A bioinformatics insight to rhizobial globins: gene identification and mapping, polypeptide sequence and phenetic analysis, and protein modeling. [version 1; referees: 2 approved]

Reinier Gesto-Borroto, Miriam Sánchez-Sánchez, Raúl Arredondo-Peter

Laboratorio de Biofísica y Biología Molecular, Centro de Investigación en Dinámica Celular, Instituto de Investigación en Ciencias Básicas y Aplicadas, Universidad Autónoma del Estado de Morelos, Colonia Chamilpa, Morelos, 62210, Mexico

Abstract
Globins (Glbs) are proteins widely distributed in organisms. Three evolutionary families have been identified in Glbs: the M, S and T Glb families. The M Glbs include flavohemoglobins (fHbs) and single-domain Glbs (SDgbs); the S Glbs include globin-coupled sensors (GCSs), protoglobins and sensor single domain globins, and the T Glbs include truncated Glbs (tHbs). Structurally, the M and S Glbs exhibit 3/3-folding whereas the T Glbs exhibit 2/2-folding. Glbs are widespread in bacteria, including several rhizobial genomes. However, only few rhizobial Glbs have been characterized. Hence, we characterized Glbs from 62 rhizobial genomes using bioinformatics methods such as data mining in databases, sequence alignment, phenogram construction and protein modeling. Also, we analyzed soluble extracts from *Bradyrhizobium japonicum* USDA38 and USDA58 by (reduced + carbon monoxide (CO) minus reduced) differential spectroscopy. Database searching showed that only *fhb*, *sdgb*, *gcs* and *thb* genes exist in the rhizobia analyzed in this work. Promoter analysis revealed that apparently several rhizobial *glb* genes are not regulated by a -10 promoter but might be regulated by -35 and Fnr (fumarate-nitrate reduction regulator)-like promoters. Mapping analysis revealed that rhizobial *fhbs* and *thbs* are flanked by a variety of genes whereas several rhizobial *sdgbs* and *gcs* are flanked by genes coding for proteins involved in the metabolism of nitrates and nitrites and chemotaxis, respectively. Phenetic analysis showed that rhizobial Glbs segregate into the M, S and T Glb families, while structural analysis showed that predicted rhizobial SDgbs and fHbs and GCSs globin domain and tHbs fold into the 3/3- and 2/2-folding, respectively. Spectra from *B. japonicum* USDA38 and USDA58 soluble extracts exhibited peaks and troughs characteristic of bacterial and vertebrate Glbs thus indicating that putative Glbs are synthesized in *B. japonicum* USDA38 and USDA58.

This article is included in the Oxygen-binding and sensing proteins channel.
Introduction

Globins (Glbs) are proteins widely distributed in organisms from the three kingdoms of life, i.e., in Archaea, Eubacteria and Eukarya. Structurally, Glbs fold into a tertiary structure known as the globin fold. This protein folding consists of six to eight α-helices (designated with letters A to H) that form a hydrophobic pocket where a heme prosthetic group is located. Two structural types of the globin fold have been identified in Glbs: the 2/2- and 3/3-fold. In the 2/2-Glbs, helices B and E overlap to helices G and H and in the 3/3-Glbs helices A, E and F overlap to helices B, G and H. Likewise, three evolutionary families have been identified in Glbs: the M, S and T Glb families. The M Glbs include flavohemoglobins (fHbs) and single-domain Glbs (sDgb), the S Glbs include globin-coupled sensors (GCSs), protoglobins and sensor single domain globins, and the T Glbs include truncated Glbs (tHbs) which are further classified into class 1, class 2 and class 3 tHbs. Canonical tHbs are ~20 to 40 amino acids shorter than the globin fold, resulting in an almost absent helix A and a helix F that is reduced to a single turn. The M and S Glbs fold into the 3/3-fold whereas the T Glbs fold into the 2/2-fold.

A variety of gaseous ligands bind to the heme Fe of Glbs, most notably O₂ and nitric oxide (NO). The reversible binding of O₂ is associated with the major function of Glbs in organisms: the transport of O₂. Binding of NO by oxygenated Glbs is essential to NO-detoxification via NO-dioxygenase activity. Several additional functions have been reported for Glbs, including dehaloperoxidase activity and reaction with free radicals, binding and transport of sulfide and lipids, and O₂-sensing (reviewed by Giardina et al. and Vinogradov et al.). This indicates that in vivo, Glbs might be multifunctional proteins.

Glbs are widespread in bacteria. A comprehensive genomic analysis revealed that glb genes exist in the genomes of 1185 Eubacteria, including several rhizobial genomes. However, only few rhizobial glb genes have been characterized. Characterizing rhizobial Glbs is of interest because rhizobia establish symbiotic relationships with leguminous plants. A result of this plant-microbe interaction is the symbiotic fixation of atmospheric N₂, which occurs within specialized plant organs called nodules. Symbiotic N₂-fixation is a process modulated by a variety of factors, such as the O₂ and NO levels in the surrounding environment. Glbs bind O₂ and NO and thus may function in some aspects of the N₂-fixation, e.g., by transporting O₂ and detoxifying NO. Modulation of O₂ levels in the plant cell cytoplasm from nodules is well characterized. A plant Glb (leghemoglobin (Lb)) that is synthesized at high (~3 to 5 mM) concentrations in nodules apparently facilitates O₂-diffusion to the symbiotic rhizobia

Forty-six years ago Appleby was the first to propose the existence of Glbs in rhizobia. This author detected absorption peaks and troughs that are characteristic of Glbs in different (dithionite reduced + CO minus dithionite reduced) spectra of soluble extracts from Bradyrhizobium japonicum 505 (Wisconsin). Subsequent spectroscopic analyses suggested the existence of soluble Glbs in Rhizobium leguminosarum bv. viciaeB, janicum NPK63 and R. etli CE3. The first rhizobial glb gene was identified in the pSymA megaplasmid of Sinorhizobium meliloti 1021. BLAST analysis revealed that this gene corresponded to an fhb gene and thus was named smfhb. A bioinformatics analysis showed that smfhb is flanked by nos and fix genes (which code for denitrification enzymes and high O₂-affinity terminal oxidases and an O₂-sensor, respectively) and that apparently it is regulated by an Fnr-like promoter. These observations suggested that smfhb is regulated by the concentration of O₂ and that SmfHb functions in some aspects of nitrogen metabolism. A transcriptomic analysis of the S. meliloti response to NO in culture showed that smfhb (also designated as a S. meliloti hup) is upregulated by NO and the analysis of a smfhb mutant exhibited a high sensitivity to NO in culture and led to a reduced N₂-fixation efficiency in planta. These observations suggested that SmfHb functions in some aspects of NO metabolism, possibly by detoxifying NO.

