Origin and Evolution of the Unique Tetra-Domain Hemoglobin from the Hydrothermal Vent Scale Worm Branchipolynoe

J. Projecto-Garcia, N. Zorn, Didier Jollivet, S. W. Schaeffer, François Lallier, Stépahne Hourdez

Origin and evolution of the unique tetra-domain hemoglobin from the hydrothermal vent scale-worm *Branchipolynoe*

SUBMISSION AS A RESEARCH ARTICLE

Projecto-Garcia[1,2], J., N. Zorn[3], D. Jollivet[1,2], S. W. Schaeffer[4], F. H. Lallier[2,5] and S. Hourdez[1,2]

[1] Équipe GAME, CNRS UMR 7144, Station Biologique de Roscoff, BP 74, Places Georges Teissier, 29682 Roscoff cedex, France

[2] Université Pierre et Marie Curie, Laboratoire Adaptation et Diversité en Milieu Marin, Place Georges Teissier, 29682 Roscoff cedex, France

[3]LDSM2 Institut de Chimie de Strasbourg, UMR 7177 CNRS-UDS, 1 rue Blaise Pascal - BP 296 R8, 67008 Strasbourg Cedex, France

[4] Department of Biology and Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pennsylvania 16802, USA

[5] Équipe EIMME, CNRS UMR 7144, Station Biologique de Roscoff, BP 74, Places Georges Teissier 29682 Roscoff cedex, France

Corresponding author:

Stéphane Hourdez (hourdez@sb-roscoff.fr)

Équipe GAME, CNRS UMR 7144, Station Biologique de Roscoff, BP 74, Places Georges Teissier, 29682 Roscoff cedex, France

Tel: +33 (0)2 98 29 23 40

Fax: +33 (0)2 98 29 23 24

Keywords: extracellular globins, tetra-domain, tandem duplication, evolution, annelids, *Branchipolynoe*

Running head: Scale-worm tetra-domain globin evolution

Abstract

Hemoglobin is the most common respiratory pigment in annelids. They can be intra- or extra-cellular, and this latter type can form large multimeric complexes. The hydrothermal vent scale-worms *Branchipolynoe symmytilida* and *B. seepensis* express an extracellular tetra-domain hemoglobin that is unique in annelids. We sequenced the gene for the single-domain and tetra-domain globins in these two species. The single-domain gene codes for a mature protein of 137 amino acids, and the tetra-domain gene codes for a mature protein of 552 amino acids. The single-domain gene has a typical 3 exons/2 introns structure, with introns located at their typical positions (B12.2 and G7.0). This structure is repeated four times in the tetra-domain gene, with no bridge introns or linker sequences between domains. The phylogenetic position of *Branchipolynoe* globins among known annelid globins revealed that, although extracellular, they cluster within the annelid intracellular globins clade, suggesting that the extracellular state of these hemoglobins is the result of convergent evolution. The tetra-domain structure likely resulted from two tandem duplications, domain 1 giving rise to domain 2 and after this the two-domain gene duplicated to produce domains 3 and 4. The high O_2 affinity of *Branchipolynoe* extracellular globins may be explained by the two key residues (B10Y and E7Q) in the heme pocket in each of the domains of the single and tetra-domain globins, which have been shown to be essential in the oxygen-avid hemoglobin from the nematode *Ascaris suum*. This peculiar globin evolutionary path seems to be very different from other annelid extracellular globins and is most likely the product of evolutionary tinkering associated with the strong selective pressure to adapt to chronic hypoxia that characterizes hydrothermal vents.

Introduction

Globins produced by invertebrates exhibit considerable heterogeneity in protein sequence and in quaternary structure when compared to vertebrates. Despite this heterogeneity, these proteins fold into the highly conserved globin fold that allows heme binding. The quaternary structure diversity encompasses simple monomers, polymeric subunits made of single-domains, and polymeric multi-domain subunits (see Weber and Vinogradov 2001 for a review). Multi-domain globins can be intracellular, such as in the bivalve *Barbatia* (Grinich and Terwilliger 1980, Suzuki and Arita 1995), or extracellular such as in crustaceans, nematodes, and molluscs (see Weber and Vinogradov 2001 for a review). This type of structure was not known in annelids until its discovery in two closely related hydrothermal vent species of annelids, *Branchipolynoe symmytilida* and *B. seepensis*, which possess tetra-domain globins (Hourdez et al. 1999a).

The presence of hemoglobin (Hb) in *Branchipolynoe* spp. may have an adaptive value in their chronically hypoxic environment (Hourdez and Lallier 2007). *Branchipolynoe* belongs to the Polynoidae, a family of scale-worms that is widely distributed in marine ecosystems from the littoral to the deep-sea. They are very diverse at hydrothermal vents (Tunnicliffe 1991) where they occupy all the microhabitats available to metazoa, from the coldest (and most oxygenated), to the warmest (and usually the most hypoxic) waters. Scale-worms in general (comprising several families, and including the Polynoidae) were known for having only tissue globins (neuroglobin and myoglobin) (Weber 1978; Dewilde et al. 1996). In contrast to their littoral relatives, most hydrothermal vent polynoid species have extracellular Hbs giving them a red-pigmented coelomic fluid (Hourdez et al. 1999a; S.H. pers. obs.). These Hbs can facilitate the diffusion of oxygen from the hypoxic environment, especially given their high affinity for oxygen (Hourdez et al. 1999b), and represent a significant oxygen storage for periods of complete anoxia (Hourdez and Weber 2005).

A phylogeny of the annelid globin sequences showed that the extracellular globins all evolved from a duplicated intracellular ancestral globin, and several duplication events followed to produce the present-day globin diversity found in the complex extracellular hexagonal bilayer Hb (HBL-Hb; Bailly et al. 2007). The original split between the intracellular and the extracellular globins occurred about 570 millions years ago in annelids (Goodman et al. 1988).

The hydrothermal vent environment was colonized by scale-worms about 65 million years ago (Hourdez et al. in prep.). All the vent-endemic subfamilies form a monophyletic group that rapidly radiated after this initial colonization event. All of these species express Hbs (single- or tetra-domain), suggesting that this trait was already present in the common vent ancestor (Hourdez et al. in prep.). We sequenced both the single- and tetra-domain (hereafter referred to SD, and TD, respectively) globin genes in *B. symmytilida* and *B. seepensis* to examine their exon/intron structure and to understand the origin and evolution of the tetra-domain globin.

