Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

The structure of the giant haemoglobin from Glossoscolex paulistus.

José Fernando Ruggiero Bachega, Fernando Vasconcelos Maluf, Babak Andi, Humberto D'Muniz Pereira, Marcelo Falsarella Carazzollea, Allen M Orville, Marcel Tabak, José Brandão-Neto, Richard Charles Garratt and Eduardo Horjales*

Wednesday 12 November 2014

Category: *research papers*

Co-editor:

Professor K. Miki

Department of Chemistry, Graduate School of Science, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan

Telephone:

Fax:

Email: miki@kuchem.kyoto-u.ac.jp

Contact author:

Eduardo Horjales

*De Física e Ciência Interdisciplinar, Instituto de Física/ Universidade de São Paulo, Av. João Dagnone n 1100,
Bairro Jardim Santa Angelina , São Carlos, SP, 13563-120, Brazil*

Telephone: +551633738730

Fax: +551633739874

Email: horjales@ifsc.usp.br

The structure of the giant haemoglobin from *Glossoscolex paulistus*.

José Fernando Ruggiero Bachega[1], Fernando Vasconcelos Maluf[1], Babak Andi[2], Humberto D'Muniz Pereira[1], Marcelo Falsarella Carazzollea[3,4], Allen M Orville[2,5], Marcel Tabak[6], José Brandão-Neto[7], Richard Charles Garratt[1], Eduardo Horjales Reboredo [1].

1- Instituto de Física de São Carlos, Universidade de São Paulo, Brazil, 2 - Photon Sciences Directorate, Brookhaven National Laboratory, Upton, New York, USA, 3-Laboratório de Genômica e Expressão, Departamento de Genética, Evolução e Bioagentes, Instituto de Biologia, Universidade Estadual de Campinas, Brazil, 4-Centro Nacional de Processamento de Alto Desempenho, Universidade Estadual de Campinas,, SP, Brazil, 5-Biosciences Department, Brookhaven National Laboratory, Upton, New York, USA, 6-Instituto de Química de São Carlos, Universidade de São Paulo, São Carlos – SP, Brazil and 7-Diamond Light Source, Harwell, UK.

**E-mail: horjales@ifsc.usp.br*

Short description: The structure of the giant haemoglobin from *Glossoscolex paulistus*.

Abstract

We report the sequences of all seven polypeptide chains from the giant haemoglobin of the free-living earthworm *Glossoscolex paulistus* (HbGp) together with the three-dimensional structure of the 3.6 MDa complex which they form. We report here the refinement of the full particle which has been solved at 3.2Å, the highest resolution reported to date for a hexagonal bilayer haemoglobin composed of twelve protomers. This has allowed for a more detailed description of the contacts between subunits which are essential for particle stability. Our interpretation of features in the electron density maps suggests the presence of metal binding sites (probably Zn^{2+} and Ca^{2+}) and glycosylation sites, some of which have not been reported previously. The former appear to be important for the integrity of the particle. The crystal structure of the isolated d chain (d-HbGp) at 2.1Å shows different inter-chain contacts between d monomers as compared to those observed in the full particle. Instead of forming trimers, as seen in the complex, the isolated d chains associate to form dimers across a crystallographic two-fold axis. These observations eliminate the possibility that trimers form spontaneously in solution as intermediates during the formation of the dodecameric globin cap and contribute to our understanding of the possible ways in which the particle self-assembles.

Keywords: Erythrocruorins; *Glossoscolex paulistus*; Giant extracellular haemoglobin; Hexagonal bilayer.

1. Introduction

Haemoglobins (Hbs) are present in organisms from all five kingdoms of life and have been widely studied, allowing for detailed descriptions of their functional properties. Invertebrates have a wide variety of haemoglobin types ranging from monomeric structures to mega Dalton multimeric complexes. These molecules occur in different anatomical sites with some being extracellular, a fact which usually confers on them special characteristics when compared to those which are present within an intracellular environment (Weber & Vinogradov, 2001). A strong motivation to study these extracellular haemoglobins, especially the giant multimeric complexes such as erythrocruorins, is related to their potential application in medicine as blood substitutes. Studies performed on the giant Hbs from *Lumbricus terrestris* (HbLt) (Hirsch et al., 1997; Elmer & Palmer, 2012) and *Arenicola marina* (HbAm) (Rousselot et al., 2006; Harnois et al., 2009) have demonstrated their greater resistance to oxidation and increased stability as compared to human haemoglobin, highlighting their potential for medical applications. (Elmer et al., 2012b).