Genomic analysis reported by Vinogradov et al. revealed that Glb sequences exist in several rhizobia. However, in spite of the above reports knowledge on the rhizobial Glbs is quite limited. Hence, in order to obtain information on the properties of rhizobial Glbs we characterized Glb sequences from selected rhizobial genomes by using bioinformatics methods. These included gene characterization, polypeptide sequence and phenetic analysis, as well as protein modeling. Also, we analyzed soluble extracts from B. japonicum USDA38 and USDA58 by differential spectroscopy. Our main results showed that only fhb, sdgb, gcs and thb genes exist in the rhizobia analyzed in this work; that several rhizobial glb genes are not regulated by a -10 promoter but might be regulated by -35 and Fnr-like promoters; that rhizobial fhs and thbs are flanked by a variety of genes whereas several rhizobial sdgb and gcs genes are flanked by genes coding for proteins involved in the metabolism of nitrates and nitrites and chemotaxis, respectively; that rhizobial Glbs segregate into the M, S and T Glb families; that predicted rhizobial SDgb and fHbs and GCSs globin domain and tHbs fold into the 3/3- and 2/2-fold, respectively, and that spectra from B. japonicum USDA38 and USDA58 soluble extracts exhibit peaks and troughs characteristic of bacterial and vertebrate Glbs.

Methods

Database search

Putative Glb sequences and Glb domains were identified in databases (Table S1) containing the genomes of rhizobial species and strains using the query sequences S. meliloti fHb; Vitreoscilla Sdgb; Agrobacterium tumefaciens GCS; Methanosarcina acetivorans protoglobin; Methylacidiphilum infernorum sensor single domain globin; Mycobacterium tuberculosis tHb class 1; A. tunefaciens fHb class 2; and M. avium fHb class 3 (Genbank accession numbers AY328026, AA75506, NP_354049, 2VEB_A, YP_001939425, NP_216058, WP_020813663 and BAN32501, respectively) and the
SUPERFAMILY database (http://supfam.mrc-lmb.cam.ac.uk). Resulting sequences were subjected to a FUGUE analysis (http:// tardis.nibio.go.jp/fugue/prfsearch.html) to determine the most similar Glb structure and presence of proximal H at the myoglobin-fold position F8. Putative Glbs had to satisfy the following criteria: length higher than or ~100 amino acids, a FUGUE Z score higher than 6 (which corresponds to 99% specificity) with known Glb structures, and the presence of proximal H at position F8.

**Gene mapping and detection of promoter sequences**

Scaffolds containing copies of the glb gene were used for mapping glbs. This included the detection of open reading frames (ORFs) ~5 kb up- and downstream to glbs and ORF length, transcription direction and localization in the +/- strand, Canonical (-10 and -35) and Fnr promoter sequences and Shine-Dalgarno sequences were searched within 130 nucleotides upstream to the rhizobial glb genes either by using the search tool of MS Word or by pairwise sequence alignments using the ClustalX program (http://www.clustal.org/clustal2/).

**Protein sequence alignments and phenetic analysis**

Pairwise and multiple sequence alignments were performed using the ClustalX program. Multiple sequence alignment was manually verified using the procedure described by Kapp et al. based on the myoglobin-fold. A phenogram was constructed from the aligned sequences using the UPGMA method from the ClustalX program. The resulting phenogram was edited using the iTOL program (http://itol.embl.de/).

**Modeling and analysis of the predicted proteins tertiary structure**

The tertiary structure of rhizobial Glbs was modeled using the automated mode of the I-TASSER server (http://zhanglab.ccmb.med.umich.edu/I-TASSER/), which also provided the best structural homologs to the query sequences. Models were edited using the VMD program (http://www.ks.uiuc.edu/Research/vmd/) and Adobe Photoshop software. Distance and dihedral angles of amino acids at the heme prosthetic group were calculated using the distance and dihedral tools of the SwissPDBViewer program (http://spdbv.vital-it.ch) as described by Gopalasubramaniam et al. and Sáenz-Rivera et al., respectively.

**Bacterial growth, cell rupture and spectral analysis**

*Bradyrhizobium japonicum* USDA38 and USDA58 were kindly provided by Drs. Donald Keister and Douglas Jones (United States Department of Agriculture, USA). All reagents were purchased from Sigma-Aldrich (St. Louis MO, USA). Bacterial growth, cell rupture and spectral analysis were obtained from commercial (Sigma-Aldrich) preparations of the sperm whale myoglobin and bovine blood hemoglobin.

**Results and discussion**

Detection of Glb sequences in the genomes of \( \alpha \)- and \( \beta \)-rhizobia

Recently, Vinogradov et al. reported that Glb sequences exist in the genomes of 96 rhizobia. However, this report did not provide the rhizobial Glb sequences or links to rhizobial scaffolds containing the Glb sequences. Hence, we searched in databases (see the Methods section and Table S1) in order to obtain rhizobial Glb sequences for analysis. We selected 62 out of the 96 rhizobial genomes reported by the above authors representing the major rhizobial genera, species and strains, which included \( \alpha \)- and \( \beta \)-rhizobia (i.e., those classified within the \( \alpha \)- and \( \beta \)-proteobacteria, respectively). A total of 197 glb sequences were detected in the 62 rhizobial genomes, corresponding to 7 \( \text{thb} \), 47 \( \text{sdgb} \), 40 \( \text{gcs} \) and 103 \( \text{thbs} \) (4 \( \text{thbs} \) class 1, 56 \( \text{thbs} \) class 2 and 43 \( \text{thbs} \) class 3). Individual Glb nucleotide and polypeptide sequences and links to rhizobial scaffolds containing the Glb sequences are provided in Dataset 1 and Dataset 2, respectively. All the rhizobial genomes analyzed in this work contained with solid ammonium sulphate between 35 and 65% saturation. The resulting pellet was resuspended in 5 ml of 50 mM Na-phosphate buffer (pH 7.2) containing 1 mM EDTA and 1 mM PMSF and dialyzed for 18 h against the same buffer to remove the excess of salts. 0.5 to 1 ml aliquots of the dialyzed solution were used to obtain the dithionite reduced + CO minus dithionite reduced differential spectra in a Beckman DU6 spectrophotometer. Control spectra were obtained from commercial (Sigma-Aldrich) preparations of the sperm whale myoglobin and bovine blood hemoglobin.

**Dataset 1. Globin genes detected in the genomes of rhizobial bacteria**

http://dx.doi.org/10.5256/f1000research.6392.d46189

Globin nomenclature corresponds to the first three binomial (genus and species) letters followed by the strain name, globin type and gene copy number. URLs indicate links to individual glb gene sequences.

**Dataset 2. Predicted Glb polypeptides detected in the genomes of rhizobial bacteria**

http://dx.doi.org/10.5256/f1000research.6392.d46190

Globin nomenclature corresponds to the first three binomial (genus and species) letters followed by the strain name, globin type and gene copy number. URLs indicate links to individual Glb polypeptide sequences.

**Dataset 3. Distance to the heme Fe and orientation of distal, proximal, B10 and CD1 amino acids in the predicted structure of selected rhizobial Glbs (Table S2)**

http://dx.doi.org/10.5256/f1000research.6392.d46191

Structural homologs (including the PDB ID number), amino acids from the structural homologs and values for the structural homologs amino acids to individual rhizobial Glbs are indicated in parenthesis for comparison.
glb sequences, thus indicating that glbs are widespread in rhizobia. However, protoglobin and sensor single domain globin sequences were not detected in the rhizobial genomes. This observation indicates that apparently only the fhb, sdgbs, gcs and thbs lineages evolved within rhizobia.

