C C C C E E E E E C C C C C C C C C C C C C C C C C C C C
AA E P G V G C A G R G V I T S I N F L E E N G A Y E G V D Y V S Y D V L G D V V C G G F A M P I R E N
60 70 80 90 100
Conf
Cart
Pred C C C E E E E E E C C H H C C
AA K A Q E I Y I V M S G E M M A
110 120 130 140 150
Legend:
Strand
Helix
Coil
Conf: - + Confidence of prediction
Cart: 3-state assignment cartoon
Pred: 3-state prediction
AA: Target Sequence
(C)

Figure 4: Secondary structure prediction of nifH protein of (A) *Rhizobium* sp. TN04; (B) *Rhizobium* sp. S1SS148; and (C) *Rhizobium rosettiformans*

The MEMSAT-SVM tool has been developed to predict the topology of transmembrane domains in proteins [51]. Although, the ni*f*H protein is not classified as a membrane protein [52], the MEMSAT-SVM software was used to identify any specific segments or features within the nifH protein that might be related to structural features or interactions associated with membranes. In this study, all the three strains showed that nifH is extracellular (Figure 5), which aligns with the results obtained from physiochemical analysis.

The functional annotations of the nifH protein from *Rhizobium* sp. TN04 was further investigated assessed using various software, namely I-TASSER, COFACTOR, and COACH. The I-TASSER modelling process initiates by utilizing structure templates identified through the LOMETS approach from the PDB library. LOMETS functions as a meta-server threading system comprising of multiple threading programs, each generating numerous template alignments. Within I-TASSER, only the most significant templates from the threading alignments are utilized, as determined by their Z-score [53]. This score represents the deviation between average and scores, focusing on the selection of the top 10 templates extracted from threading programs (Figure 6).

Rank	PDB Hit	Iden1	Iden2	Cov	Norm. Z-score	Download Align.
1	1de0A	0.77	0.75	0.96	2.11	Download
2	7t4hA	0.79	0.76	0.96	5.07	Download
3	6nzjA	0.65	0.61	0.95	2.29	Download
4	6nzjA	0.65	0.61	0.95	1.53	Download
5	6nzjA	0.65	0.61	0.95	3.43	Download
6	6nzjA	0.65	0.61	0.95	2.53	Download
7	6n4lA	0.80	0.80	1.00	1.83	Download
8	6nzjA	0.65	0.61	0.95	2.41	Download
9	7t4hA	0.78	0.76	0.96	0.80	Download
10	6uykA	0.36	0.40	0.98	4.84	Download

(a) All the residues are colored in black; however, those residues in template which are identical to the residue in the query sequence are highlighted in color. Coloring scheme is based on the property of amino acids, where polar are brightly coloured while non-polar residues are colored in dark shade.
(b) Rank of templates represents the top ten threading templates used by I-TASSER.
(c) Ident1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence.
(d) Ident2 is the percentage sequence identity of the whole template chains with query sequence.
(e) Cov represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of query protein.
(f) Norm. Z-score is the normalized Z-score of the threading alignments. Alignment with a Normalized Z-score >1 mean a good alignment and vice versa.
(g) Download Align. provides the 3D structure of the aligned regions of the threading templates.
(h) The top 10 alignments reported above (in order of their ranking) are from the following threading programs:
1: FFAS-3D 2: SPARKS-X 3: HHSEARCH2 4: HHSEARCH I 5: Neff-PPAS 6: HHSEARCH 7: pGenTHREADER 8: wdPPAS 9: PROSPECT2 10: SP3