The giant extracellular Hbs of annelids, also known as *Hexagonal Bilayer Haemoglobins (HBL Hbs)*, or erythrocruorins, together with chlorocruorins (a variant of HBL Hbs, restricted to four families of polychaetes) are recognized to form a single group according to their physicochemical properties (Vinogradov, 1985). Early studies using electron microscopy applied to haemoglobins from annelids revealed that they are huge complexes, containing a large number of polypeptide chains divided into two different types, globins and linkers, which together form a characteristic quaternary structure. This consists of two superimposed hexagonal layers with a height of ~20 nm and a diameter of ~30 nm, displaying an overall D_6 (or 622) symmetry (Kapp et al., 1982). The structural characterization of the Hb of *L.terrestris* (HbLt) by electron microscopy, revealed that the hexagonal structure is comprised of multimeric units called protomers which present a pseudo 3-fold axis of symmetry. Twelve such protomers are necessary to complete the hexagonal bilayer and the haemoglobin from *L.terrestris* presents the two layers staggered with respect to one another by approximately 16°. This arrangement is called the type I (or non-eclipsed) form. A similar quaternary structure has been found in giant haemoglobins from marine organisms, such as *Riftia pachyptila* (a vestimentiferan), *Alvinella pompejana* (a polychaete), and the clorocruorin from *Eudistylia vancouverii* (also a polychaete) (de Haas et al., 1996a; b; c). However, the structure from *A.pompejana* (HbAp) lacks the 16° rotation between the two hexagonal layers. This finding has been extended to several other Hexagonal Bilayer Haemoglobin structures from polychaetes, including the species *Arenicola marina*, HbAm (Royer et al., 2007). This suggests that polychaetes, including those which inhabit the deep ocean (but excluding the four families containing chlorocruorins), present structures with no staggering between the two layers. This second arrangement is termed eclipsed or type II.

Vestimentiferans have two types of extracellular Hbs. One, named "heavy", has a molecular mass between 3.0 and 4.0 MDa, corresponding to an hexagonal bilayer structure, while the other, named "light", is of the order of 400 kDa. Pogonophorans (a group of animals close to vestimentiferans and polychaetes) on the other hand, present only the "light" form, as described below. Two crystal structures have been obtained for these lighter complexes, from the pogonophoran *Oligobrachia mashikoi*, Hb_LOm *(Numoto et al., 2005, 2008)* and from *R.pachyptila (Flores et al., 2005)*. In both cases the complex is composed of six copies each of four globin subunits (A1, A2, B1, and B2) forming a pair of dome-shape structures associated to form a hollow spherical assembly. The oxygen binding properties of these light complexes are qualitatively similar to those of annelid giant Hbs. Both the oxygen affinity and cooperativity of Hb_LOm are enhanced by the addition of Ca^{2+} and/or Mg^{2+}, or by an increase in pH. These similarities have been related to the similar structure of the tetrameric units of the four globin chains present in all of these complexes.

The crystal structures of the erythrocruorins from *L.terrestris* (HbLt) at 3.5 Å (Royer et al., 2006) and *A.marina* (HbAm) at 6.2 Å (Royer et al., 2007) have been reported and are representatives of the type I and type II arrangements respectively. With the crystal structure of HbLt, *Royer et al.*

greatly contributed to clarifying the relative stoichiometry of the constituent subunits, and also to the hierarchical packing of the complex. In the structure there are four different types of haem-containing globin chains, named *a*, *b*, *c* and *d*, which together, form a hetero-tetramer, in which *a, b* and *c*, are linked by disulphide bonds. The heterotetramer is then repeated three times to form a structure with C3 symmetry called the dodecamer or "cap", with the following stoichiometry, $[ab\mathrm{c}]_3[d]_3$. This dome-shaped structure, with a molecular mass of approximately 200 kDa, has also been characterized in isolation by X-ray crystallography. The structure solved at 2.4 Å (Strand et al., 2004) gave a detailed model of the cap, showing that it is very similar to those found in the extra cellular "light" Hbs of vestimentiferans and pogonophorans.

In the case of HbLt, as well as other hexagonal bilayer haemoglobins and chlorocruorins, the dodecameric cap is associated with a hetero-trimer of linker chains formed of a single copy each of L1, L2 and L3. This structure with stoichiometry $([abc]_3[d]_3)$(L1L2L3) forms the previously mentioned protomer. The globular parts of the linkers, together with the cap, form the head of the protomer from which the tails of the linkers protrude into the centre of the particle generating a mushroom-like appearance, In summary, therefore, the hexagonal bilayer is composed of twelve protomers, including a total of 144 globin chains and 36 linkers. The interactions formed by the latter are primarily responsible for stabilizing the structure of the entire particle.

The oxygen binding properties of erythrocruorins exhibit features significantly different from those of vertebrate tetrameric haemoglobins. Heterotropic effectors are particularly different. Whilst the cooperative oxygen binding properties of vertebrate tetrameric haemoglobins are enhanced by small organic anions or chloride, in the case of annelid giant haemoglobins inorganic cations are the best characterized heterotropic effectors. In general, divalent cations are more effective than monovalent species, accompanying the following series: $Ba^{2+} > Ca^{2+} > Sr^{2+} > Mg^{2+} > Li^{+} > Na^{+} > K^{+}$, (Weber & Vinogradov, 2001). This effect not only depends on the ionic strength, but also appears to involve the ionic radius of each species. Besides revealing the intricate ways in which all 180 subunits associate in order to form a stable and biologically functional unit, the crystal structure also provided information concerning the presence and localization of the calcium binding sites (Royer et al., 2007). Little is known about the effects of other metal ions but the presence of the divalent cations Zn^{2+} and Cu^{2+} has been identified in the chemical composition of Hbs samples from many oligochaetes, including HbLt (Standley et al., 1988). Furthermore, the addition of Zn^{2+} increases the affinity for O_2, but in a manner different to that observed for the other divalent cations mentioned above. Zn^{2+} increases the association constant of O_2 for the T-state (Ochiai et al., 1993) but without affecting that of the R-state, suggesting a distinct mechanism of action from other divalent cations.