A distribution analysis showed that most (61) of the rhizobial genomes analyzed in this work contain thbs, either as single thbs (13) or in combination with fhbs, sdgbs and/or gcs (48). Furthermore, one rhizobial genome contained only a gcs and none contained only fhbs and sdgbs and the combinations fhbs + sdgbs, fhbs + gcs and sdgbs + gcs (Figure 1). These observations indicate that in the rhizobia analyzed in this work thbs predominate over other glbs and that in these bacteria fhbs, sdgbs and gcs mostly exist in combination with thbs. Also, analysis of the glb copy number showed that in the rhizobia analyzed in this work fhbs mostly exist as single copy (ranging from one to two copies), sdgbs mostly exist as two copies (ranging from one to four copies), gcs exist as either single or two copies (ranging from one to two copies) and thbs mostly exist as two copies (ranging from one to three copies) although quite a few thbs exist as single copy (Table 1). Thus, apparently rhizobial glbs mostly exist as either single or two copies.

Mapping of glb genes in the rhizobial genomes

The glb genes detected in this work were mapped within the rhizobial genomes in order to identify genes that flank nearby to and could coexpress with glbs. Mapping analysis showed that rhizobial glb copies are located in different scaffolds and that they are not tandemly arrayed. Figure S1A shows that neither ORFs or ORFs coding for hypothetical or non-identified proteins are located nearby most of the rhizobial fhbs genes. However, genes coding for the transcriptional regulator NsrR, 2-nitropropane dioxygenase and NosR, Z, D, F, Y and X are located nearby cupnecN1fhb1, rhilegUPM1137fhb and sinmel1021fhb, respectively. Figure S1B shows that B. elkanii and B. japonicum sdgbs are mostly flanked by genes coding for proteins that function in nitrate/nitrite metabolism and sugar transport. Figure S1C shows that genes coding for proteins that function in chemotaxis are located nearby several rhizobial gcs, although genes coding for a peptide deformylase, sugar and nitrate transport proteins and NAD(P)H nitrate reductase are located nearby some other rhizobial gcs. Figure S1D shows

| glb/no. of copies | No. of genomes |
|-------------------|----------------|
| fhbs              |                |
| 1                 | 5              |
| 2                 | 1              |
| sdgbs             |                |
| 1                 | 5              |
| 2                 | 11             |
| 3                 | 4              |
| 4                 | 2              |
| gcs               |                |
| 1                 | 12             |
| 2                 | 14             |
| thbs              |                |
| 1                 | 22             |
| 2                 | 36             |
| 3                 | 3              |

Figure 1. Venn diagram illustrating the distribution of glb genes in the rhizobial bacteria analyzed in this work. Numbers correspond to rhizobial genomes containing glbs.
that genes flanking the rhizobial *thbs* are rather variable. However, *B. japonicum* *thbs* are often flanked by genes coding for the transcriptional regulator Rieske Fe-S, shikimate kinase and alcohol dehydrogenase; mesorhizobial *thbs* are often flanked by genes coding for permeases and tRNA-Trp, and *R. leguminosarum* *thbs* are often flanked by genes coding for membrane proteins. Thus, if *glb* and flanking genes coexpress in rhizobia, and proteins coded by these genes function within the same metabolic pathways, the above observations suggest that rhizobial *Glbs* could play a variety of roles in rhizobial physiology, including nitrate/nitrite metabolism, transport processes, gene regulation and chemotaxis. Interestingly, with the exception of *simmel1021thb* which is flanked by *nos* and *fix* genes (Figure S1A)^96^, *nif* and *fix* genes coding for proteins that function in N₂-fixation were not detected nearby the rhizobial *glb* genes. This observation suggests that rhizobial *Glbs* might not directly function in N₂-fixation.

**Detection of promoter sequences upstream to the rhizobial *glb* genes**

Identification of promoter sequences is crucial to an understanding of gene regulation and ultimately protein function within the cell’s physiology. Hence, we searched for canonical (-10 and -35) promoters and the O₂- and NO-regulated Fnr promoter^10,11,12^ within 130 nucleotides upstream to 44 selected rhizobial *glb* genes (i.e. those representative of major rhizobial Glb clades identified in this work (see Figure 2)). Also, we searched for Shine-Dalgarno sequences within the same region, which indicate that *Glb* transcripts could be translated into proteins. Results showed that, with the exception of *burphySTM815thb1*, *burphySTM815thb2* and *rhiupHPC(L)thb1*, a -10 promoter is absent upstream of the selected rhizobial *glbs*. In contrast, with the exception of *cupnecN1thb1* and *rhiupHPC(L)thb2*, a -35 promoter exists upstream of the selected rhizobial *glbs*. Searching for Fnr promoter sequences revealed that Fnr-like promoters exist upstream to 30 out of the 44 selected rhizobial *glbs*, including *fhhb*, *sdgb*, *gcs* and *thb* genes. A Shine-Dalgarno sequence was detected upstream to most of the selected rhizobial *glbs* (Table 2). These observations suggest that the -35 promoter is a major canonical promoter that regulates most of the rhizobial *glbs*, that it is likely that several rhizobial *glbs* are regulated by levels of O₂ and NO throughout an FNR mechanism^13,14^ and that rhizobial *Glb* transcripts are translated into proteins.

**Table 2. Position of canonical and Fnr-like promoter sequences and Shine-Dalgarno sequence within 130 nucleotides upstream to selected rhizobial *glb* genes.** Consensus sequences are indicated in parenthesis. Identical and non-identical nucleotides into the Fnr-like promoter sequences to the consensus Fnr promoter sequence are indicated with upper- and lowercase letters, respectively. N.D., non-detected.

|  | **Canonical promoters** | **Fnr promoter** | **Shine-Dalgarno sequence (AGGAGG)** |
|---|------------------------|-----------------|-----------------------------|
|  | -10 promoter | -35 promoter | Sequence | Position |
|  | (TATAAT) | (TTGACA) | (TTTAAGAGGCAAT) |
| *fhbs* | | | |
| burphySTM815thb | N.D. | -36 to -41 | TcTAAgCcGcTggAT | -102 to -115 | -10 to -13 |
| cupnecHPC(L)thb | N.D. | -43 to -48 | N.D. |
| cupnecJMP134thb | N.D. | -46 to -54 | TTtAaAcGgagcc | -5 to -18 | -10 to -15 |
| cupnecN1thb1 | N.D. | -46 to -52 | N.D. |
| cupnecN1thb2 | N.D. | -29 to -36 | aTcAAgGcGgAg | -64 to -77 | -8 to -12 |
| rhiupHPC(L)thb1 | N.D. | -32 to -37 | N.D. |
| simmel1021thb | N.D. | -48 to -56 | gTcAAgGgGCCAa | -12 to -25 | -8 to -12 |
|  |  |  | ggTtgGgGtCCAcT | -61 to -74 |
| *sdgb* | | | |
| azodoeUFLA1-100sdgb | N.D. | -38 to -42 | gccAgGAGtCCGAT | -2 to -15 | -8 to -12 |
| braelkUSDA9sdgb2 | N.D. | -36 to -43 | TtAAGgacatcAT | -114 to -127 | -7 to -11 |
| braelkUSDA1741sdgb2 | N.D. | -34 to -40 | N.D. |
| braelkUSDA3254sdgb1 | N.D. | -61 to -66 | TTTtGgGCaAAT | -71 to -84 | N.D. |
| braelkUSDA3254sdgb2 | N.D. | -40 to -44 | TTTAcGAGGCtgcT | -11 to -24 | -16 to -22 |
| braelkUSDA3259sdgb1 | N.D. | -41 to -46 | TTTcAGAacIcAT | -22 to -35 | -8 to -12 |
| brajapUSDA38sdgb2 | N.D. | -53 to -58 | N.D. |
| brajapUSDA124sdgb1 | N.D. | -60 to -65 | N.D. |
| brajapUSDA124sdgb2 | N.D. | -53 to -58 | N.D. |
| brajapUSDA124sdgb1 | N.D. | -60 to -65 | N.D. |
Sequence alignments and phenetic analysis of rhizobial Glbs