Figure 6: Top 10 templates extracted from threading programs

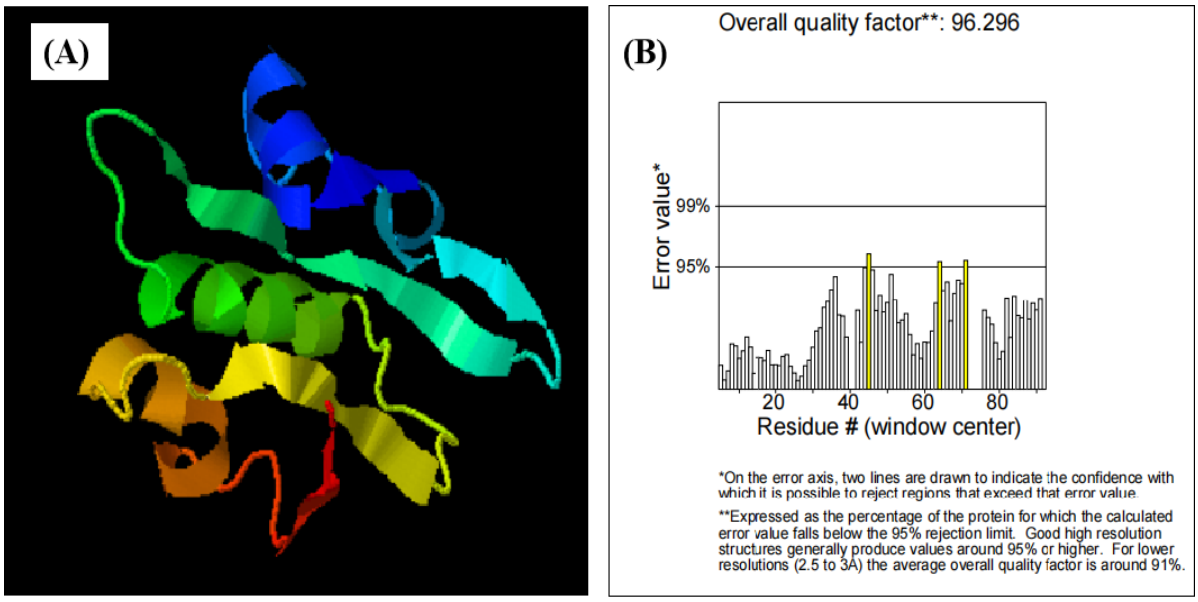

Figure 7: I-TASSER and ERRAT analysis of nifH protein of TN04 (A) 3D Structure prediction (B) ERRAT verification

