Biochimica et Biophysica Acta 1834 (2013) 1789–1800

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Bacterial and archaeal globins — A revised perspective ☆

Serge N. Vinogradov [a], Mariana Tinajero-Trejo [b], Robert K. Poole [b], David Hoogewijs [c,*]

[a] Department of Biochemistry and Molecular Biology, Wayne State University School of Medicine, Detroit, MI 48201, USA
[b] Institute of Biology and Biotechnology, Department of Molecular Biology and Biotechnology, The University of Sheffield, Sheffield, United Kingdom
[c] Institute of Physiology and Zürich Center for Integrative Human Physiology, University of Zürich, Zürich, Switzerland

ARTICLE INFO

Article history:
Received 14 December 2012
Received in revised form 12 February 2013
Accepted 16 March 2013
Available online 27 March 2013

Keywords:
Bacteria
Archaea
Hemoglobin
Prokaryote

ABSTRACT

A bioinformatics survey of putative globins in over 2200 bacterial and some 140 archaeal genomes revealed that over half the bacterial and approximately one fifth of archaeal genomes contain genes encoding globins that were classified into three families: the M (myoglobin-like), and S (sensor) families all exhibiting the canonical 3/3 myoglobin fold, and the T family (truncated myoglobin fold). Although the M family comprises 2 subfamilies, flavohemoglobins (FHbs) and single domain globins (SDgbs), the S family encompasses chimeric globin-coupled sensors (GCSs), single domain Pgbs (protoglobins) and SSDgbs (sensor single domain globins). The T family comprises three classes TrHb1s, TrHb2s and TrHb3s, characterized by the abbreviated 2/2 myoglobin fold. The Archaea contain only Pgbs, GCSs and TrHb1s. The smallest globin-bearing genomes are the streamlined genomes (~1.3 Mbp) of the SAR11 clade of alphaproteobacteria and the slightly larger (ca. 1.7 Mbp) genomes of Aquificae. The smallest genome with members of all three families is the 2.3 Mbp genome of the extremophile *Methylacidiphilum infernorum* (Verrumicrobia). Of the 147 possible combinations of the eight globin subfamilies, only 83 are observed. Although binary combinations are infrequent and ternary combinations are rare, the FHb + TrHb2 combination is the most commonly observed. Of the possible functions of bacterial globins we discuss the two principal ones — nitric oxide detoxification via the NO dioxygenase or denitrosylase activities and the sensing of oxygen concentration in the environmental niche. In only few cases has a physiological role been demonstrated *in vivo*. This article is part of a Special Issue entitled: Oxygen Binding and Sensing Proteins.

1. Introduction

The first accounts of haemoglobins in microbes were those of Keilin in the 1930s, who used a hand spectroscope, in which absorbance bands of a cell suspension are seen as black bands at characteristic wavelengths superimposed on a spectrum of visible colours [1–3]. This work lay fallow until B. Chance et al. in the 1970s investigated in biochemical detail the yeast protein and showed it to comprise not only a haem domain but also a flavin chromophore [4,5]; this was the first description of a flavohaemoglobin.

Recognition of haemoglobins in bacteria was slow to arrive. Dale Webster published several seminal papers in the 1970s describing a 'soluble cytochrome *o*' in the microaerobic bacterium *Vitreoscilla* [6–9]. The haem protein (variously named as Vgb, VHb etc.) was observed in intact cells as a stable oxygenated species [4–6]. The sequence of this protein [10] provided the first definitive proof that haemoglobins occur also in bacteria [11]. The first nucleotide sequence of a bacterial globin was that of Hmp in *E. coli*, the bacterial homologue of Keilin's and Chance's yeast protein. The sequence [12] clearly revealed the chimeric nature of the protein with an N-terminal globin domain and a C-terminal reductase domain with binding sites for NAD(P)H and FAD. More detailed accounts of the development of our now-extensive knowledge of bacterial flavohaemoglobins can be found elsewhere [13,14].