Pairwise sequence alignments showed that the rhizobial fHbs, SDgbs, GCSs and tHbs analyzed in this work are 34.6 to 85.4%, 6.7 to 100%, 10.9 to 100% and 3.5 to 100% identical, respectively. This indicates that variability among the rhizobial Glb sequences is high. Moreover, identity values for the fHbs globin and flavin domains were 39.1 to 93.7% and 26.5 to 81.1%, respectively, and identity values for the GCSs globin and transmitter domains were 17.5 to 100% and 5.9 to 100%, respectively. Thus, apparently in the rhizobial fHbs and GCSs analyzed in this work the globin domain is more conserved than the flavin and transmitter domains.

The average length and molecular mass for the rhizobial fHbs, SDgbs, GCSs and tHbs analyzed in this work are 400 amino acids and 44 kDa, 141 amino acids and 15 kDa, 510 amino acids and 55 kDa and 149 amino acids and 17 kDa, respectively. However, sequence analysis revealed that globin domain from BraelkUSDASDgb2, BraelkUSDASDA94thb1 and BraelkWSM1741SDgb2, RhieltCFN42GCS1, BrajapUSDA4tHb2 and BrajapWSM2793tHb3 contains 27 to 73 extra amino acids at the N-terminal. In contrast, a large deletion comprising helices A and B, CD loop and part
of helix E was detected in the BraelkUSDA94SDgb2 sequence indicating that BraelkUSDA94SDgb2 is 89 amino acids in length (Figure S2).

Multiple sequence alignment showed that, with the exception of 21 GCSs, in the rhizobial Glbs analyzed in this work, the proximal (F8, located at position 322/323 in Figure S2) amino acid to the heme Fe is H. Apparently, in the above rhizobial GCSs, F8 is E. Amino acids other than H occupying the F8 position in bacterial Glbs were previously reported by Vinogradov et al. However, because H F8 is absolutely conserved in Glbs (i.e. from bacteria to mammals), assigning E F8 to rhizobial (and other bacterial) GCSs should be taken with caution as this assignment might result from a sequence alignment artifact.

Ideally, F8 from rhizobial GCSs should be identified by experimental methods, such as x-ray crystallography. Multiple sequence alignment also showed that in the rhizobial Glbs analyzed in this work, the distal (E7, located at position 285/289/290 in Figure S2) amino acid to the heme Fe is Q in fHbs, can be Q/R/K/M/L in SDgbs, Q in GCSs and can be H/F/L/V/R in tHbs. This indicates that distal Q is conserved in rhizobial fHbs and GCSs and that amino acids occupying the distal position in rhizobial SDgbs and tHbs are variable. The B10 and CD1 amino acids (located at positions 257 and 270/271/273 in Figure S2, respectively), which also participate in binding of ligands to the heme Fe, are Y and F in most of the rhizobial Glbs analyzed in this work followed by (in order of abundance) F, S and V and H, I, S and Y, respectively.

Figure 2. Phenetic relationships among Glbs detected in the genomes of rhizobial bacteria. Phnogram was obtained from the Glbs sequence alignment shown in Figure S2. The fHb, SDgb, GCS, tHb class 1, tHb class 2 and tHb class 3 clusters are indicated with light blue, dark blue, red, light green, bright green and dark green, respectively. Stars indicate Glbs selected for the detection of promoter sequences upstream to the glb genes and Glb protein modeling.
A phenogram was constructed from the above multiple sequence alignment. Figure 2 shows that the rhizobial Glbs analyzed in this work segregate into two main lineages: one containing fHbs, SDgbs and GCSs, and the other containing tHbs (the fHb/SDgb/GCS and tHb lineages, respectively). This is consistent with the main evolutionary lineages identified in bacterial Glbs\(^{31,32}\) thus indicating that major evolutionary patterns for rhizobial Glbs were identical to those for other bacterial Glbs. Rhizobial fHbs and GCSs cluster with rhizobial SDgbs within the fHb/SDgb/GCS lineage owing to the similarity between the fHb and GCS globin domains and SDgbs. This has been postulated to be the result of an early divergence from a common ancestor to the bacterial fHb and GCS globin domains and SDgbs\(^{31,32}\). The tHb lineage segregates into rhizobial tHbs class 1, tHbs class 2 and tHbs class 3. Within this lineage the rhizobial tHbs class 3 segregate in ancestral position to the rhizobial tHbs class 1 and tHbs class 2. Also, the bradyrhizobial, azorhizobial, mesorhizobial, rhizobial and burkholderial tHbs class 3 segregate from each other; the segregation within rhizobial, sinorhizobial, mesorhizobial and β-rhizobial tHbs class 2 is rather conserved, and bradyrhizobial tHbs class 2 and class 3 segregate into the B. elkanii and B. japonicum tHb sublineages. These observations indicate that rhizobial tHbs evolved similarly to other bacterial tHbs\(^{31,32}\) and that evolution of rhizobial tHb sublineages was rather conserved.

**Modeling and analysis of the predicted rhizobial Glbs**

**tHb**

Structure elucidation is essential to a full understand of a protein’s function within the cell’s physiology. The structure of a considerable number of bacterial and non-bacterial Glbs has been elucidated by x-ray crystallography. However, with the exception of a S. meliloti fHb whose tertiary structure was predicted using bioinformatics methods\(^{36}\), the structure of rhizobial Glbs is not known. Hence, we used bioinformatics methods to predict and analyze the tertiary structure of 44 selected rhizobial Glbs (i.e. those representative of major rhizobial Glb clades identified in this work (see Figure 2 and Table S2)) using the best structural homologs as templates (Dataset 3).