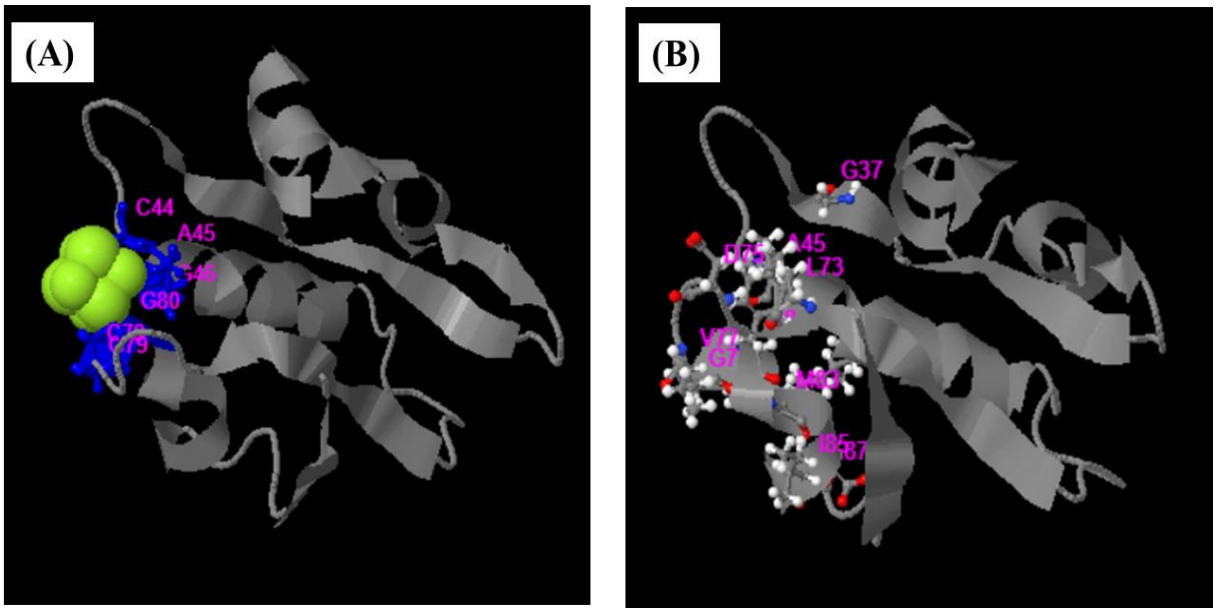

Figure 8: Functional annotation of TN04 nifH protein (A) Ligand binding sites (B) Homology with dintrogenase enzyme

Level	CATH Code	Description
C	3	Alpha Beta
A	3.40	3-Layer(aba) Sandwich
T	3.40.50	Rossmann fold
H	3.40.50.300	P-loop containing nucleotide triphosphate hydrolases

Figure 9: CATH classification of nifH protein of TN04.