The giant extracellular haemoglobin from *Glossoscolex paulistus* (HbGp), an earthworm found in the south-east of Brazil, is similar in structure to HbLt and is one of the best characterized erythrocruorins (Santiago et al., 2010a; b; Carvalho et al., 2013). HbGp has a molecular weight of 3.6 MDa (Carvalho et al., 2009) and consists of 144 haem-containing subunits with molecular masses in the range of 15–19 kDa and 36 linker subunits with masses between 24 and 26 kDa. Preliminary results of the crystallization, x-ray diffraction data collection and molecular replacement solution of the HbGp complex (using the whole protomer structure from HbLt as a model), have been reported previously (Bachega et al., 2011). Like HbLt, the hexagonal layers of HbGp are also staggered with respect to one another by approximately 16° which characterizes this structure as type I. Several other details were observed in agreement with the HbLt, such as the presence of disulphide linkages connecting the a, b and c chains within the cap and similar calcium binding sites. Nevertheless, until the present work, the only sequence available for HbGp has been that of the globin d subunit (Bosch Cabral et al., 2002), - which made impossible the refinement and interpretation of the crystal structure. We report here the amino acid sequencing of all of the seven chains which comprise the complex together with the full hexagonal bilayer structure refined at 3.2 Å resolution, its description and the comparison with the homologous structure of HbLt.

2. Materials and Methods

2.1. Sequencing of the HbGp subunits

Total RNA (< 10 μg) was extracted from *Glossoscolex paulistus* using the RNeasy kit from Qiagen. The quality and quantity of RNA was determined by monitoring the absorbance at 260 nm and on 1% agarose gels under denaturing conditions. It was subsequently subjected to cDNA synthesis prior to sequencing. The cDNA was sequenced using a high throughput sequencing platform (Illumina HiSeq 2000) generating 2x100 bp paired-end reads (250 bp of insert length). Sequencing was performed at the *High Throughput Sequencing Facility (HTSF) of the University of North Carolina - Chapel Hill. De novo* assembly of paired-end reads was performed using the Trinity assembler (Grabherr et al., 2011) setting the maximum insert parameter to 300 bp and considering a minimum contig size of 500 bp. The polypeptide chains of *Lumbricus terrestris* haemoglobin were used to identify homologous transcripts within the *Glossoscolex paulistus* transcriptome assembly using BLASTx (Altschul et al., 1997), applying an e-value cut-off of 10^{-5}. Hits were subsequently manually verified.

2.2. Structure determination and crystallographic refinement of the full biological particle.

Some of the details of the data collection and structure solution have been described in our preliminary report (Bachega et al., 2011). Briefly, the whole HbGp complex was purified directly from adult earthworms and crystallized as previously described (Bachega et al., 2011). X-ray diffraction images were collected at 93°K with an ADSC Quantum 315 detector using synchrotron radiation of wavelength 0.920 Å at beamline X29A of the NSLS (Brookhaven National Laboratory, Upton, New York, USA).

Diffraction data processing to 3.2 Å was performed using iMOSFLM (Leslie, 2006; Battye et al., 2011) and SCALA from the CCP4 package (1994). The data collection and processing statistics have been published previously (Bachega et al., 2011). Molecular replacement with Phaser (McCoy et al., 2007) employing a single protomer from HbLt (PDB code 2gtl) (including both main and side chains) was used to determine the location of the three protomers within the asymmetric unit. The haem groups were omitted during this procedure. After amino acid substitution (where necessary), the program PHENIX (Adams et al., 2002) was used to refine the structure. Three runs of 3 cycles each were alternated with local real space refinement and visual evaluation with the program COOT (Emsley & Cowtan, 2004). It should be borne in mind that the asymmetric unit contains three protomers, each of which is composed of three [*abcd*] tetramers and an [L1,L2,L3] trimer of linkers. The large size of the structure and the limited resolution (3.2 Å) therefore required some special care as follows:

a) Non crystallographic symmetry (NCS) was essential to inhibit over-refinement. During the first 3 cycle steps using PHENIX we tested two different possible sets of NCS. The first used a full protomer (three globin tetramers and three linkers) as the reference structure whilst the other used only one tetramer and three linkers as reference structures (the seven unique chains). The first set of NCS did not generate significant differences among the three equivalent globins within the protomer (for example among the different copies of the *a* chain) and the refinement evolution was slower than the second set. Thus we continued the refinement using NCS based on one sequence-one reference structure. This procedure required 13 Gb of RAM memory to run. We used a computer with a Intel CPU Core I7 2600 (sandbridge) with 16Gb RAM memory running linux UBUNTU. Under these conditions the