Predicted structures for selected rhizobial SDgbs and fHbs and GCSs globin domain and tHbs fold into the 3/3- and 2/2-globin fold, respectively (Figure 3 to Figure 8). Figure 3 shows that structures among the predicted rhizobial fHbs are highly similar. Yet major differences were detected in the Bur phySTM815tHb, CupnecHPC(L)tHb and RhilegUMP1137tHb flavin domains, which exhibited two additional helices. Dataset 3 shows that among globin domains from predicted rhizobial fHbs the distance of the proximal H and distal Q to the heme Fe is 1.44 to 2.47 Å and 6.71 to 15.35 Å, respectively. This observation suggests that the heme Fe in rhizobial fHbs is pentacoordinate.

Figure 4 shows that 3/3-globin folding is highly conserved in the predicted structure of the rhizobial SDgbs AzodoeUFLA1-100SDgb, BraelkUSDA3254SDgb2, BraelkUSDA3259SDgb1 and Braja pUSDA38SDgb2. Major variations to 3/3-globin folding from predicted rhizobial SDgbs consisted of the existence of an unusually short helix E in BraelkUSDA94SDgb2, a long helix H in BraelkUS DA3254SDgb1 and BrajapUSDA124SDgb1, and the existence of a pre-helix A followed by a long loop at the N-terminal of Braelk WSM1741SDgb2. Dataset 3 shows that among the predicted rhizobial SDgbs the distance of proximal H and distal Q/R/K/M to the heme Fe is 2.11 to 4.44 Å and 5.08 to 6.63 Å, respectively. This observation suggests that the heme Fe in rhizobial SDgbs is either penta- or hexacoordinate.

Only the globin domain from bacterial GCSs has been crystallized and analyzed by x-ray crystallography\(^{33,34}\) (Dataset 3). Crystal structure for the bacterial GCSs transmitter domain has not been elucidated. Hence, we only predicted and analyzed the tertiary structure of globin domains from the selected rhizobial GCSs. Figure 5 shows that the predicted rhizobial GCSs globin domain exhibits a 1.5- to 3-turn pre-helix A, that (with the exception of SinfreGR64GCS) no loop exists between helices A and B, and that helix H is unusually long in Rhi teli83GCS, Rhi telCIAT652GCS2 and RhilegGB30GCS2. Dataset 3 shows that among the predicted rhizobial GCSs globin domain distance of proximal H/E and distal Q to the heme Fe is 1.77 to 5.56 Å and 4.09 to 9.04 Å, respectively. This observation suggests that the heme Fe in the rhizobial GCSs globin domain is either penta- or hexacoordinate.

Figure 6 to Figure 8 show that 2/2-globin folding is highly conserved in the predicted rhizobial tHbs class 1, class 2 and class 3. Major variations to 2/2-globin folding from predicted rhizobial tHbs consisted of the existence of a 2.5-turn pre-helix A followed by a long loop at the N-terminal of (class 1) CupnecN1tHb1 (Figure 6); the existence of a one-turn pre-helix F (designated as φ in Figure 7) in the rhizobial tHbs class 2; the existence of a long and extended C-terminal region in (class 2) BraelkUSDA94tHb1 (Figure 7), and the substitution of helix A by a long loop that connects to helix B through a 1- to 2.5-turn pre-helix B in (class 3) BraelkUSDA76tHb2, BrajapUSDA123tHb1, Bur phySTM815tHb1, Meslo tNZP2037tHb2 and Sin mel1021tHb2 (Figure 8). Dataset 3 shows that among the predicted rhizobial tHbs, the distance of proximal H and distal H/L/F to the heme Fe is 1.77 to 7.51 Å and 4.09 to 8.25 Å, respectively. This observation suggests that the heme Fe in the rhizobial tHbs is either penta- or hexacoordinate.

The above observations suggest that in spite of sequence variability (see the *Sequence alignments and phenetic analysis of rhizobial Glbs* subsection) the structure of rhizobial Glbs is similar to the canonical 3/3- or 2/2-globin folding of bacterial and non-bacterial Glbs. However, a number of predicted rhizobial Glbs exhibited variations at the N- and C-terminal regions suggesting that their structural properties could be different to those of canonical Glbs.
Figure 3. Predicted structure of rhizobial fHbs (blue) overlapped to structural homologues (green). Structural homologues are indicated in Dataset 3. Distal and proximal amino acids to the heme Fe and amino acids that interact with the FAD cofactor are shown in brown. Heme and FAD are shown in red and yellow, respectively. Helices within the globin domain are indicated with letters A to H. All structures are displayed in the same orientation.
Figure 4. Predicted structure of selected rhizobial SDgbs (blue) overlapped to structural homologues (green). Structural homologues are indicated in Dataset 3. Distal and proximal amino acids to the heme Fe are shown in brown. Heme is shown in red. Helices are indicated with letters A to H. All structures are displayed in the same orientation.
Figure 5. Predicted structure of selected rhizobial GCSs globin domain (blue) overlapped to structural homologues (green). Structural homologues are indicated in Dataset 3. Distal and proximal amino acids to the heme Fe are shown in brown. Heme is shown in red. Helices are indicated with letters A to H. All structures are displayed in the same orientation.

Figure 6. Predicted structure of class 1 CupnecN1tHb1 (blue) overlapped to the structural homologue *Tetrahymena pyriformis* tHb (PDB ID 3AQ5) (green). Distal and proximal amino acids to the heme Fe are shown in brown; only potential distal E11 is shown in the CupnecN1tHb1 structure. Heme is shown in red. Helices are indicated with letters A to H.
Figure 7. Predicted structure of selected rhizobial thbs class 2 (blue) overlapped to structural homologues (green). Structural homologues are indicated in Dataset 3. Distal and proximal amino acids to the heme Fe are shown in brown; only potential distal E11 is shown in the thbs structure. Heme is shown in red. Helices are indicated with letters A to H. Pre-helix F is indicated with the Greek letter $\phi$. All structures are displayed in the same orientation.
Figure 8. Predicted structure of selected rhizobial tHbs class 3 (blue) overlapped to structural homologues (green). Structural homologues are indicated in Dataset 3. Distal and proximal amino acids to the heme Fe are shown in brown; only potential distal E11 is shown in the tHbs structure. Heme is shown in red. Helices are indicated with letters A to H. All structures are displayed in the same orientation.
Data also shows that (with few exceptions) in addition to proximal and distal amino acids the distance of B10 and CD1 amino acids to the heme Fe and the orientation of proximal, distal, B10 and CD1 amino acids are similar within and among the predicted rhizobial SDgbs, fHbs and GCSs globin domain and tHbs. These amino acids participate in the binding of ligands to the heme Fe. Thus, these observations suggest that the mechanisms and chemistry for ligand binding are similar among the rhizobial Glbs.