Material and methods

Animal collection

Branchipolynoe symmytilida specimens were collected from the 9˚50' N locality on the East Pacific Rise in 2001 (9º46' N, 104˚21' W, 2515 m depth), and *B. seepensis* were collected from the Lucky Strike site in 2001 (Mid-Atlantic Ridge, 37º18' N, 32º16' W, 1700 m depth). The worms were removed from the pallial cavity of their host mussels. The specimens were identified and immediately frozen in liquid nitrogen on board, transported back to the laboratory and stored at -80ºC until analyzed.

Nucleic acids extraction, and cDNA synthesis

DNA, RNA extractions, and cDNA synthesis were performed with standard procedures that are described in Supplementary materials.

Globin cDNA and gene sequencing for *Branchipolynoe symmytilida*

- Protein microsequencing and PCR primer design

The two coelomic hemoglobins (HbC1 and HbC2) from *Branchipolynoe symmytilida* were purified as described earlier (Hourdez et al. 1999a). We used Edman degradation for microsequencing and the released amino acids were identified by HPLC (High Performance Liquid Chromatography). The N-terminus in HbC1 was blocked but HbC2 yielded the microsequence N-terminus VSAAQKAAIK. Based on this microsequence, degenerate primers were designed to amplify HbC2 by PCR (Table S1, Supplementary material).

- Initial globin amplification and sequencing

Primers BSY_E1D1_C2F and an anchored oligo(dT) (Table S1) were used on the synthesized cDNA, to amplify part of the tetra-domain globin gene. Detailed PCR conditions and cloning procedures are given in Supplementary material. This initial PCR reaction

produced two insert size classes: 460-600 bp inserts that include one domain (400 bp) with different sized 3'UTRs, and 900 bp inserts corresponding to two domains (800 bp) with different sized 3'UTRs (See Results, Fig. S1). Two sequences representing the two size classes of inserts were chosen because they were the most abundant among the clones, designated BSY3 and BSY7. These sequenced clones were used to design specific primers to amplify the coding sequence and introns of the corresponding genes (single-domain and tetra-domain) (Table S1).

- Bridging oligonucleotide rapid amplification of cDNA ends (BO-RACEing) for the single-domain globin

The 5'UTR for sequence BSY3 was amplified by BO-RACEing (Shi et al. 2002). Specific primers and conditions are given in Supplementary material. This approach did not yield any amplification for sequence BSY7 and chromosome walking was used instead to obtain the missing portion of the sequence.

- Chromosome walking on genomic DNA

Sequencing of globin BSY7 from *B symmytilida* was completed by directional genome walking using PCR (Mishra et al., 2002). The detailed procedure is given in the Supplementary material. Directional genome walking was also used to obtain the promoter regions of both the TD and SD globins, in *B. symmytilida* and in *B. seepensis*.

- Northern blot

A standard northern blot protocol (Sambrook et al. 1989) was used for *Branchipolynoe symmytilida*. The detailed procedure is given in the Supplementary material.

Globin amplification and sequencing in *Branchipolynoe seepensis*

The amplification of the SD and TD globin genes (cDNA and gDNA) from *B. seepensis* took advantage of the primers designed for the same genes in *B. symmytilida* and of an EST

library from which 2500 clones were sequenced (Hourdez and Tanguy, unpub. data). Amplification and sequencing followed a standard procedure detailed in Supplementary materials. Chromosome directional walking (see above) was used to sequence the 5'UTR and the promoter region of the globin genes.

Phylogenetic Analyses

The sequences were assembled and nucleotide positions with conflicts were resolved based on their chromatograms with CodonCode Aligner© 2.0.6 (http://www.codoncode.com/aligner/index.htm). Multiple nucleotide and amino acid sequence alignments were performed by using ClustalX 2.0.10 (Larkin et al. 2007) and, when necessary, manually optimized by using the sequence aligner editor Se-Al 2.0a11 Carbon (http://tree.bio.ed.ac.uk/software/seal/). During optimization, we minimized the number of indels and nucleotide alignment was constrained by the amino acid sequences alignment.

Before all analyses, we verified that there was no saturation of the signal by plotting K2P pairwise differences using all nucleotide sites against K2P pairwise differences estimated only on the first two codon positions (data not shown).

jModelTest was used to choose a model of nucleotide substitution for use in phylogenetic analysis that would best fit our data (Guindon and Gascuel 2003, Posada 2008). Neighbor-joining (Saitou and Nei 1987) trees were constructed using MEGA4 (Tamura et al. 2007), maximum-likelihood trees were computed using Phylip (http://evolution.genetics.washington.edu/phylip.html) and Phyml Online (http://atgc.lirmm.fr/phyml/, Guindon and Gascuel 2003, Guindon et al. 2005). Finally, a Bayesian analysis was performed with Mr. Bayes (Ronquist and Huelsenbeck 2003, Huelsenbeck and Ronquist 2001). PAML 3.14 (http://abacus.gene.ucl.ac.uk/software/paml.html, Yang 1997) was later used to test different

tree topologies issued from the different phylogenetic analyses, by maximum likelihood relative ratio tests, and the codon substitution model derived from Goldman & Yang (1994). The tree topologies were rooted only when a molecular clock hypothesis was being tested, and under every test condition the single-domains were forced to be monophyletic. To test for molecular clock and possible differences in selective pressures between paralogous domains of the tetra-domain globin, we calculated pairwise distances between species for each domain, using the K2P model for nucleotides and the Dayhoff matrix for amino acids (MEGA4).

Results

cDNA cloning and sequencing

Amplification of *B. symmytilida* cDNA with degenerate primers and oligo(dT) yielded six distinct sequences that can be separated into two types (Fig. S1). The first type has a 400 bp coding region, corresponding to a globin domain, followed by a 3'UTR sequence that differed in length (sequences BSY1 and BSY3). The second type has a 800 bp coding region, corresponding to two globin domains (sequences BSY2, BSY4, BSY5, and BSY7), followed by a 3'UTR sequence also presenting differences in length. These two globin domains will be hereafter referred to as D3 and D4 (from 5' to 3'), in reference to the *Branchipolynoe* hemoglobin model composed of 4 domains (Hourdez et al. 1999a) (the same nomenclature will be used for the upstream domains: D1 and D2). The 3'UTR for sequences BSY2, BSY4, and BSY7 are nearly identical and a few differences appeared in their coding regions. The three remaining 3'UTR sequences (corresponding to sequences BSY1, BSY3, and BSY5) are clearly different. A Northern blot on *B. symmytilida* cDNA samples probed with a portion of sequence BSY3 and BSY7 (the most common sequences among the clones) revealed that these correspond to different transcript sizes. The BSY3 probe revealed a single band of ~610 bp in length, the size expected for a full-length mRNA coding for a single-domain globin (Fig. S2A). The BSY7 probe revealed a single band of ~1980 bp in length, consistent with the expected size for full-length mRNA coding for a tetra-domain globin (Fig. S2B).