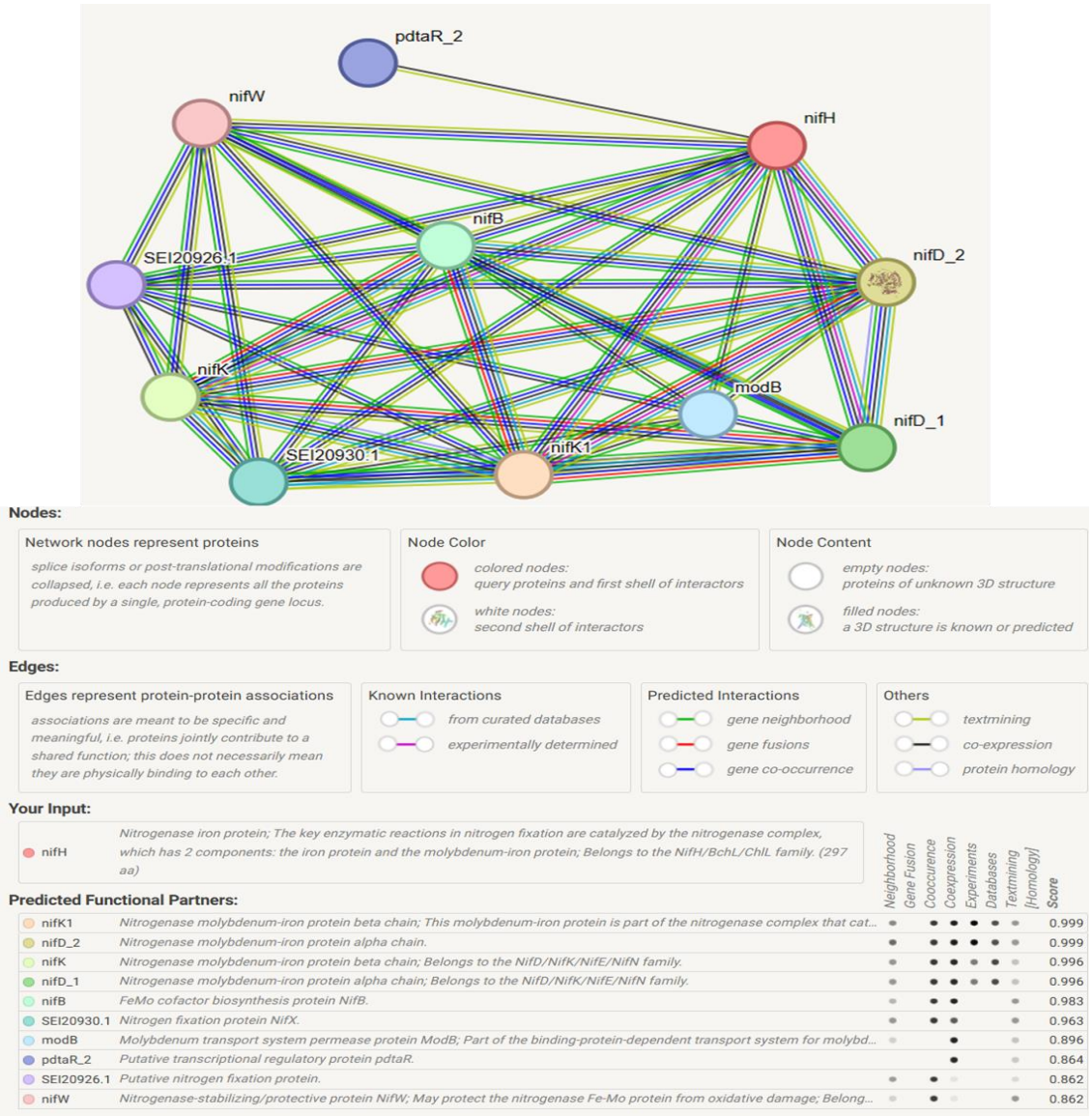

Figure 10: STRING analysis for interaction nifH protein with other proteins.

In the present study, the template with highest Z-score was chosen.I-TASSER predicted the 3D structure of the nif*H* protein with a C-score of 0.57, TM-score of

0.79±0.09, and RMSD value of 2.8±2.0Å (Figure 7A). The confidence level of each model was quantitatively evaluated using the C-score, which is determined based on the significance of threading template alignments and the convergence parameters observed during structure assembly simulations. It ranges between -5 to 2; a higher C-score indicates a model with increased confidence, whereas a lower value suggests lower confidence [54]. The nifH protein of *Rhizobium* sp. TN04 falls within this range. Furthermore, RMSD and TM-score values are estimated from the protein length and C-score, as these metrics show correlation with the observed qualities of the modelled structures [27]. Moreover, this 3D structure was validated using ERRAT to detect potential misconfigurations of atoms. The ERRAT analysis yielded an overall quality factor of 96.29 (Figure 7B), indicating the high quality of the 3D model depicted in Figure 7A.

The biological annotations of the nifH protein from *Rhizobium* sp. TN04 were conducted using COFACTOR and COACH, based on the I-TASSER structure prediction. The COACH tool uses meta-server approach that combines multiple function annotation results, specifically concentrating on the analysis of ligand-binding sites [55]. The analysis showed structural similarity between TN04 and a known structure in the PDB (1m1yN) database (Figure 8). Results showed that TN04 can bind to a ligand called SF4 at specific amino acid positions (44-46 and 78-80) within the protein structure with high confidence level (C-score of 0.83).

Moreover, the COFACTOR software used protein-protein networks and structural comparisons to deduce various protein functions such as EC numbers [56]. The COFACTOR analysis of TN04 showed a low RMSD value of 1.20 Å and a high TM-score of 0.914, indicating minimal structural divergence between the atomic arrangements of the two proteins (TN04 and Pbd hit: 1cp2B), suggesting considerable structural similarity. Furthermore, IDENa (sequence identity) of 0.734 indicates significant similarity within the aligned regions of amino acid sequences.

COFACTOR identified an enzyme homolog of the TN04 nifH protein, displaying a C-score of 0.507, closely resembling with EC number of 1.18.6.1 (Figure 8B). This enzyme functions as a complex comprising two distinct components: dinitrogenase and dinitrogen reductase.