X-ray/*stereochemistry* weight was adjusted to optimize simultaneously the geometry of the model and the fit to the X-ray diffraction data.

b) During the first PHENIX run, a wider radius was defined (A larger) in the bulk solvent mask calculation in order to avoid losing surface details such as the sugars that clearly became well defined later in the refinement process.

c) During the final steps of refinement we reduced the NCS restraints from medium to lose without changes to the observed RMSDs.

2.3 Crystallization, structure determination and crystallographic refinement of the isolated *d* chain.

The isolated *d* chain of HbGp (d-HbGp) was purified according to a method previously described (Cabral *et al,* 2002). Oxy-state crystals were obtained by hanging drop vapour diffusion at 291K in the presence of 1.4 M sodium citrate, 50 mM Tris-HCl, pH 7.5. The addition of 10% ethylene glycol to the original growing condition was used for cryoprotection. Diffraction data from crystals of d-HbGp were collected at the Diamond Light Source, beamline I04-1, equipped with a Pilatus 2M detector using a wavelength of 0.9200 Å, oscillation range of 0.4 ° and a 1 sec. exposure time per image at a temperature of 80 K. The search model used for molecular replacement with PHASER was the isolated *d* subunit from the full particle structure solved at 3.2 Å and was refined using PHENIX, reserving 10% of the reflections for the calculation of R_{free}.

3. Results and Discussion

3.1 Sequence determination and analysis.

The *ab-initio* transcriptome assembly generated 97,639 contigs larger than 500bp representing the most expressed isoforms of *Glossoscolex paulistus*. The *Lumbricus terrestris* haemoglobin were used to identify homologous transcripts within the *Glossoscolex paulistus* resulting in six complete globin chains (Figure 1) and five complete linker chains (Figure 2), all of which included N-terminal sequences rich in hydrophobic residues (Table S1). These were predicted to be signal peptides by the signalP 4.1 server (www.cbs.dtu.dk/services/SignalP/), which are responsible for the migration of these subunits to the extracellular environment and which are subsequently edited. The assignment of the different chains was attributed according to their sequence identity when compared to their counterparts in HbLt. The six globin sequences correspond to subunits *a1*, *a2*, *b*, *c*, *d1* and *d2,* where *a1* and *a2* are isoforms of the *a* subunit and *d1* and *d2* are isoforms of the *d* subunit. HbLt also presents isoforms of the *d* chain (Xie et al., 1997) which share 80% sequence identity whereas in *G.paulistus* this value falls to only 58%. The isoform we name *d1* is equivalent to the *d* subunit sequence previously reported (Bosch Cabral et al., 2002). As expected the ½-cystine residues which form both intra- and inter-molecular disulphide bridges are conserved in all isoforms.

The complete dataset of RNA-seq reads has been deposited in SRA under accession number SRR1519963.

Overall, the globin chains share sequence identities of at least 45%, similar to that observed for HbLt. However, comparisons made between equivalent globin subunits in HbGp and HbLt, show a greater level of sequence identity for the *a*, *b* and *c* chains than was previously estimated based on the *d1* sequence reported by Cabral *et al.* As also observed for HbLt, the sequences from HbGp show no D helix, a region which is highly variable in globins in general. As might be anticipated, the residues around the sixth coordination site are highly conserved. These are positions 37, 51, 67, 71 and 113 using the numbering system given in the alignment of Fig. 1. With the exception of position 67, which corresponds to the distal histidine, the remaining four positions form a hydrophobic pocket which is also conserved in human globins (not shown in the alignment).

For the sequences identified as linkers, three were readily classified as L1, L2 and L3. These sequences share about 60% identity with their homologues in HbLt but only approximately 30% amongst themselves. A fourth sequence was also identified which is similar to that of L4 in HbLt. Here we consider it to be an isoform of the L3 subunit and have named it L3b. The sequence identity between L3 and L3b is 43%, and neither share more than 30% with the remaining linkers. The fifth and final sequence identified shares 35% and 34% respectively with L1 and L2 and is thus ambiguous with respect to its most appropriate classification. Here it has been arbitrarily named L1b.