The complete BSY3 cDNA sequence encodes a SD globin (coding sequence of 417 bp, including the stop codon), with a 67-base 5'UTR (Fig. S3), and a 58-base 3'UTR. After removal of the initial methionine, the deduced amino acid sequence (137 codons) would produce a protein with a molecular weight (MW) of 13826.8 Da.

The complete BSY7 cDNA sequence has a 79-base 5'UTR, a 1674 bp coding sequence (including the stop codon), and a 102-base 3'UTR (polyA tail not included). The mature

protein encoded corresponds to a 4-domain globin (552 codons) (Fig. S4), with a MW of 57691.9 Da, a mass that closely matches the value found for the native HbC2 subunit (Hourdez et al, 1999a).

The cDNA of *B. seepensis* for both the SD and TD globins have the same length and number of codons as the homologous sequences from *B. symmytilida* (Figs. S3 and S4). The corresponding proteins have a MW of 13798.7 and 57742.9 Da for the SD and TD globins, respectively.

Protein primary structure

The protein sequences for both *Branchipolynoe* species were aligned with other intracellular and extracellular annelid globins, two nematode (*Ascaris suum*) globin sequences, and the myoglobin from the sperm-whale *Physeter catodon* (Fig. 1). Only two positions were invariant: the proximal histidine (F8H, i.e. helix F, position 8 with the *Physeter* myoglobin sequence as a reference), and phenylalanine on the corner between helices C and D (CD1F). A tryptophan in helix A (A12W) was conserved in all but the nematode sequences. For all *B. symmytilida* and *B. seepensis* sequences, the distal histidine is replaced by a glutamine (E7Q), a feature also observed in *Ascaris*. Similarly, the B-helix tyrosine (B10Y), essential in modulating the oxygen affinity in *Ascaris* (De Baere et al. 1994), is also found in all *Branchipolynoe* spp. sequences. A tryptophan (H8W) is present in all sequences except in *P. catodon*. The two conserved cysteines in extracellular globins that form the characteristic intrachain disulfide bridge (A2C and H11C) are not found in any of the *Branchipolynoe* sequences.

The TD sequences also have a cysteine (E18C) 11 residues after E7Q, a feature that is found in some globins of the tubeworms *Riftia pachyptila* and *Lamellibrachia* sp. (from hydrothermal vents and cold seeps, respectively). For all *Branchipolynoe* sequences, there is no pre A helix, signal peptide, or linker sequences.

Origin of the globin gene and relationships with other annelid globins

For both *Branchipolynoe* species, polymorphism was present in the different amplified sequences, therefore, a consensus nucleotide sequence was produced based on the majority of clones obtained for each species. These consensus sequences were translated and used in all phylogenetic analyses.

The unrooted phylogenetic tree clearly separates the globins that are typical extracellular HBL-Hbs from all intracellular globins (Fig. 2). The SD globin and the four domains of the TD globin from both *Branchipolynoe* species clearly group with the intracellular annelid globins, indicating that these extracellular globins have a distinct origin from all other annelid extracellular globins.

Evolution of the tetra-domain structure

A phylogeny of the various *Branchipolynoe* globin sequences was determined using the SD as an outgroup because of its more ancestral status compared to the tetra-domain sequences (Fig. 2). Several nucleotide substitution models were used, and when using Mr. Bayes the codon model was also chosen. These analyses yielded different tree topologies depending on the phylogenetic method used and the selected nucleotide substitution model (Fig. S5).

The phylogenetic tree based on amino-acid sequences suggests that D4 is the most ancestral of the domains followed by D3, D2, and D1 in the TD gene. Although this phylogeny is well supported, it may not reflect the history of domain duplications as these can be going through periods of relaxed selective pressure and/or positive selection. Analysis based on nucleotides provided a different result, suggesting that D1 (or D3) is the most ancestral domain. The best resolved trees correspond to 4 equiprobable topologies (Fig. S5).

In each tree topology, the globin domains of both species form well-supported pairs, indicating that duplications giving birth to the TD occurred prior to the radiation between the two *Branchipolynoe* species. The deeper branches of the nucleotide-based topologies, however, were not well supported (bootstrap values usually no higher than 65%, and at best 73%), suggesting that duplications occurred very close to each other over time, or that variable selective pressure over the various branches reduced the phylogenetic signal.

The different topologies were tested against each other using maximum likelihood relative ratio tests in CodeML (Yang 1997). Pairwise comparison of these 4 main topologies revealed topology (c) was the most probable, regardless of the d_N/d_S ratio model selected. The pairwise differences between the different TD domains and the SD indicated that D1 possesses the fewest differences with SD, followed by D3 (Table 1). In addition, D1 also displayed several motifs of adjacent amino acids that were nearly identical with the SD globin, reinforcing the hypothesis that D1 is more closely related to the ancestral state. D1 and D3 were also the domains with the smallest divergence between species for either nucleotide or amino acid sequences, suggesting that these domains are more constrained by purifying selection. To verify this hypothesis, we used BaseML (PAML) to test for a molecular clock using the theoretical topology D1D3 vs D2D4. The results indicate that the molecular clock is rejected in the evolution of the TD ($L_{clock=0}$ = -2171.78932, $L_{clock=1}$ = -2190.984727; RRT = 2*($L_{clock=0}$ - $L_{clock=1}$) = 38.390814 >> 10.83 (p = 0.001)). This supports the idea that D1 and D3 may be under stronger selective constraints and evolving slower than the other domains.

Promoters and 5'UTRs

A portion of the promoter region was sequenced for the SD globin from *B. symmytilida*, and for the TD globin for both species. For both genes, the TATA box is located 32 bp upstream of the transcription start in *B. symmytilida*. For the SD sequence, only one base was

different for the portion of 5'UTR for which there are data for both species (48 nucleotides, i.e. 97.9% identity) (Fig. S3). The promoter for the TD globin, as well as the 5'UTR, from both species, also exhibit a high identity level (97.5% for the 5'UTR) (Fig. S4). A search for transcription regulatory signal binding sites in *B. symmytilida* revealed a site for CP2, a factor that is known to enhance the transcription of globin genes in erythroid cells (Chae and Kim 2003). This element was however not found in *B. seepensis*. Other universal transcription factors, such as Oct-1, were identified in the promoter of both species (data not shown). Although these are extracellular Hbs no signal sequence for secretion was found. In *B. seepensis*, we sequenced a 700 bp fragment of the promoter. This region contains motifs for the binding of two transcription factors relevant for this gene: one site for hypoxia inducible factor 1 (HIF-1); and two sites for GATA-1, which plays an important role in erythroid development (De Maria et al. 1999) and has been reported to enhance erythropoeisis in response to tissue hypoxia (Krantz 1991, Zon et al. 1991). The sequenced promoter region for *B. symmytilida* does not cover the region where these transcription factors' binding sites are located in *B. seepensis*.