Dinitrogen reductase, characterized as a [4Fe-4S] protein, facilitates the transfer of an electron from ferredoxin to the dinitrogenase component upon stimulated by the presence of two ATP molecules [57]. The provided active site residues (37, 45, 50, 73, 75, 77, 79, 81, 83, 85, 87) are particularly important for the aligned protein enzymatic activity or substrate interaction. The nifH protein from *Rhizobium* sp. TN04 was classified using the CATH database which showed 10 matching domains. For this study, the classification was predicted using the 6n4lA00 matching domain, which showed a significant E value of 7.8 x 10^{-41}. This protein is classified as "Alpha Beta" at the Class level (Class 3), indicating its overall structural fold (Figure 9). At the Architecture level, it showed "3-Layer(aba) Sandwich" (Architecture 3.40), specifying the arrangement of its secondary structures. Further, at the level of Topology, it exhibits a "Rossmann fold" (Topology 3.40.50), elucidating its specific arrangements and associations within this architecture. Lastly, it falls under the "P-loop containing nucleotide triphosphate hydrolases" homologous superfamily (Homologous Superfamily 3.40.50.300), indicating a group of proteins with shared structural and evolutionary characteristics. This hierarchical classification system aids in understanding its structural nature and evolutionary relationships [58]. Moreover, CATH also determine the functional protein family, which matched with "Nitrogenase iron protein 1 (3.40.50.300/FF/1379)".

Rhizobium sp. TN04 showed a P-loop NTPase domain, a conserved motif involved in binding nucleotide phosphates, as outlined in previous studies [59]. This domain houses Walker A and B motifs, which are known for their role in nucleotide binding and specifically serving as binding sites for Mg^{2+} ions. Studies have established the significance of Mg^{2+} ions as essential cofactors crucial for nitrogenase activity, particularly in conjunction with the Fe-protein of dinitrogenase reductase [60, 61]. This structural similarity suggests a potential association between the characteristics observed in TN04 and its ability to facilitate nitrogen fixation.

The STRING programme was used to predict the protein-protein interaction of TN04 nifH protein. Results showed that TN04 interacts with various nif proteins including nifK, nifK1, nifD_1 and 2, nifB, nifX, and ifW (Figure 10), suggesting its function as potential N-fixer similar to leguminous *Rhizobium* strains. Studies have reported that majority of these

associated *nif* genes, including *nif*N, *nif*E, *nif*K, *nif*D, *nif*X and *nif*B, within this network are major elements of the *nif* operon and play direct roles in the process of N-fixation [62, 63]. Thus, it can be concluded that nifH protein of *Rhizobium* sp. TN04 has similar N-fixation potential to that observed in leguminous *Rhizobium* strains. However, this study has several limitations. It only included partial CDS sequences, which may overlook essential regulatory elements or other functional segments associated with N-fixation. Furthermore, this study is based on computational analysis only, therefore, further validation through *in vivo* experiments is required.

Conclusion

In order to get effective outcomes in *Rhizobium*-mediated biological N-fixation in non-leguminous plants, an in-depth understanding of protein structure is necessary. *In silico* analysis provide a valuable approach for modelling protein structures, offering a promising way to enhance our understanding of this biological process. In the present study, *nif*H sequence of *Rhizobium* sp. TN04 isolated from non-leguminous potato plant was selected to investigate its evolutionary relationship, physicochemical characteristics and different structural aspects of the protein utilizing *in silico* approaches. Primary structural analysis indicated that the nifH protein of TN04 is stable, hydrophobic and had acidic nature. The nifH protein can be sensitive to oxygen due to cysteine residues in its structure. Secondary structure prediction showed that the presence of random coils, β-turns and α-helix in the nifH protein sequences. The 3D structure of the nifH protein was predicted using I-TASSER, resulting in a 0.57 C-score, TM-score of 0.79±0.09, and 2.8±2.0Å RMSD value. ERRAT analysis confirmed the good quality of the protein structure and suggested potential interactions with various nif proteins including nifK, nifK1, nifD_1 and 2, nifB, nifX, and nifW, suggesting it role as potential N-fixer similarl to leguminous strains. This study contributed in identifying and characterizing nifH protein using computational approaches, however, further validation is required through *in vivo* experiments and modelling techniques to authenticate the findings presented in this study. Moreover, further research is also needed to better understand the molecular pathways involved in N-fixation in the *Rhizobium* sp. TN04 focusing on predicting the quaternary structure of its nifH protein.