Concomitantly, a shorter than usual 118-residue globin, GlbN, was observed in *Nostoc commune* (Cyanobacteria: Nostocales) [15] and related cyanobacteria *Synechococcus* and *Synechocystis* [16]. Furthermore, a haemoprotein with kinase activity was found encoded by the oxygen sensor of *Rhizobium meliloti* [17]. Additional genomic information revealed the presence of single domain globins homologous to the globin domain of FHbs [14,18] and of a family of globin-coupled sensors involved in chemotaxis in Archaea and Bacteria [19,20] and related single domain protoglobins [21].

A survey of putative globins in some 250 bacterial and ca. 40 archaeal genomes showed that they all could be classified into one of three families and two structural classes [22,23]. The myoglobin-like family comprises chimeric (flavohemoglobin) and single domain

Abbreviations: Adgb, androglobin; Cygb, cytoglobin; FHb, flavohaemoglobin; GbE, globin E; GbX, globin X; GbY, globin Y; GCS, globin-coupled sensor; Hb, haemoglobin; HGT, horizontal gene transfer; LECA, Last Universal Eukaryote Common Ancestor; Mb, myoglobin; Ngb, neuroglobin; Pgb, protoglobin; SDgb, single domain 3/3 globin related to the N-terminal of FHbs; SSDgb, sensor single domain 3/3 globin related to the N-terminal of GCSs.

☆ This article is part of a Special Issue entitled: Oxygen Binding and Sensing Proteins.

* Corresponding author at: Institute of Physiology, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland. Tel.: +41 44 635 5075; fax: +41 44 635 6814.
E-mail address: david.hoogewijs@access.uzh.ch (D. Hoogewijs).

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.03.021

globins as well as the sensor family comprising globin-coupled sensors (GCS), protoglobins (Pgbs), and sensor single domain globins (SSDgbs) all exhibiting the canonical secondary structure of Hbs, the 3/3 Mb-fold. The third family comprises the truncated globins consisting of three subgroups, all sharing an abbreviated 2/2 globin fold, with a vestigial or absent helix A and a loop substituting for helix F [24–26].

2. Material and methods

2.1. Identification of globin sequences

The identification of putative globins and globin domains was carried out using two approaches. In one, we employed the globin gene assignments provided on the SUPERFAMILY site (http://supfam.mrc-lmb.cam.ac.uk), based on a library of hidden Markov models [27], discarded sequences shorter than 100 aa and subjected the remaining sequences to a FUGUE search [28] (www-cryst.bioc.cam.ac.uk). FUGUE scans a database of structural profiles, calculates the sequence–structure compatibility scores for each entry, using environment-specific substitution tables and structure-dependent gap penalties, and produces a list of potential homologs and alignments. FUGUE assesses the similarity between the query and a given structure via the Z score, the number of standard deviations above the mean score obtained by chance: the default threshold Z = 6.0 corresponds to 99% probability [28].

We also performed BLASTP and PSIBLAST searches with pairwise alignment [29] of the NCBI non-redundant protein sequence database (www.ncbi.nlm.nih.gov/BLAST/), using known globins as queries. PSIBLAST searches were carried out using the improved version with composition-based statistics [30]. The match between the query sequence and each of the sequences in a database is assessed by the Expect value (E), which is related to the probability of finding a false positive. Thus, E = 0.1 signifies that the probability of finding by chance, in a given database, another match with the query sequence having the same score, is 1 in 10. Iterations were carried out until all the sequences comprising the desired query group had E > 0.005, the default threshold. Recognition is defined to be a hit with E > 0.005, and with the pairwise alignment fulfilling the following two criteria: proper alignment of the F8 His residues and of helices BC through G.