Introns and gene structure

In both the SD and TD globin genomic sequences from both species, the typical 3 exons/2 introns globin gene structure was present, and this motif is repeated four times in the TD gene (Fig. 3). For all the genes, the introns are located at positions B12.2 and G7.0, corresponding to the typical globin introns positions. In the SD globin gene, the second intron exhibits a very high level of identity between the two species (94.4%). Although this is also true for the first 160 bp in intron 1, the identity then drops sharply because *B. seepensis* possesses a large (440 bp) insertion (Table S2). In the TD globin genomic sequence, each intron was usually 300 bp long, with one exception at 700 bp again for the first intron of D1 (Table S2). Interestingly,

neither introns nor linker segments separate the four globin domains (Fig. S4). Pairwise intron sequence comparisons between the two species revealed a very high level of conservation (on average 90% of identity) of all but the second intron of D4. The major differences in the orthologous introns correspond to indels of varying size, from 3 to 41 bp. Comparisons of intron sequences among domains failed to reveal significant sequence similarity. In intron 2 from D1 and D2 from *B. seepensis*, there is a regulatory signal for GATA-1 and HIF-1, respectively.

Discussion

The coding sequence for the unique TD and SD globins genes, as well as the position and size of the introns have been determined to understand the origin and evolution of these respiratory pigments. The TD globin with a multi-domain gene structure represents the emergence of a new type of hemoglobin in annelids (Terwilliger 1992).

Clarification of the protein structure

Hourdez et al. (1999a) showed that the *Branchipolynoe* extracellular TD globins form dimers (HbC2) or trimers (HbC1) and our results suggest that we most likely amplified the former because the estimated MW most closely matches the one for HbC2 subunits. Large polypeptidic assemblages are not unusual in invertebrate respiratory pigments (Terwilliger, 1998). In all cases, the formation of large polypeptide complexes has been attributed to the necessity to avoid their excretion (Weber and Vinogradov, 2001). In *Daphnia* and *Artemia*, the didomain and nine-domain Hbs, respectively, can form homo- or hetero-dimers (Dewilde et al. 1999, Mansfield et al. 1980). In the nematode *A. suum* the didomain Hbs assemble into octamers of 328 kDa, through a C-terminus extension that forms a β-strand and each strand binds to each other with H-bonds in a zipper-like structure (Darawshe et al. 1987, De Baere et al. 1992). In *Branchipolynoe*, there is no such extension and there are no interchain disulfide bridges to assemble into trimers or dimers, and only inter-subunit non-covalent bonds are involved.

Based on subtilisin partial digestion patterns and heme content determination, Hourdez et al. (1999a) suggested that each Hb subunit was composed of 4 globin domains (with one likely truncated) each with one heme group. The sequences we obtained clearly show that there are indeed 4 domains but each of these domains is full-length. This may indicate that the

interdomain regions are not well exposed in the native Hb, and susceptible to peptidic cleavage by a wide-spectrum peptidase.

Interestingly, the globins from *Branchipolynoe* do not possess the two cysteine residues necessary to form the conserved disulfide bridge conserved in all other extracellular annelid globins (Bailly et al., 2007). This indicates that this character is not necessary for the globins to function outside of cells, although its conservation in the lineage of the HBL-Hb globins suggests that purifying selection is constraining these amino acid positions.

Structure/function relationship

The amino acid alignment revealed the presence of a glutamine instead of a distal histidine (E7H). Although this amino acid is usually conserved in vertebrates, some invertebrate species use other amino acids. In the nematode *A. suum* Hb, known to have one of the highest oxygen affinities among invertebrates (see Weber and Vinogradov 2001 for a review), the position is also occupied by a glutamine (E7Q). This high affinity can be explained in part by the interaction of the E7Q with a tyrosine in position B10Y (De Baere et al. 1994, Yang et al. 1995). Interestingly, *Branchipolynoe* also has a tyrosine in position B10. This could explain the high affinity for oxygen that was found in *B. symmytilida* Hbs (Hourdez et al. 1999b).

In the TD globins, but not in the SD ones, there is a cysteine located 11 residues after the E7Q (E18C). This cysteine is known to be under strong selection in some annelids living in reduced habitats (Bailly et al. 2003), suggesting a key function for E18C in these habitats. It was hypothesized to be a key residue for reversible sulfide binding, a characteristic that is essential for vestimentiferan tubeworms that live symbiotically with sulfide-oxidizing bacteria (Zal et al. 1998). It was later shown that the E18 cysteine was not likely the binding site for transporting sulfide in *Riftia pachyptila* and that, at least for the 400-kDa Hb, zinc atoms were involved (Flores et al. 2005, Flores and Hourdez 2006). The SD globin from *Branchipolynoe*

is devoid of this specific residue, indicating that it most likely represents an adaptive convergence to life in a sulfide-rich habitat in the TDs. Its presence in the heme pocket may have a protective role for the heme group that usually reacts with sulfide to irreversibly form sulfhemoglobin, an altered molecule incapable of binding oxygen.

Secretion into the coelomic fluid

Although *Branchipolynoe* SD and TD Hbs are extracellular (secreted in the coelomic cavity), their genes do not possess a signal peptide for secretion whereas all other annelid extracellular globins have such a signal (Riggs 1991; Bailly et al. 2007). The absence of a signal peptide in all *Branchipolynoe* globins could be due to the fact that we did not obtain the whole sequence and that the upstream sequence (promoter) actually corresponds to an intron. This however is unlikely as we did locate the TATA box and the sequence corresponding to the 5'UTR, and a 700 bp stretch upstream did not reveal any other open reading frame in *B. seepensis* TD globin gene. This may indicate that the secretion of *Branchipolynoe* Hbs is holocrine (by rupture of the cell membrane, and the release of all the components of the cytoplasm). This is supported by the fact that, in at least one other vent species (*Lepidonotopodium piscesae*), the SD globin expressed in the muscles (myoglobin) has exactly the same molecular weight as the one found in the coelomic cavity (Hourdez, unpub. data).