Spectroscopic identification of putative Glbs in soluble extracts from *Bradyrhizobium japonicum* USDA38 and USDA58

The prerequisites for being able to infer a protein’s function are isolating and characterizing either native or recombinant proteins and detecting protein synthesis in *vivo*. No rhizobial Glb has been isolated and characterized thus far. However, spectroscopic evidence indicates that putative Glbs exist in soluble extracts from *B. japonicum* 505 (Wisconsin), *R. leguminosarum* bv. viciae, *B. japonicum* NPK63 and *R. etli* CE3 (see the Introduction section). In order to extend these analyses to other rhizobia, we analyzed soluble extracts from *B. japonicum* USDA38 and USDA58 by (dithionite reduced + CO minus dithionite reduced) differential spectroscopy using as controls the sperm whale myoglobin and bovine blood hemoglobin. Table 3 shows that absorption peaks and troughs in the Soret and Q regions for the *B. japonicum* USDA38 and USDA58, *B. japonicum* 505 (Wisconsin), *R. leguminosarum* bv. viciae, *B. japonicum* NPK63 and *R. etli* CE3 soluble extracts, *Vitreoscilla* VHb, *E. coli* K12 Hmp, sperm whale myoglobin and bovine blood hemoglobin are nearly identical. This preliminary evidence indicates that putative soluble Glbs are synthesized in *B. japonicum* USDA38 and USDA58. Interestingly, genes coding for SDgbs (*brajapUSDA38SDgb1* and *brajapUSDA38SDgb2*) and tHbs (*brajapUSDA38tHb1* and *brajapUSDA38tHb2*) were identified in the *B. japonicum* USDA38 genome (*Dataset 1*). Thus, it is likely that putative *B. japonicum* USDA38 Glbs corresponds to a combination of SDgbs and tHbs. Inferences from the preliminary results reported here should be confirmed by Glb detection, isolation and unequivocal identification after protein sequencing. This may open the possibility to carry out further experimental analyses on rhizobial Glbs.

| Rhizobial soluble extract/Glb | Soret region | | Q region | | Reference |
|------------------------------|-------------|---|---|---|---|
|                              | Peak (nm)   | Trough (nm) | Peak (nm) | Trough (nm) | |
| Rhizobial soluble extracts    |             |             |             |             | |
| *B. japonicum* USDA38        | 425         | 448         | 535         | 573         | 549 | 600 | This work |
| *B. japonicum* USDA58        | 416         | 437         | 535         | 573         | 554 | 601 | This work |
| *B. japonicum* NPK63         | 422         | 443         | 529         | 574         | 558 | 598 | This work |
| *B. japonicum* 505 (Wisconsin)| 417        | 434         | 540         | 569         | 556 | n.i. | 22 |
| *R. etli* CE3                | 421         | 439         | 539         | 563         | 547 | 590 | 25 |
| *R. leguminosarum* bv. viciae96 | 424       | 443         | 535         | 574         | 555 | n.i. | 23 |
| Bacterial Glbs               |             |             |             |             | |
| *Vitreoscilla* VHb            | 418         | 436         | 534         | 567         | 551 | 590 | 25 |
| *E. coli* K12 Hmp            | 420         | 437         | 530         | 570         | 555 | 592 | 55 |
| Vertebrate Glbs              |             |             |             |             | |
| Sperm whale myoglobin        | 419         | 436         | 538         | 578         | 558 | 596 | This work |
| Bovine blood hemoglobin      | 417         | 432         | 533         | 570         | 554 | 588 | This work |

n.i., non-identified
Conclusions
Rhizobial Glbs have been poorly studied. However, results reported in this work provide molecular and biochemical data from a bioinformatics perspective that contribute to a better understanding of these proteins. For example, the distribution and outline for the evolution of glb genes and Glb proteins among rhizobia was clarified, genes that could coexpress with the rhizobial glbs were identified and the predicted tertiary structure for rhizobial Glbs was elucidated. Also, spectroscopic analysis suggested that soluble Glbs are synthesized in free-living *B. japonicum* USDA38 and USDA58. This information will be useful in designing future experimental work focused on clarifying Glb functions within the physiology of free-living and symbiotic rhizobia.

Data availability
F1000Research: Dataset 1. Globin genes detected in the genomes of rhizobial bacteria. 10.5256/f1000research.6392.d46189

F1000Research: Dataset 2. Predicted Glb polypeptides detected in the genomes of rhizobial bacteria. 10.5256/f1000research.6392.d46190

F1000Research: Dataset 3. Distance to the heme Fe and orientation of distal, proximal, B10 and CD1 amino acids in the predicted structure of selected rhizobial Glbs (Table S2). 10.5256/f1000research.6392.d46191

Supplementary materials
Supplementary File S1. Mapping of the *fhb* (A), *sdgb* (B), *gcs* (C) and *thb* (D) genes in the genomes of rhizobial bacteria. DNA fragments correspond to ~5 kb up- and downstream to *glb* genes. Arrows indicate the transcription orientation. The *glb* genes are shown in red, predicted polypeptides functioning in nitrogen metabolism are shown in blue and predicted polypeptides functioning in chemotaxis are shown in green. The ORF sizes and distances between ORFs are shown at an approximate scale. Abbreviations for the predicted polypeptides are indicated at the end of the figure. Location of DNA fragments in the rhizobial genome: Chr, chromosome; Pmd, plasmid; n.i., non-identified.

Supplementary File S2. Sequence alignment of Glbs detected in the genomes of rhizobial bacteria. The tHb, fHb, SDgb and GCS sequences are shown in green, light blue, dark blue and red, respectively. Distal and proximal amino acids located in helices E and F, respectively, are indicated within black boxes; potential distal E7 and E11 are indicated in the tHb sequences. Amino acids that interact with FAD and NAD(P)+ cofactors in the fHb flavin domain are indicated within gray boxes. Limits for the globin domain are indicated with right- and left-oriented arrows within black circles. Helices are indicated with letters A to H within the 2/2- and 3/3-fold of the tHb and SDgb and fHb and GCS globin domains, respectively. Outgroups for fHbs correspond to Bacs3Hb, Escc3Hb and SaccerHb (Genbank accession number YP_003865693, NP_289108 and NP_011750, respectively); outgroup for SDgbs corresponds to VitSDgb (Genbank accession number AAA75506); outgroups for GCSs correspond to AgrtumGCS and Bacs3GCS (Genbank accession number NP_354049 and NP_388919, respectively); outgroups for tHbs correspond to Myc2tHb class1, Myc3tHb class 2, AgrtumtHb class 2 and Myc3tHb class 3 (Genbank accession number NP_216058, NP_216986, WP_020813663 and BAN32501, respectively).

Author contributions
RGB, MSS and RAP conceived the study. RGB and MSS executed the experiments. RAP prepared the first draft of the manuscript. RGB, MSS and RAP revised the draft manuscript and have agreed to the final content.

Competing interests
No competing interests were disclosed.

Grant information
This work was partially financed by SEP-PROMEP (grant number UAEMor-PTC-01-01/PTC23) and Consejo Nacional de Ciencia y Tecnología (CoNaCyT grant numbers 25229N and 42873Q), México. R. Gesto-Borroto is a graduate student financially supported by CoNaCyT (registration no. 293307).