References

1. Zhang, X., et al., *Quantification of global and national nitrogen budgets for crop production.* Nature Food, 2021. **2**(7): p. 529-540.
2. Akter, Z., et al., *Biological nitrogen fixation and nif H gene expression in dry beans (Phaseolus vulgaris L.).* Canadian Journal of Plant Science, 2014. **94**(2): p. 203-212.
3. Shamseldin, A., *The role of different genes involved in symbiotic nitrogen fixation—review.* Global Journal of Biotechnology & Biochemistry, 2013. **8**(4): p. 84-94.
4. Thakur, S., A.K. Bothra, and A. Sen, *Exploring the genomes of symbiotic diazotrophs with relevance to biological nitrogen fixation.* Agricultural Bioinformatics, 2014: p. 235-257.
5. Imran, A., et al., *Diazotrophs for lowering nitrogen pollution crises: looking deep into the roots.* Frontiers in Microbiology, 2021. **12**: p. 637815.
6. Jacobson, M.R., et al., *Physical and genetic map of the major nif gene cluster from Azotobacter vinelandii.* Journal of bacteriology, 1989. **171**(2): p. 1017-1027.
7. Fischer, H.-M., *Genetic regulation of nitrogen fixation in rhizobia.* Microbiological reviews, 1994. **58**(3): p. 352-386.
8. Yan, Y., et al., *Nitrogen fixation island and rhizosphere competence traits in the genome of root-associated Pseudomonas stutzeri A1501.* Proceedings of the National Academy of Sciences, 2008. **105**(21): p. 7564-7569.
9. Clúa, J., et al., *Compatibility between legumes and rhizobia for the establishment of a successful nitrogen-fixing symbiosis.* Genes, 2018. **9**(3): p. 125.
10. Andrews, M. and M.E. Andrews, *Specificity in legume-rhizobia symbioses.* International journal of molecular sciences, 2017. **18**(4): p. 705.
11. Sprent, J.I., J. Ardley, and E.K. James, *Biogeography of nodulated legumes and their nitrogen-fixing symbionts.* New Phytologist, 2017. **215**(1): p. 40-56.
12. Yanni, Y.G., et al., *The beneficial plant growth-promoting association of Rhizobium leguminosarum bv. trifolii with rice roots.* Functional Plant Biology, 2001. **28**(9): p. 845-870.
13. Mishra, R.P., et al., *Rice–rhizobia association: evolution of an alternate niche of beneficial plant–bacteria association.* Plant-Bacteria Interactions: Strategies and Techniques to Promote Plant Growth, 2008: p. 165-193.
14. Mehboob, I., et al., *Comparative effectiveness of different Rhizobium sp. for improving growth and yield of maize (Zea mays L.).* Soil & Environment, 2012. **31**(1).
15. Wijesinha-Bettoni, R. and B. Mouillé, *The contribution of potatoes to global food security, nutrition and healthy diets.* American Journal of Potato Research, 2019. **96**: p. 139-149.
16. Zaheer, K. and M.H. Akhtar, *Potato production, usage, and nutrition—a review.* Critical reviews in food science and nutrition, 2016. **56**(5): p. 711-721.
17. Calvo, P., et al., *Characterization of Bacillus isolates of potato rhizosphere from andean soils of Peru and their potential PGPR characteristics.* Brazilian Journal of Microbiology, 2010. **41**: p. 899-906.
18. Kesaulya, H., B. Zakaria, and S.A. Syaiful, *Isolation and physiological characterization of PGPR from*