A comparison of the predicted physicochemical properties of the sequences of HbGp and HbLt shows that there are no major differences with respect to the total masses and isoelectric points (Table S1**.** The N-terminal regions of the three linker subunits are characterized by long α-helices which come together to form a trimeric coiled-coil. This plays an important role in the formation of the hexagonal bilayer as will be described below. This region shows the characteristic heptad repeat pattern (*abcdefg*) in which hydrophobic residues at *a* and *d* form the interior of the coiled-coil. Figure 2 shows the alignment of the HbGp linker sequences compared to other annelids, highlighting the residues that form the coiled-coil in the N-terminal portion. Type I (uneclipsed) erythrocruorins are characterized by a deletion at positions 65 and 66 which results in a breakdown in the phase of the heptad repeats whereas type II erythrocruorins (HbAp and HbAm) retain the phase of the coiled coil throughout this region. This is consistent with low resolution maps of the structure of HbAm (Royer et al., 2007) which show a contiguous α-helix rather than one divided into two segments as seen in the type I structures (see below). This structural difference is believed to be related to the formation of eclipsed or uneclipsed particles. The presence of gaps at positions 65 and 66 in all HbGp linker sequences is thus consistent with it being a type I erythrocruorin, as confirmed by the crystal structure described below. The remaining two domains of the linkers (the LDL-A domain and the β-barrel domain) can be readily delimited from the alignment with the HbLt sequences.

In Table S1 the experimental masses obtained by MALDI-TOF-MS analysis [Oliveira et al., 2007; Carvalho et al, 2011] are summarized together with those derived from the sequences reported here.. For the linkers L1 and L2 the masses are quite similar. The positive difference observed for the experimental masses could be due to either glycosylation, as observed for L1, L2 and L2b, or to cations, such as Ca^{2+} or Zn^{2+}, bound to the corresponding chain. For L3 the mass from the sequence is quite different from that from mass spectrometry. Probably, L3 was not detected in mass spectrometry experiments due to its low concentration, and the mass of 32kDa assigned as due to a mixture of L3 and a dimer of d monomers, corresponds solely to the latter. In the case of chains a and d mass spectral analysis detected up to four isoforms, two of them being predominant. As noticed in Table S1, masses of chains a and c were exchanged in the experimental mass assignments. The mass differences (between mass spectrometry and DNA sequencing) for the a isoforms suggest the presence of glycosylation, which has been observed also for the a chains of HbLt [Ownby et al, 1993]. The largest discrepancy is noticed for chain c. The explanation could be due to multiple glycosylations at additional sites, or longer chains at fewer sites. The latter possibility is consistent with disorder as the glycosylation chain length extends further. Disorder is not visible in crystallography.. In summary, considering all of the effects, described above, the agreement between the directly determined experimental masses and those derived from the deduced amino acid sequences is rather good.

3.2. Crystallographic refinement

A preliminary report has already described the basic structure solution (Bachega *et al.,* 2011). Briefly, the problem of phase ambiguity and the space group (I222 or $I2_12_12_1$) was solved by molecular replacement using the HbLt protomer (PDB code 2gtl) as template. The search model therefore corresponded to only 1/12th of the biological particle from which the haem groups and expected non-haem metals had been excluded. Subsequently, the appearance of density corresponding to these groups was used as an evaluation criterion for the molecular replacement solution. Three protomers were located in the asymmetric unit, a quarter of the biological particle, with two coming from one hexagonal disc and one from the other. The biological unit (with 622 point group symmetry), is therefore generated by the crystallographic symmetry present in space group I222. This requires that the particle sits on a special position, namely the intersection of the three 2-fold axes. The location of the first protomer in the asymmetric unit generated an R value of 57.2% and an LLG of 1,446. The location of the second protomer resulted in an R value of 53.5% and an LLG of 5,509, and subsequently when the third and final protomer was located, the R value dropped to 49.6% and LLG rose significantly to 12,050.

The initial electron density maps (2Fo-Fc) showed clear signs of many side chains and haem groups enhancing confidence in the MR solution. The asymmetric unit consists of a total of 45 subunits (36 globins and 9 linkers) totalling 7,200 amino acid residues. In the a, L1, L2 and L3 subunits, the first 2, 3, 17 and 5 N-terminal residues respectively are not observed in the electron density maps. In all other subunits the main chain is complete if we use the predicted cleavage position of the signal peptide as a criterion for determining the N-terminus (Table S1). Where different isoforms existed, the choice of sequence was taken to be that which was most consistent with the calculated electron density. Thus, all *a* and *d* globin subunits were modelled using sequences *a2* and *d1*, respectively (Figure S1 - supplementary material). Linkers L1 and L3 were best accounted for by the sequences labelled as such in Figure 2, with the isoforms L1b and L3b considered less compatible with the maps.

The limited resolution and large number of atoms in the asymmetric unit (59,091) demanded the use of NCS restraints and group temperature factor refinement. The final R and R_{free} values were 21.7% and 23.5% respectively. The weights applied in order to maintain standard stereochemistry during refinement can be considered satisfactory, since R_{free} fell at each stage of the refinement process reaching an acceptable final value (Kleywegt & Jones, 2002) and simultaneously resulted in good RMSDs for both bond lengths and angles. The Φ/Ψ distribution is also consistent with that expected for structures of similar resolution. The main data for the refinement process is presented in Table 1.