Origin and evolution of the tetra-domain globin

Phylogenetic relationships among annelid globins indicate that extracellular globins have evolved independently from the intracellular (circulating or non-circulating) ones, and diverged about 570 millions years ago (Goodman et al. 1988), an origin distinct from extracellular globins found in other phyla (Bailly et al. 2007). *Branchipolynoe* extracellular

globin sequences have a distinct history, not only from other phyla's extracellular globins (including multidomain globins), but also from the typical extracellular annelid globins. It appears that they arose from an intracellular annelid globin more recently than the original split between intra and extracellular globins in annelids. Despite this different evolutionary history, the SD and TD globins conserved the basic globin gene structure of 3 exons and 2 introns, including intron positions. This arrangement differs from that of *Aphrodite aculeata* myoglobin (Dewilde et al. 1996) in which the first intron is missing. This species belongs to the scale-worm family Aphroditidae, a close related family to the Polynoidae. This indicates that the ancestor to all scale-worms most likely did possess this first (B12.2) intron, and that it was later lost in the lineage leading to *Aphrodite aculeata*. Some of the introns showed a high level of conservation between the two species, but in both the SD and the TD globin genes there is a strongly divergent intron (intron 1 in the SD, intron 2 in D4 for the TD). This may indicate strong selective constraint on the intron sequence for a possible role related to a regulatory function.

The phylogenetic analyses yielded trees with only limited support for the internal branches. This could be indicative of a rapid succession of the duplication events to lead to this structure, and/or to different selective pressure affecting the various domains. We found that D1 and D3 are probably under stronger purifying selection than D2 and D4. This would then yield to different branch lengths, and the absence of a molecular clock. Although the phylogenetic trees do not allow us to clearly decide on a likely duplication scenario, other observations (pairwise distances, conservation of some amino acid motifs), suggest that the TD structure is likely the result of two tandem duplications, the first one giving rise to D1D2, and the second one affecting these two domains together to generate the whole TD. Domain D1 is likely the most ancestral domain, as (1) in both species, intron 1 from D1 has the same size as in the SD globin, (2) pairwise differences between the SD and D1 are smaller than all

the other domains, and (3) there are some conserved amino acids motifs. This duplication scenario shows that duplication moved in a 5' to 3' direction where downstream duplicates were likely joined by intron losses. The alternative scenario where domain 4 is ancestral, and duplications proceeded in a 3' to 5' direction seems less likely because with each subsequent domain duplication the 5' regulatory region may have had to be reconstructed.

Duplication events

The mechanism responsible for these duplication events remains unclear for the moment. Interestingly, the TD globin gene from these scale-worms does not have any bridge introns. This contrasts with the didomain globins from the bivalve *Barbatia reeveana* (Naito et al. 1991), the water-flea *Daphnia pulex* (Dewilde et al. 1999), and some nematodes (*Ascaris suum* and *Pseudoterranova decipiens*; Dixon et al. 1992, Sherman et al. 1992). Similarly, in mollusk hemocyanins, the 7 or 8 domains are separated by bridge introns that are thought to be the remnants of the duplication of the gene (Lieb and Todt 2008). Amino acid linker sequences are also missing between the domains of the TD globins from *Branchipolynoe*. This kind of structure is known in other invertebrate multidomain proteins such as the nine-domain Hb from *Artemia* (Maning et al. 1990, Trotman et al. 1994), and the didomain Hb of the bivalve *B. reeveana* (Naito et al. 1991). It is however not found in the nematodes two-domain Hb (De Baere et al. 1992, Sherman et al. 1992). These linker sequences are thought to be the remnants of an inter-domain intron that lost its splicing signals and was eventually integrated into the coding sequence. All *Branchipolynoe* TD domains are full-length, indicating that there likely was inter-domain introns that were secondarily integrally lost, as suggested by Naito et al. (1991) for *B. reeveana*, Dewilde et al. (1999) for *D. pulex* Hb, and Dixon et al. (1992) for *P. decepiens*. The alternative, less likely, hypothesis would require two end-to-end tandem duplications that preserved (1) the whole length of the domains, and (2)

the reading frame. In another water-flea, *Moina macrocopa*, the two-domain globin also lacks the bridge intron (Kato et al. 2001), and the authors suggest unequal crossing-over as the mechanism for the duplication of the gene, with a subsequent loss of the bridge intron when the Cladocera (water-fleas) families diverged. Finally, the TD structure could also correspond to the result of unequal crossing-overs between two copies of an initial didomain structure but this would yield chimeric domains and our search for such possible chimeras proved unsuccessful (data not shown).

Concluding remarks

Our work shows that *Branchipolynoe* extracellular Hbs have a different origin and evolutionary history from other annelid extracellular Hbs. It appears that an ancestral intracellular myoglobin was duplicated and evolved to form an extracellular globin by evolutive tinkering. The evolution of these Hbs, constitutively expressed at high levels, is most likely the result of strong selective pressure due to the chronic hypoxia that characterizes hydrothermal vents (Hourdez and Lallier, 2007). The multidomain structure allows higher concentration of oxygen binding sites (and thereby transport/storage capacity), without increasing the colloid osmotic pressure of the coelomic fluid.

Acknowledgments

The authors would like to thank the crews of the ships and submersibles, as well as the chief scientists, of the cruises ATOS 2001 (project funded by Ifremer and INSU), and EPR 2001 (project funded by a NSF grant to C.R. Fisher (NSF OCE-0002729)). We would also like to thank Isabelle Boutet-Tanguy and Arnaud Tanguy for technical advice in lab. This work is part of the project HYPOXEVO (Région Bretagne), and was supported by the ESTeam research Marie Curie grant under the 6th framework program from the European Commission.

Supplementary material

Materials and methods

Nucleic acids extraction, and cDNA synthesis

Initial *B. symmytilida* globin amplification and sequencing

Bridging oligonucleotide rapid amplification of cDNA ends (BO-RACEing) for the single-domain globin

Northern-blot on Branchipolynoe symmytilida RNA

Directional chromosome walking

B. seepensis globin sequencing

Table S1, Table S2

Figures S1, S2, S3, S4, and S5

Literature cited

Bailly X, Leroy R, Carney S, Collin O, Zal F, Toulmond A, Jollivet D. 2003. The loss of the hemoglobin H2S-binding function in annelids from sulfide-free habitats reveals molecular adaptation driven by Darwinian positive selection. P Natl Acad Sci USA. 100:5885-5890.

Bailly X, Chabasse C, Hourdez S, Dewilde S, Moens L, and Zal F. 2007. Globin gene family evolution and functional diversification in annelids. FEBS J. 274: 2641-2652.