2.2. Multiple sequence alignments and phylogenetic analysis

Multiple sequence alignments were carried out using MUSCLE [31], COBALT [32], MAFFT [33] PROBCONS [34] and TCOFFEE [35]. The quality of the alignments was assessed by MUMSA [36]. Neighor-joining (NJ) analyses were performed using MEGA version 5.05 [37]. Distances were corrected for superimposed events using the Poisson method. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (pairwise deletion option). The reliability of the branching pattern was tested by bootstrap analysis with 1000 replications. Bayesian inference trees were obtained employing MrBayes version 3.1.2 [38], assuming the WAG model of amino acid substitution and a gamma distribution of evolutionary rates, as determined by the substitution model testing option in MEGA 5.05. Two parallel runs, each consisting of four chains were run simultaneously for 8×10^6 generations and trees were sampled every 100 generations generating a total of 80,000 trees. The final average standard deviations of split frequencies were stationary and posterior probabilities were estimated on the final 50,000 trees. The CIPRES web portal was used for the Bayesian analyses [39] and MEGA version 5.05 [37] was used to visualize these analyses.

3. Results

3.1. Distribution of globins in bacterial and archaeal phyla

A list of >2000 bacterial genomes included in our survey is provided in Appendix A Supplementary Table A1. Globins were found in 1185 out of 2275 bacterial genomes (52%) and in 32 of 140 archaeal genomes (23%). Overall, all the globins found in this survey belong to one of the three families found earlier [22,23], along with the eight subfamilies. Table 1 lists the three families, the eight subfamilies and the estimated number of sequences; a small subset (1–5%) of sequences from each of the eight subfamilies lack the F8 His, the essential determinant of a globin, even though their FUGUE scores are >6. Fig. 1 shows a Bayesian phylogenetic tree of about 150 manually selected representative globins, illustrating clearly the presence of three families and eight subfamilies.

A summary of the distribution of the eight globin subfamilies among the major bacterial and archaeal phyla is provided in Tables 2 and 3, respectively. The highest percent of globin-containing genomes are found in the Aquificae (100%) and the Deinococcus/Thermus group (88%). About 75% of alpha-, beta- and gamma-proteobacterial genomes and of planctomycete genomes have globins. Only about one half of the genomes in Cyanobacteria, delta- and epsilon-proteobacteria have globins. Fewer globins are found in Archaea (see Table 3). They are limited to Pgbs and SSDgbs in a few Crenarchaeota and to GCSs, Pgbs and TrHb1s in the Halobacteria (Euryarchaeota). The decrease in the percent of globin-containing bacterial genomes from 65% to 53% found earlier [22] can be accounted for by the recent infusion of genomes from the Human Microbiome Project [40], the majority of which do not have globins.

3.2. How complete is the bacterial species coverage?

An appropriate question to ask at this juncture would be whether a new survey of globins in bacteria was necessary? We believe that it was, given that the earlier, limited survey [22,23], missed the presence of sensor single domain globins, the SSDgbs. At that time, their absence represented a substantial obstacle to the formulation of a model of globin evolution based on endosymbiotic events, particularly in view of the discovery of bacterial-like SSDgbs in fungi [41]. Thus, we feel justified in having carried out the new survey, to ensure that we would obtain the most complete coverage of bacterial globins possible. What about the bacterial species coverage? An evaluation of the approximately 1000 bacterial genomes available in 2007 [42] carried out by Wu et al. in 2009, concluded that the data suffered from a pronounced sampling bias, and suggested ways of correcting it [43]. Since then, more than 2000 completed genomes, including the Human Microbiome Project genomes [40] and those discussed in ref [43] have become available. Thus, although bias towards major bacterial groups such as Actinobacteria,

Table 1
Summary and structural properties of the three bacterial globin families and eight subfamilies.

Fold	3/3 Mb-fold		3/3 Mb-fold			2/2 Mb-fold		
Family name	Myoglobin-like		Globin-coupled sensor			Truncated		
Subfamily name	FHb (chimeric)	SDgb (single domain)	GCS (chimeric)	Pgb (single domain)	SSDgb	TrHb1 N	TrHb2 O	TrHb3 P
Estimated number of sequences	533	221	420	40	35	235	622	334