In the region of the haem groups, the electron density is sufficiently clear to determine the orientation of the haem plane in all subunits. Since the sample crystallized was the cyano form of HbGp, a CN moiety was modelled as the sixth ligand to the iron atoms. However, it should be mentioned that this density is insufficiently clear to be able to unambiguously establish the precise chemical nature of the sixth ligand. This is due to a series of factors including the proximity of the density to the iron, the poor resolution of the data and to redox effects due to the ionizing radiation (Andi *et* al, in preparation). The electron density map also confirmed the presence of Ca^{2+} ions coordinated to the LDL-A domain in all linker chains (Figure 3a). These are present in equivalent positions to those found in the linkers of HbLt. Electron density for all three L3 type subunits also revealed two further peaks indicating the presence of other heavy atoms (Figure 3b and 3c). These were modelled as Zn^{2+} ions after taking into account the nature of the residues that comprise their chemical environment (particularly the first coordination sphere), and due to the fact that Zn^{2+} was detected in the sample by Inductively Coupled Plasma Atomic Emission Analysis (data not shown).

On the other hand, the only Zn^{2+} ion present in the crystal structure of HbLt lies in the cavity ("mouth") formed by the β-barrel of subunit L2 (see below). The amino acid residues which compose this site are conserved in the L2 sequence of HbGp, and the Fo-Fc difference maps indicate the presence of a scattering ion at this position. However the chemical environment is rich in positively charged residues (Arg123 and Arg209) and seems therefore inconsistent with the presence of a Zn^{2+} cation. This density was left uninterpreted but we suggest that it is unlikely to be a metal ion as reported previously.

The sequences of L2 and L3 are predicted to be N-glycosylated based on the amino acid sequences. However, the electron density is only able to confirm glycosylation to subunits L3 at Asn121 which was modelled as an N-acetylglucosamine (Figure 3d). The glycosylation and the presence of the Zn^{2+} site are strong evidence that it is the L3 sequence which is dominant for the crystallographic model of HbGp, since the sequence of the L3b isoform lacks both of these features.

3.3. Complex organization (structural overview).

The structure of HbGp is based on a two-layered hexagonal doughnut in which each layer consists of six protomers. HbGp, like HbLt, has a relative rotation of 16° between the two hexagonal layers (Royer *et al.*, 2000). This characterizes the HbGp erythrocruorin as being of the type I, or "non-eclipsed" form consistent with the idea that all earthworms present erythrocruorins of this type (Weber & Vinogradov, 2001). The biological unit (or full particle) is a 180 subunit structure (144 globins and 36 linkers) with an overall diameter of approximately 290 Å and a height of 190 Å (Figure 4). The protomers, which form the two hexagonal layers, are oligomeric structures with a mushroom-like format and which are approximately 130 Å in length. For descriptive purposes, the protomer can be conveniently divided into two parts; the cap (a dodecamer of globins), and a hetero-trimer of linkers. The cap is a dome shaped structure, as observed in HbLt, and is composed of a trimer of tetramers, $(abcd)_3$, where the subunits a_1, b_1, and c_2 (*c* from a neighbouring tetramer - see below) are connected by disulphide bonds. The trimer of linkers is a stoichiometric association of L1, L2 and L3 subunits (labelled M, N and O respectively in the PDB file), forming a funnel shaped structure whose "head" (the C-terminal portion, composed of an LDL-A domain and a β-barrel domain) interacts with the dodecameric globin cap. The N-terminal tails of the trimer form coiled-coils which point, like spokes, towards the centre of the structure. Figure 4 establishes the standard colours used in this work to distinguish between the different kinds of subunit as well as the various levels within the structural hierarchy which will be described below.

3.4. Globins and the cap formation.

In HbGp, the globin chains present the typical fold described by Perutz (see for ex. Perutz, M. 1989) containing all the expected secondary structures with the exception of the small D helix, which is replaced by a short loop. The C helix is a 3_{10} helix, as in mammalian globins. Like other annelids, the globins in HbGp have a significantly larger F helix when compared with human globins. This is due to an insertion of three residues within the short loop connecting helices E and F, which leads to an N-terminal extension to the F-helix. The presence of hydrophobic residues forming the haem group cavity is conserved, and particularly noticeable is the aromatic residue at position 37 in the *a*, *b* and *d* subunits (Figure 1).