Chae JH, Kim CG. 2003. CP2 Binding to the Promoter is Essential for the Enhanced Transcription of Globin Genes in Erythroid Cells. Mol Cells. 15: 40-47.

Church GM, Gilbert W. 1984. Genomic sequencing. P Natl Acad Sci USA 81:1991-1995.

Darawshe S, Tsafadyah Y, Daniel E. 1987. Quaternary structure of erythrocruorin from the nematode *Ascaris suum*. Biochem J. 242: 689-694.

De Baere I, Liu L, Moens L, Van Beeumen J, Gielens C, Richelle J, Trotman C, Finch J, Gerstein M, Perutz M. 1992. Polar zipper sequence in the high-affinity hemoglobin of *Ascaris suum*: Amino acid sequence and structural interpretation. P Natl Acad Sci USA. 89:4638-4642.

De Baere I, Perutz MF, Kiger L, Marden MC, Poyart C. 1994. Formation of two hydrogen bonds from the globin to the heme-linked oxygen molecule in *Ascaris* hemoglobin. P Natl Acad Sci USA. 91:1594-1597.

De Maria R, Zeuner A, Eramo A, Domenichelli C, Bonci D, Grignani F, Srinivasula SM, Alnemri

ES, Testa U, Peschle C. 1999. Negative regulation of erythropoiesis by caspase-mediated cleavage of GATA-1. Nature. 401:489-493.

Dewilde S, Blaxter M, Van Hauwaert M-L, Vanfleteren J, Esmans EL, Marden M, Griffon N, Moens L. 1996. Globin and Globin Structure of the Nerve Myoglobin of *Aphrodite aculeata*. J Biol Chem. 271:19865-19870.

Dewilde S, Van Hauwaert M-L, Peeters K, Vanfleteren J, Moens L. 1999. *Daphnia pulex* didomain hemoglobin: structure and evolution of polymeric hemoglobins and their coding genes. Mol Biol Evol. 16:1208-1218.

Dixon B, Walker B, Kimmins W, Pohajdak B. 1992. A nematode gene contains an intron previously thought to be unique to plants. J Mol Evol. 35:131-136.

Flores JF, Fisher CR, Carney SL, Green BN, Freytag JK, Schaeffer SW, Royer Jr WE. 2005. Sulfide binding is mediated by zinc ions discovered in the crystal structure of a hydrothermal vent tubeworm hemoglobin. P Natl Acad Sci USA. 102:2713-2718.

Flores JF, Hourdez S. 2006. The zinc-mediated sulfide-binding mechanism of hydrothermal vent tubeworm 400-kDa hemoglobin. Cah Biol Mar. 47: 371-377.

Goldman N, Yang Z.1994 A codon based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736

Goodman M, Pedwaydon J, Czelusniak J, Suzuki T, Gotoh T, Moens L, Shishikura F, Walz D,

Vinogradov SN. 1988. An evolutionary tree for invertebrate globin sequences. J Mol Evol. 27:236-249.

Grinich N P, Terwilliger RC. 1980. The quaternary structure of an unusual high-molecular-weight intracellular hemoglobin from the bivalve mollusc Barbatia reeveana. Biochem J. 189: 1-8.

Guindon S, Gascuel O. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst Biol 52: 696-704.

Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online-a web server for fast maximum likelihood-based phylogenetic inference. Nucl. Acids Res. 33:W557-W559.

Hourdez S, Lallier FH, Green BN, Toulmond A. 1999a. Hemoglobins from deep-sea hydrothermal vent scale-worms of the genus *Branchipolynoe*: A new type of quaternary structure. Proteins. 34:427-434.

Hourdez S, Lallier FH, Martin-Jézéquel V, Weber RE, Toulmond A. 1999b. Characterization and functional properties of the extracellular coelomic hemoglobins from the deep-sea, hydrothermal vent scale-worm *Branchipolynoe symmytilida*. Proteins. 34:435-442.

Hourdez S, Lallier FH. 2007. Adaptations to hypoxia in hydrothermal vent and cold-seep invertebrates. Rev Env Sci Biotech. 6: 143-159.

Hourdez S, Weber RE. 2005. Molecular and functional adaptations in deep-sea hemoglobins. J Inorg Biochem. 99(1): 130-141

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17: 754-755.

Kato K, Tokishita S-i, Mandokoro Y, Kimura S, Ohta T, Kobayashi M, Yamagata H. 2001. Two-domain hemoglobin gene of the water flea *Moina macrocopa*: duplication in the ancestral Cladocera, diversification, and loss of a bridge intron. Gene. 273:41-50.

Krantz SB. 1991. Erythropoietin. Blood. 77(3): 419-434.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. (3 co-authors). 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948.

Lieb B, Todt C. 2008. Hemocyanin in mollusks - A molecular survey and new data on hemocyanin genes in *Solenogastres* and *Caudofoveata*. Mol Phylogenet Evol. 49:382-385.

Manning A, Trotman C, Tate W. 1990. Evolution of a polymeric globin in the brine shrimp Artemia. Nature. 348: 653-656.

Mansfield BC, Krissansen GW, Smith MG, Tate WP. 1980. Isolation of intact larval haemoglobin from the brine shrimp *Artemia salina*. Prevention of degradation in vitro by proteases induced during larval development. Biochem Biophys Acta. 623:163-170.

Mishra RN, Singla-Pareek SL, Nair S, Sopory SK, Reddy MK. 2002. Directional Genome Walking

Using PCR. BioTechniques. 33:830- 834.

Naito Y, Riggs C, Vandergon T, Riggs A. 1991. Origin of a "bridge" intron in the gene for two domain globin. P Natl Acad Sci USA. 88: 6672-6676.

Posada D. 2008. jModelTest: Phylogenetic Averaging. Mol Biol Evol. 25:1253-1256.

Riggs AF. 1991. Aspects of the Origin ad Evolution of Non-Vertebrate Hemoglobins. Am Zool. 31: 535-545.

Rodionov MA, Johnson MS. 1994. Residue-residue contact substitution probabilites derived from aligned three-dimensional structures and the identification of common folds. Protein Sci. 3: 2366-2377.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: 1572-1574.

Saitou N, Nei M. 1987. The neighbor-joining method: A nem method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406-425.

Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular cloning: a laboratory manual. Nolan C. ed. Cold Spring Harbour :Cold Springs Harbor Press.

Sherman DR, Kloek AP, Krishnan BR, Guinn B. 1992. *Ascaris* hemoglobin gene: Plant-like structure reflects the ancestral globin gene. P Natl Acad Sci USA. 89:11696-11700.