potato plant rhizosphere in medium land of Buru Island. Procedia Food Science, 2015. **3**: p. 190-199.
19. Pathak, D., et al., *Plant growth promoting rhizobacterial diversity in potato grown soil in the Gwalior region of India.* Biotechnology Reports, 2022. **33**: p. e00713.
20. Naqqash, T., *Biodiversity of Diazotrophs in Rhizosphere of Potato (Solanum tuberosum L.)*. 2017, Pakistan Institute of Engineering and Applied Sciences.
21. Singh, N., et al., *In silico analysis of protein.* J Bioinform Genomics Proteomics, 2016. **1**(2): p. 1007.
22. Tamura, K., G. Stecher, and S. Kumar, *MEGA11: molecular evolutionary genetics analysis version 11.* Molecular biology and evolution, 2021. **38**(7): p. 3022-3027.
23. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Molecular biology and evolution, 1987. **4**(4): p. 406-425.
24. Hillis, D.M. and J.J. Bull, *An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis.* Systematic biology, 1993. **42**(2): p. 182-192.
25. Gasteiger, E., et al., *Protein identification and analysis tools on the ExPASy server*. 2005: Springer.
26. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-405.
27. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC bioinformatics, 2008. **9**: p. 1-8.
28. Sillitoe, I., et al., *CATH: comprehensive structural and functional annotations for genome sequences.* Nucleic acids research, 2015. **43**(D1): p. D376-D381.
29. Szklarczyk, D., et al., *STRING v10: protein–protein interaction networks, integrated over the tree of life.* Nucleic acids research, 2015. **43**(D1): p. D447-D452.
30. Naveed, M., et al., *Structural and functional annotation of conserved virulent hypothetical proteins in Chlamydia trachomatis: an in-silico approach.* Current Bioinformatics, 2019. **14**(4): p. 344-352.
31. Huang, T.-C., et al., *Organization and expression of nitrogen-fixation genes in the aerobic nitrogen-fixing unicellular cyanobacterium Synechococcus sp. strain RF-1.* Microbiology, 1999. **145**(3): p. 743-753.
32. Consortium, U., *UniProt: a hub for protein information.* Nucleic acids research, 2015. **43**(D1): p. D204-D212.
33. Ueda, T., et al., *Remarkable N2-fixing bacterial diversity detected in rice roots by molecular evolutionary analysis of nifH gene sequences.* Journal of bacteriology, 1995. **177**(5): p. 1414-1417.
34. Celador-Lera, L., et al., *Rhizobium zeae sp. nov., isolated from maize (Zea mays L.) roots.* International Journal of Systematic and Evolutionary Microbiology, 2017. **67**(7): p. 2306-2311.
35. Norris, D., *Acid production by Rhizobium a unifying concept.* Plant and Soil, 1965. **22**: p. 143-166.
36. Righetti, P.G., *Determination of the isoelectric point of proteins by capillary isoelectric focusing.* Journal of chromatography A, 2004. **1037**(1-2): p. 491-499.
37. Adhikari, S., et al., *A unified method for purification of basic proteins.*