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
We thank Drs. Donald Keister and Douglas Jones (United States Department of Agriculture, USA) for kindly providing the *Bradyrhizobium japonicum* USDA38 and USDA58 strains, and Dr. Serge N. Vinogradov (Wayne State University, Detroit MI, USA) for critical reading of this article and providing constructive comments.
Table S1. Classification of the α- and β-rhizobia and rhizobial genomes analyzed in this work.

| Family                          | Genus                    | Species/strain                                                                                               | Database                                      |
|---------------------------------|--------------------------|--------------------------------------------------------------------------------------------------------------|-----------------------------------------------|
| **α-rhizobia**                  |                          |                                                                                                             |                                               |
| **Domain Eubacteria**           |                          |                                                                                                             |                                               |
| **Phylum Proteobacteria**       |                          |                                                                                                             |                                               |
| **Class Alphaproteobacteria**   |                          |                                                                                                             |                                               |
| **Order Rhizobiales**           |                          |                                                                                                             |                                               |
| **Family**                      | **Genus**                | **Species/strain**                                                                                            | **Database**                                  |
| Bradyrhizobiaceae               | Bradyrhizobium           | B. elkanii USD3254, 587, USDA3259, USDA76, USDA94, WSM1741, WSM2783                                        | JGI Genome Portal ([http://genome.jgi.doe.gov/](http://genome.jgi.doe.gov/)) |
|                                 |                          | B. japonicum USDA110, 22, USDA122, USDA123, USDA124, USDA 135, USDA38, USDA4, USDA6-7488, USDA6-8372, WSM1743, WSM2793, in8p8, is5 | Rhizobase ([http://genome.microbedb.jp/rhizobase](http://genome.microbedb.jp/rhizobase)) |
| Phyllobacteriaceae              | Mesorhizobium            | M. ciceri CM06, WSM4083, bv. biserrulae WSM1271 M. loti CJ3sym, MAFF303099, NZP2037, R7A, R88b, USDA3471 | JGI Genome Portal ([http://genome.jgi.doe.gov/](http://genome.jgi.doe.gov/)) |
|                                 |                          |                                                                                                             | Rhizobase ([http://genome.microbedb.jp/rhizobase](http://genome.microbedb.jp/rhizobase)) |
| Rhizobiaceae                    | Rhizobium                | R. etli BC-3, Brasil 5, CFN42, CIAT894, CNNAF512, GR56, IE4771, Kim5 R. leguminosarum bv. viciae UPM1137, bv. viciae UPM1137, bv. viciae GB30, bv. viciae UF39, bv. viciae WSM1455, bv. viciae WSM1481, bv. viciae 248, bv. viciae 3841, bv. viciae Ps8, bv. viciae TdM, bv. viciae Vc3, bv. viciae Vc2 R. lupini HPC(L) | CCG-UNAM ([http://www.cifn.unam.mx/](http://www.cifn.unam.mx/)) |
|                                 |                          |                                                                                                             | Sanger Institute ([http://www.sanger.ac.uk](http://www.sanger.ac.uk)) |
|                                 |                          |                                                                                                             | NEERI ([http://www.neeri.res.in/](http://www.neeri.res.in/)) |
| Sinorhizobium                   | S. fredii               | GR64, HH103, USDA257 S. melloti 1021                                                                         | GGL ([http://appmibio.uni-goettingen.de/](http://appmibio.uni-goettingen.de/)) |
|                                 |                          |                                                                                                             | INRA ([https://iant.toulouse.inra.fr](https://iant.toulouse.inra.fr)) |
| Xanthobacteriaceae              | Azorhizobium            | A. doebereinerae UFLA1-100                                                                                   | JGI Genome Portal ([http://genome.jgi.doe.gov/](http://genome.jgi.doe.gov/)) |
| **β-rhizobia**                  |                          |                                                                                                             |                                               |
| **Domain Eubacteria**           |                          |                                                                                                             |                                               |
| **Phylum Proteobacteria**       |                          |                                                                                                             |                                               |
| **Class Betaproteobacteria**    |                          |                                                                                                             |                                               |
| **Order Burkholderiales**       |                          |                                                                                                             |                                               |
| **Family**                      | **Genus**                | **Species/strain**                                                                                            | **Database**                                  |
| Burkholderiaceae                | Burkholderia            | B. phytophilum STM815                                                                                       | JGI Genome Portal ([http://genome.jgi.doe.gov/](http://genome.jgi.doe.gov/)) |
| Cupriavidus                     | C. necator              | HPC(L), JMP134, N-1 ATCC43291                                                                                | Goettingen Genomics Laboratory ([http://appmibio.uni-goettingen.de/](http://appmibio.uni-goettingen.de/)) |

*Formerly classified as *Alcaligenes eutrophus* and *Ralstonia eutropha*. 

*β-rhizobia*
Table S2. Predicted structures of rhizobial Glbs deposited in the Caspur protein model dataBase (http://bioinformatics.cineca.it/PMDB/).

| Glb                | ID number        |
|--------------------|------------------|
| **fHbs**           |                  |
| BurphySTM815fHb    | PM0079658        |
| CupnecHPC(L)fHb    | PM0079659        |
| CupnecJMP134fHb    | PM0079660        |
| CupnecN-1 ATCC43291fHb1 | PM0079661 |
| CupnecN-1 ATCC43291fHb2 | PM0079662 |
| RhilegUPM1137fHb   | PM0079663        |
| Sinmel1021fHb      | PM0079672        |
| **SDgbs**          |                  |
| AzodoeUFLA1-100SDgb | PM0079664       |
| BraelkUSDA94SDgb2  | PM0079665        |
| BraelkUSDA3254SDgb1 | PM0079666      |
| BraelkUSDA3254SDgb2 | PM0079667      |
| BraelkUSDA3259SDgb1 | PM0079668      |
| BraelkWSM1741SDgb2 | PM0079669        |
| BrajapUSDA38SDgb2  | PM0079670        |
| BrajapUSDA124SDgb1 | PM0079671        |
| **GCSs**           |                  |
| Brajapin8p8GCS globin domain | PM0079673 |
| RhietCIAT652GCS1 globin domain | PM0079674 |
| RhietCIAT652GCS2 globin domain | PM0079675 |
| RhietI8C-3GCS globin domain | PM0079676 |
| RhietCFN42DSM 11541GCS1 globin domain | PM0079677 |