The first level of oligomeric association between globin chains is the generation of the *a*/*d* and *b*/*c* dimers. These are equivalent structures, where the interface occurs directly via haem groups and residues located in helices E and F (residues Gln98 and His99) as can be seen in Figure S2. The association between a pair of these dimers results in a hetero-tetramer (Figure S2) in which subunit *a* forms a disulphide bond with subunit *b*. The hetero-tetramer is repeated 3 times to form the dodecameric cap. The contacts between hetero-tetramers that generate the cap occur between subunits a_1 and c_2, (including a disulphide bridge) and between d_1, d_2 and d_3 (where the subscripts refer to different tetramers). The latter contacts result in a trimer of *d* subunits sitting on the three-fold axis of symmetry which the cap presents. It is known that the *abc* trimer when isolated from the entire particle of HbLt displays a small cooperative behaviour (C_{Hill} ~ 1.3) (Fushitani & Riggs, 1991), while the fraction containing the *d* subunit displays no cooperativity. However, when the *abc* trimer is re-joined with the *d* subunit, the resulting cap displays, at pH 6.8, a cooperative behaviour that is indistinguishable from the whole particle (C_{Hill} = 7.8). Thus, the structure of the cap is pointed to as being the basic cooperative unit in HbLt. This fact is further supported by the "light" Hbs (HB_L), found in vestimentiferans and pogonophorans (Negrisolo et al., 2001). Despite the low sequence identity between the subunits in HB_L and those of HbGp and HbLt, the same hierarchal association is preserved resulting in a dodecameric structure equivalent to the cap, suggestive of a similar cooperative mechanism. In $HB_{L,}$ two dodecamers join to form a 24-mer with a hollow cavity. The superposition of these dodecamers with the caps of HbGp and HbLt yields RMSDs between 1.3 and 1.6 Å on C_αs where the orientation of the subunits is effectively identical. Figure 5 shows a comparison between the structure of the cap in HbGp and the equivalent structures in *L.terrestris*, *Oligobrachia mashikoi* (Hb_LOm) and *Riftia pachyptila* (Hb_LRp). The latter two present metal binding sites (Zn^{2+} in Hb_LRp and Ca^{2+} in Hb_LOm) that are not conserved in the land-living species, indicating that these metals may have a different influence over cooperativity and allosterism in these systems.

3.5. Linkers and the heterotrimer.

The three types of linker in HbGp show a similar structure, divided into four domains (Royer et al., 2006). The first two domains are responsible for forming the coiled-coil and consist of two helices connected by a loop. These loops are the most variable regions of the coiled-coil structure when comparing one linker to another. This variation leads to differences in the relative orientations of the two helices in different linkers (Figure S3) and is necessary for producing the inclined orientation of the second coiled coil region with respect to the first. This variation is conserved on comparing HbGp with HbLt suggesting that it is relevant for the correct assembly of the trimer (Figure S3). The third domain is cysteine-rich and, similar in fold to LDL-A domains, as observed in the LDL receptor. The LDL-A domain contains the Ca^{2+} binding site formed by six oxygen ligands, forming a distorted octahedral complex. Such a configuration is also compatible with a Mg^{2+} ion, and it is worth noting that Mg^{2+} has a very similar effect on HbGp. However, ICP analysis of the cyano-HbGp sample, used in this work for crystallization, excluded the presence of magnesium. The fourth and final domain is a typical eight-stranded antiparallel β-barrel which plays an important role in stabilizing the protomer. This globular domain is responsible for the main interactions between the trimer of linkers and the cap.

In the L3 subunits, besides confirmation of the Ca^{2+} site in the LDL-A domain, two other intense peaks were found which were attributed to Zn^{2+} ions, based on coordination environment and ICP analysis (which exclude other metals different than Ca and Zn). The first site (zinc site 1) resides between the LDL-A domain and the β-barrel domain. In this case the Zn^{2+} is coordinated by the side chains of His64, Asp90, His94 and the fourth ligand has been modelled as a water molecule (Figure 3a). Based on the alignment (Figure 2), this site is present only in HbGp. The second Zn^{2+} site, is distant from the first and found between subunit L3 and L2 from neighbouring protomers of the same hexagonal layer (a fuller description of the interactions made between different protomers will be given later). This site is formed by His142 and His144 from L3, His236 from L2 and a water molecule (Figure 3b**).** Liochev and co-workers reported a superoxide dismutase activity for HbLt, which is in agreement with the presence of Cu^{2+} and Zn^{2+} species experimentally determined for this system (Liochev et al., 1996). This activity was attributed to linkers, since the isolated globin portion showed no SOD activity. However, our ICP analysis showed no evidence for copper in HbGp and there is also no evidence in our structure for a bimetallic site as seen in Cu/Zn superoxide dismutases. Our results shed no light on the observations of Liochev *et al*, which require further investigation and may be a peculiarity of HbLt and not erythrocruorins in general.

The three different linkers interact to form a trimer stabilized principally by hydrophobic interactions from the first two domains which form the coiled-coil. This is directed towards the centre of the structure and is important for particle stability. The coiled coil is interrupted around residue 45 (alignment position 65 in Fig. 2), dividing it into two portions (domains 1 and 2). The first helix is the larger of the two and corresponds to approximately four heptad repeats making eight sets of hydrophobic contacts (four at the *a* position and four at the *d* position). There is no clear pattern of β-branched and non β-branched residues at the *a* and *d* positions, an observation consistent with the formation of a trimeric coiled-coil over a dimer or tetramer (Harbury et al., 1993). Overall, leucine and isoleucine are the most abundant residues but at the first *d* level two hydrophilic residues, Gln17 and Asn24 are observed in the L1 and L2 chains respectively. This arrangement is stabilized by direct hydrogen bonding which also extends to include Asn15 from the *e* position of the equivalent heptad in the L3 chain (Figure 6). These hydrophilic residues are conserved in sequences from other species suggesting they represent a common feature. At the C-terminus of the first helix, Arg35 which occupies a *d* position in the L3 chain is charge compensated by the carbonyl groups coming from the remaining two *d* residues, Leu38(L1) and Ala45(L2). This arginine is well conserved in the L3 chains of different species and its role in helix disruption is presumably of structural importance. However, different from *G.paulistus*, the sequence of *A.marina* in this region shows a contiguous pattern of heptads (Fig. 2), consistent with the observed presence of a continuous helix in the crystal structure (Royer *et al.,* 2007). The disruption of the coiled coil may therefore be necessary for determining the relative offset of the protomers from one disc with respect to the other, which is the major structural difference observed between type I particles such as that observed here for *G. paulistus* and type II particles seen in *A. marina*.