Analytical biochemistry, 2010. **400**(2): p. 203-206.
38. Bachmair, A., D. Finley, and A. Varshavsky, *In vivo half-life of a protein is a function of its amino-terminal residue.* science, 1986. **234**(4773): p. 179-186.
39. Gamage, D.G., et al., *Applicability of instability index for in vitro protein stability prediction.* Protein and peptide letters, 2019. **26**(5): p. 339-347.
40. Jaspard, E. and G. Hunault, *Comparison of amino acids physico-chemical properties and usage of late embryogenesis abundant proteins, hydrophilins and WHy domain.* PloS one, 2014. **9**(10): p. e109570.
41. Goshe, M.B., J. Blonder, and R.D. Smith, *Affinity labeling of highly hydrophobic integral membrane proteins for proteome-wide analysis.* Journal of proteome research, 2003. **2**(2): p. 153-161.
42. Ikai, A., *Thermostability and aliphatic index of globular proteins.* The Journal of Biochemistry, 1980. **88**(6): p. 1895-1898.
43. Nilsson, I. and G. von Heijne, *Fine-tuning the topology of a polytopic membrane protein: role of positively and negatively charged amino acids.* Cell, 1990. **62**(6): p. 1135-1141.
44. Pace, C.N., et al., *How to measure and predict the molar absorption coefficient of a protein.* Protein science, 1995. **4**(11): p. 2411-2423.
45. Livingstone, C.D. and G.J. Barton, *Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation.* Bioinformatics, 1993. **9**(6): p. 745-756.
46. Tripet, B., et al., *Effects of side-chain characteristics on stability and oligomerization state of a de novo-designed model coiled-coil: 20 amino acid substitutions in position "d".* Journal of molecular biology, 2000. **300**(2): p. 377-402.
47. Park, W.M., *Coiled-coils: The molecular zippers that self-assemble protein nanostructures.* International journal of molecular sciences, 2020. **21**(10): p. 3584.
48. Kokkinidis, M., N. Glykos, and V. Fadouloglou, *Protein flexibility and enzymatic catalysis.* Advances in protein chemistry and structural biology, 2012. **87**: p. 181-218.
49. Kumar, L.S., T. Ramakrishna, and C.M. Rao, *Structural and functional consequences of the mutation of a conserved arginine residue in αA and αB crystallins.* Journal of Biological Chemistry, 1999. **274**(34): p. 24137-24141.
50. Roy, C.R. and J. Cherfils, *Structure and function of Fic proteins.* Nature Reviews Microbiology, 2015. **13**(10): p. 631-640.
51. Nugent, T. and D.T. Jones, *Transmembrane protein topology prediction using support vector machines.* BMC bioinformatics, 2009. **10**: p. 1-11.
52. Hurek, T., et al., *Induction of complex intracytoplasmic membranes related to nitrogen fixation in Azoarcus sp. BH72.* Molecular microbiology, 1995. **18**(2): p. 225-236.
53. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction.* Nature protocols, 2010. **5**(4): p. 725-738.
54. Yang, J. and Y. Zhang, *I-TASSER server: new development for protein structure and function predictions.* Nucleic acids research, 2015. **43**(W1): p. W174-W181.

55. Edgar, R.C. and K. Sjölander, *COACH: profile–profile alignment of protein families using hidden Markov models.* Bioinformatics, 2004. **20**(8): p. 1309-1318.
56. Zhang, C., P.L. Freddolino, and Y. Zhang, *COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information.* Nucleic acids research, 2017. **45**(W1): p. W291-W299.
57. Burgess, B.K. and D.J. Lowe, *Mechanism of molybdenum nitrogenase.* Chemical reviews, 1996. **96**(7): p. 2983-3012.
58. Dawson, N.L., et al., *CATH: an expanded resource to predict protein function through structure and sequence.* Nucleic acids research, 2017. **45**(D1): p. D289-D295.
59. Leipe, D.D., et al., *Classification and evolution of P-loop GTPases and related ATPases.* Journal of molecular biology, 2002. **317**(1): p. 41-72.
60. Lee, J.H., et al., *Constitutive ATP hydrolysis and transcription activation by a stable, truncated form of Rhizobium meliloti DCTD, a sigma 54-dependent transcriptional activator.* Journal of Biological Chemistry, 1994. **269**(32): p. 20401-20409.
61. Kozlova, M.I., et al., *Common Patterns of Hydrolysis Initiation in P-loop Fold Nucleoside Triphosphatases.* Biomolecules, 2022. **12**(10): p. 1345.
62. Lahiri, S., *Molecular-genetic and structural analyses of the NifHDKX proteins of the nitrogenase system*. 2006: Mississippi State University.
63. De Bruijn, F.J., *The quest for biological nitrogen fixation in cereals: a perspective and prospective.* Biological nitrogen fixation, 2015: p. 1087-1101.