Shi X, Karkut T, Chahmanhkah M, Alting-Mees M, Hemmingsen SM, Hegedus D. 2002. 5'-RACEing Across a Bridging Oligonucleotide. Biotechniques. 32:480-482.

Suzuki T, Arita T.1995. Two-Domain Hemoglobin from the Blood Clam, *Barbatia lima*. The cDNA-Derived Amino Acid Sequence. J Protein Chem. 14(7): 499-502.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 24: 1596-1599.

Terwilliger NB. 1992. Molecular Structure of the Extracellular Heme Proteins Vol 13. In: Mangum CP editor. Advances in Comparative and Environmental Physiology. Blood and Tissue Oxygen Carriers, Berlin Heidelberg: Spriger-Verlag. 459p.

Terwilliger NB. 1998. Functional adaptations of oxygen-transport proteins. J Exp Biol. 201:1085-1098.

Trotman C, Manning A, Bray J, Jellie A., Moens L., Tate W. 1994. Interdomain Linkage in the Polymeric Hemoglobin Molecule of *Artemia*. J Mol Evol. 38:628-636.

Tunnicliffe V. 1991. The Biology of Hydrothermal Vents: Ecology and Evolution. Oceanogr Mar Biol. 29:319-407.

Weber RE, Vinogradov SN. 2001. Nonvertebrate hemoglobins: functions and molecular adaptations. Physiol Rev. 81:569-628.

Weber RE. 1978. Respiratory Pigments. Physiology of Annelids. P. J. Mill. London, Academic Press Inc.

Yang J, Kloek A., Goldberg D, Mathews F. 1995. The structure of *Ascaris* hemoglobin domain I at 2.2 Å resolution: molecular features of oxygen avidity. P Natl Acad Sci USA. 92:4224-4228.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput App BioSci. 13: 555-556.

Zal F, Lallier FH, Green BN, Vinogradov SN, Toulmond A. 1996. The Multi-hemoglobin System of the Hydrothermal Vent Tube Worm *Riftia pachyptila*. II. Complete polypeptide chain composition investigated by maximum entropy analysis of mass spectra. J Biol Chem. 271:8875-8881.

Zal F, Leize E, Lallier FH, Toulmond A, Van Dorsselaer A, Childress J. 1998. S-Sulfohemoglobin and disulfide exchange: The mechanisms of sulfide binding by *Riftia pachyptila* hemoglobins. P Natl Acad Sci USA. 95:8997-9002.

Zon LI, Youssoufian H, Mather C, Lodish HF, Orkin S. 1991. Activation of the erythropoietin receptor promoter by transcription factor GATA-1. P Natl Acad Sci USA. 88:10638-10641.

Table

Table 1. Pairwise differences between orthologous domains and the SD. Nucleotide differences were calculated by the Kimura-2 parameter model and amino acid differences based on the Dayhoff index matrix. BSY: *Branchipolynoe symmytilida*, BSE: *B. seepensis*, D: domain, SD: single-domain. Shaded values represent the lowest values between the SD and all the other domains.

BSY/BSE	Nucleotides (K2P)	Amino Acids (Dayhoff matrix)
D1/D1	0.030	0.022
D2/D2	0.035	0.058
D3/D3	0.025	0.022
D4/D4	0.065	0.058
SD/SD	0.017	0.007
SD/D1	0.458	0.592
SD/D2	0.506	0.725
SD/D3	0.485	0.653
SD/D4	0.564	0.689
D1/SD	0.462	0.549
D2/SD	0.511	0.713
D3/SD	0.480	0.625
D4/SD	0.554	0.657

Figures captions

Figure S1. Schematic structure of the six initial cDNA sequences obtained for *B. symmytilida*. The coding sequence is shown in white and the 3'UTR in grey. The length of each part of the sequences is shown. BSY: *Branchipolynoe symmytilida*

Figure S2. Northern-Blot results using probes from the sequences BSY3 (A) and BSY7 (B). The size of the obtained bands is indicated (the molecular weight marker is not visible). Lane 1: 20 µg of total RNA; Lane 2: 1 µg of poly A + mRNA; Lane 3: 3 µg of poly A + mRNA.

Figure S3. cDNA sequence of the single-domain globin of *B. symmytilida* (BsySD) and *B. seepensis* (BseSD). The 5'UTR sequence and STOP codon are in lower-case. TATA box highlighted. Intron positions indicated by vertical lines. 3'UTR not shown.

Figure S4. cDNA sequence of the tetra-domain globins of *B. symmytilida* (BsyTD) and *B. seepensis* (BseTD). The 5'UTR and STOP codon are in lower-case. The TATA promoter binding site is highlighted. Domains separated by vertical lines. The 3'UTR not shown.

Figure S5. Different topologies obtained from the various phylogeny softwares and used in the PAML analysis. All trees are unrooted. Used softwares are indicated (the method and nucleotide substitution model are shown between parentheses). The bootstrap or posterior probability (Mr. Bayes) values are shown above/under the branches and the nucleotide substitution model to which they correspond is underlined. (a) to (c): topologies based on nucleotide sequences, (d): topology based on amino acid sequences. Bsy: *B. symmytilida*; Bse: *B. seepensis*; TD: tetra-domain; SD: single-domain; D1-D4: domains 1 through 4.

Figure 1. Globin sequences from annelids, nematodes and a vertebrate. *Branchipolynoe* globin sequences are shaded, TD globins shaded in light grey and SD globins in dark grey. Conserved residues shown in bold (CD1F and F8H), heme pocket residues that may explain high O_2 affinity in *Branchipolynoe* are boxed. Cysteines forming an intrachain disulfide bridge in typical extracellular annelid globins (A2C and H11C) underlined. Cysteine E18 underlined in the TD globins from *Branchipolynoe* spp., and for *Riftia* and *Lamellibrachia*. Intron (I1 and I2) conserved positions shown above the sequences. Bsy: *B. symmytilida*; Bse: *B. seepensis*; TD: tetra-domain; SD: single-domain; D1-D4: domains 1 through 4; AacuNg:

Aphrodite aculeata neuroglobin; Gly: *Glycera* sp.; Lumt: *Lumbricus terrestris*; Tubifex: *Tubifex tubifex*; Phese: *Pheretima seiboldi*; Tylo: *Tylorhynchus heterochaetus*; Rifb: *Riftia pachyptila*; Lam: *Lamellibrachia* sp.; Asuum: *Ascaris suum*; Phyca: *Physeter catodon*.