| Glb                | ID number        |
|--------------------|------------------|
| RhilegGB30GCS1 globin domain | PM0079678 |
| RhilegGB30GCS2 globin domain | PM0079679 |
| SinfreGR64GCS globin domain | PM0079680 |
| Sinmel1021GCS globin domain | PM0079681 |
| **tHbs**           |                  |
| AzodoeUFLA1-100tHb1 | PM0079682       |
| AzodoeUFLA1-100tHb2 | PM0079683       |
| BraelkUSDA76tHb2    | PM0079684       |
| BraelkUSDA94tHb1 (globin domain + C-terminal extension) | PM0079701 |
| BrajapUSDA38tHb2    | PM0079685       |
| BrajapUSDA123tHb1   | PM0079686       |
| BurphySTM815tHb1    | PM0079687       |
| BurphySTM815tHb2    | PM0079688       |
| CupnecN-1 ATCC43291tHb1 | PM0079689 |
| CupnecN-1 ATCC43291tHb2 | PM0079690 |
| MescicCMG6tHb       | PM0079691       |
| MeslotNZP2037tHb2   | PM0079692       |
| RhietCNPAF512tHb    | PM0079693       |
| RhietKim5tHb        | PM0079694       |
| RhilegGB30tHb2      | PM0079695       |
| RhilegVC2tHb1       | PM0079696       |
| RhilupHPC(L)tHb1    | PM0079697       |
| RhilupHPC(L)tHb2    | PM0079698       |
| SinfreH1103tHb      | PM0079699       |
| Sinmel1021tHb2      | PM0079700       |
45. Vinogradov SN, Fernández I, Hoogewijs D, et al.: Phylogenetic relationships of 3/3 and 2/2 hemoglobins in Archaeplastida genomes to bacterial and other eukaryote hemoglobins. Mol Plant. 2011; 4(1): 42–58. PubMed Abstract | Publisher Full Text
46. Vinogradov SN, Waltz DA, Pohajdak B, et al.: Adventitious variability? The amino acid sequences of nonvertebrate globins. Comp Biochem Physiol B. 1993; 106(1): 1–26. PubMed Abstract | Publisher Full Text
47. Weber RE, Vinogradov SN: Nonvertebrate hemoglobins: functions and molecular adaptations. Physiol Rev. 2001; 81(2): 569–628. PubMed Abstract
48. Gardner AM, Martin LA, Gardner PR, et al.: Steady-state and transient kinetics of Escherichia coli nitric-oxide dioxygenase (flavohemoglobin). The B10 tyrosine hydroxyl is essential for dioxygen binding and catalysis. J Biol Chem. 2000; 275(17): 12581–12589. PubMed Abstract | Publisher Full Text
49. Igarashi J, Kobayashi K, Matsuoka A: A hydrogen-bonding network formed by the B10–E7–E11 residues of a truncated hemoglobin from Tetrahymena pyriformis is critical for stability of bound oxygen and nitric oxide detoxification. J Biol Inorg Chem. 2011; 16(4): 599–609. PubMed Abstract | Publisher Full Text
50. Ouettel Y, Millani M, Couture M, et al.: Ligand interactions in the distal heme pocket of Mycobacterium tuberculosis truncated hemoglobin N: roles of TyrB10 and GinE11 residues. Biochemistry. 2006; 45(29): 8770–8781. PubMed Abstract | Publisher Full Text
51. Vinogradov SN, Hoogewijs D, Bailly X, et al.: A model of globin evolution. Gene. 2007; 398(1–2): 132–142. PubMed Abstract | Publisher Full Text
52. Wu G, Wainwright LM, Poole RK: Microbial globins. Adv Microb Physiol. 2003; 47: 255–310. PubMed Abstract | Publisher Full Text
53. Pesce A, Thijs L, Nardini M, et al.: HisE11 and HisF8 provide bis-histidyl heme hexa-coordination in the globin domain of Geobacter sulfurreducens globin-coupled sensor. J Mol Biol. 2009; 386(1): 246–260. PubMed Abstract | Publisher Full Text
54. Zhang W, Phillips GN Jr: Structure of the oxygen sensor in Bacillus subtilis: signal transduction of chemotaxis by control of symmetry. Structure. 2003; 11(9): 1097–1110. PubMed Abstract | Publisher Full Text
55. Vasudevan SG, Armarego WL, Shaw DC, et al.: Isolation and nucleotide sequence of the hmp gene that encodes a haemoglobin-like protein in Escherichia coli K-12. Mol Gen Genet. 1991; 226(1–2): 49–58. PubMed Abstract | Publisher Full Text
56. Gesto-Borroto R, Sánchez-Sánchez M, Arredondo-Peter R: Dataset 1 in: A bioinformatics insight to rhizobial globins: gene identification and mapping, polypeptide sequence and phenetic analysis, and protein modeling. F1000Research. 2015. Data Source
57. Gesto-Borroto R, Sánchez-Sánchez M, Arredondo-Peter R: Dataset 2 in: A bioinformatics insight to rhizobial globins: gene identification and mapping, polypeptide sequence and phenetic analysis, and protein modeling. F1000Research. 2015. Data Source
58. Gesto-Borroto R, Sánchez-Sánchez M, Arredondo-Peter R: Dataset 3 in: A bioinformatics insight to rhizobial globins: gene identification and mapping, polypeptide sequence and phenetic analysis, and protein modeling. F1000Research. 2015. Data Source
Open Peer Review

Current Referee Status:  

Version 1

Referee Report 08 July 2015

doi:10.5256/f1000research.6858.r9023

Paul Twigg
Department of Biology, University of Nebraska at Kearney, Kearney, NE, USA

I find this paper by Gesto-Borroto et al. to be well written and analyzed. This paper fills a gap in the knowledge base for what is known about bacterial or more specifically rhizobial globins. The authors seem to have taken great care to analyze all available globins from various rhizobial species and biovars. I have no major revisions for this work. It is comprehensive and demonstrates some interesting points about globins in rhizobia. The coexistence of various globins in the bacteria begs questions about their control and functions that undoubtedly other researchers will address. The work with the promoter analysis was particularly interesting to me indicating more than once that the globin expression is likely tied to nitrate/nitrogen metabolism. The absence of the -10 promoter area was also unexpected. The amino acid sequence alignment data also shows interesting information about the conserved positions in the sequence. The phylogenetic relationships are also interesting and appropriately analyzed. The structural modeling is also well done and reveals interesting points about the structure and function of the various globins. Lastly, the authors back up some of their proposals with spectroscopic data from globin extracts. Again, overall I thought that the work was well done and needs no major revisions.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 08 Jul 2015

Raul Arredondo-Peter, Universidad Autonoma del Estado de Morelos, Mexico

We thank Dr. Twigg for evaluating this article and constructive comments.

Competing Interests: No competing interests were disclosed.

Referee Report 26 May 2015

doi:10.5256/f1000research.6858.r8649

Manuel Becana
Department of Plant Nutrition, Spanish National Research Council, Zaragoza, Spain
This is a well-written paper on a subject of great interest. There is very little information of rhizobial globins and the authors have done a good job by systematically analyzing the composition of globin genes of 62 genomes in various genera, species and biovars of rhizobia. The authors are experts in the phylogeny and evolution of plant hemoglobins, and I have no major comments to improve this work. It will nevertheless be of interest for future work to address the issue of why several types of hemoglobins coexist in rhizobia. For example, both truncated hemoglobins and flavohemoglobins seem to be present within the same species and strain, although this would have to be verified by identifying the proteins themselves rather than by only gene sequencing or by analyzing differential spectra (reduced + CO vs reduced) in bacterial extracts. Both classes of hemoglobins have been proposed to act as modulators of NO concentration, but they are unlikely to have redundant functions. An interesting, additional aspect of the work is the mapping analysis, including the report of flanking sequences. This hints to a role of at least some rhizobial globins in nitrogen metabolism. This observation is very timing because of the recent discovery that truncated hemoglobins of Chlamydomonas regulate nitrate reductase.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

**Competing Interests:** No competing interests were disclosed.

---

Author Response 26 May 2015

**Raul Arredondo-Peter**, Universidad Autonoma del Estado de Morelos, Mexico

We thank Dr. Becana for evaluating this article and constructive comments.

**Competing Interests:** No competing interests were disclosed.