Figure 2. Bayesian phylogenetic tree based on annelid globins corresponding to the alignment in Fig. 1. White bubble: extracellular globins; light-grey bubble: *Branchipolynoe* globins (single-domain: SD and tetra-domain: TD); dark-grey bubble: annelid intracellular globins. Posterior probability values are indicated above the branches. See Fig. 1 for abbreviations and accession numbers.

Figure 3. Gene structure from the single- (A) and tetra-domain (B) globins from *Branchipolynoe* spp. Promoter region: dark grey; UTR: medium grey; exons: light grey; introns: dark grey with "I" or "Intron". Gene sizes and components are not to scale.

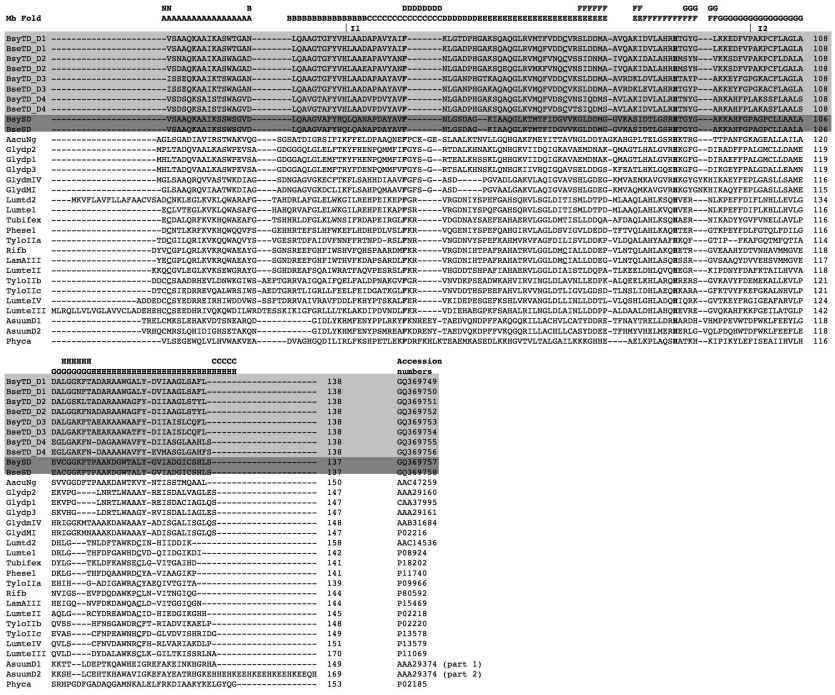

Figure 1. Globin sequences from annelids, nematodes and a vertebrate. Branchipolynoe globin sequences are shaded, TD globins shaded in light grey and SD globins in dark grey. Conserved residues shown in bold (CD1F and F8H), heme pocket residues that may explain high O2 affinity in Branchipolynoe are boxed. Cysteines forming an intrachain disulfide bridge in typical extracellular annelid globins (A2C and H11C) underlined. Cysteine E18 underlined in the TD globins from Branchipolynoe spp., and for Riftia and Lamellibrachia. Intron (I1 and I2) conserved positions shown above the sequences. Bsy: B. symmytilida; Bse: B. seepensis; TD: tetra-domain; SD: single-domain; D1-D4: domains 1 through 4; AacuNg: Aphrodite aculeata neuroglobin; Gly: Glycera sp.; Lumt: Lumbricus terrestris; Tubifex: Tubifex tubifex; Phese: Pheretima seiboldi; Tylo: Tylorhynchus heterochaetus; Rifb: Riftia pachyptila; Lam: Lamellibrachia sp.; Asuum: Ascaris suum; Phyca: Physeter catodon.

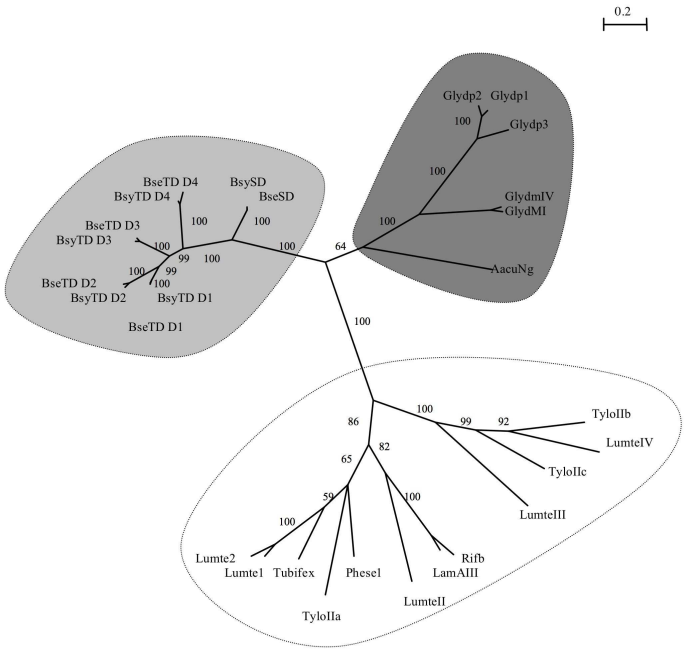

Figure 2. Bayesian phylogenetic tree based on annelid globins corresponding to the alignment in Fig. 1. White bubble: extracellular globins; light-grey bubble: Branchipolynoe globins (single-domain: SD and tetra-domain: TD); dark-grey bubble: annelid intracellular globins. Posterior probability values are indicated above the branches. See Fig. 1 for abbreviations and accession numbers.

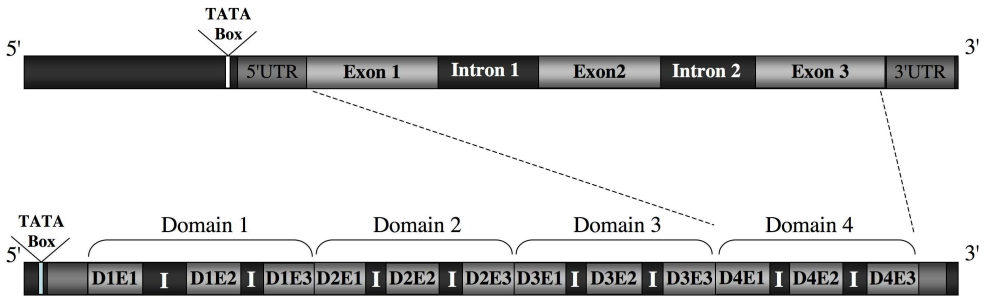

Figure 3. Gene structure from the single- (A) and tetra-domain (B) globins from Branchipolynoe spp. Promoter region: dark grey; UTR: medium grey; exons: light grey; introns: dark grey with "I" or "Intron". Gene sizes and components are not to scale.