Aspartic acid residues are found in all three linkers at the beginning of the second helix but due to variation in the connecting loop these do not occupy equivalent positions in the sequence alignment. In the case of the L2 and L3 chains, these are well conserved and form classical N-caps

(Richardson & Richardson, 1988). It is of note, however, that the L3b isoform lacks this aspartic acid (alignment position 61 in Fig 2) which together with other sequence differences in this region suggests possible conformational variation. The second helix is shorter and composed of only two heptad repeats. At the first *a* level, the helices are splayed apart such that a fourth residue (Phe46(L2) occupying a *g* position) contributes to the hydrophobic core. At the C-terminal end, the coiled coil is stabilized by classical Arg (at *g*) – Glu (at *e*) salt bridges.

3.6. Interactions between the dodecameric caps and linkers.

The protomer is formed by the interaction between the cap and the trimer of linkers, producing a structure with pseudo three-fold symmetry. The contacts between the cap and linkers occur between the outer rim of the head (globular region) of the linker trimer and the inside edge of the cap, with the LDL-A and β-barrel domains of each linker interacting with one of the globin trimers [$a_1b_1c_2$]. As a result of the dome shape of the cap, a large cavity is created within the protomer (Figure S4). The cavity has a volume of around 40,000 A^3 which is approximately twice the volume occupied by a globin subunit alone. There are 5 different types of channel which give access to the cavity, giving rise to 13 in total. The first channel has a diameter of approximately 6.5 Å and is unique in that it lies along the three-fold axis at the top of the cap, formed at the interface between the *d* subunits. The second group corresponds to the channels formed at the interface between neighbouring *abcd* tetramers and has a diameter of approximately 7.0 Å. In the Hb_LOm structure, this channel is occupied by a Ca^{2+} binding site that is not conserved in HbGp or HbLt. The third group is formed at the interface between the second helical domain, the LDL-A domain and the β-barrel domain of a given linker. The fourth group of channels is formed at the interface between two linkers, specifically between the β-barrel of one linker and the LDL-A domain of its neighbour, resulting in small spaces that give access to the central cavity. The group 3 and group 4 channels have average diameters of 10 Å. The fifth and final group of access channels is the result of the interaction between the cap and the heterotrimeric linkers. This channel is formed at the interface between the *bc* globin dimer, the β-barrel from one linker and the LDL-A domain of an adjacent linker. These latter channels are the widest, having diameters of the order of 15Å. No significant electron density is observed within this cavity providing no evidence for the presence of small ligands. However, the multiple access channels which arise from defects at subunit or domain interfaces is suggestive of a carrier function for the cavity. However, this is speculation and if the cavity has any physiological significance, it remains to be elucidated.

3.7 Interactions between protomers from the same layer.

The linker subunits provide the primary contacts between the protomers which generate the full biological particle. The lateral contacts between protomers within the same hexagonal layer occur mainly between L2 of one protomer and L3 from another. The main contacts are hydrogen bonds provided by the anti-parallel alignment of β-strand 2 from L2 with β-strand 4 from L3 of the neighbouring protomer. The contact is further stabilized by a salt bridge between Lys113 (L2) and Glu159 (L3). The main difference with respect to HbLt is the presence of zinc site 2, as described above and shown in Figure 7. The three histidine residues which form the metal binding site (His 142 and His 144 from L2 and His 236 from L3 of the neighbouring protomer) are conserved in other terrestrial species (Figure 2), suggesting the presence of Zn^{2+} at this interface in all of them.

Zn^{2+} has been described as being different from other divalent cations in terms of its effector properties, at least in some species. According to Ochiai and co-workers, in Pheretima hilgendorfi, Zn^{2+} binds in a different site from the alkaline earth metals, and acts on the T-state increasing its affinity for oxygen (Ochiai et al., 1993). It was also shown that zinc is capable of increasing the stability of the particle under conditions which otherwise lead to fragmentation of the complex, for example at alkaline pH. The conservation of the residues involved at the second zinc site and its

strategic location between protomers suggest that it may well be this site which is responsible for contributing to the stability of the particle.

Lateral contacts between protomers also include minor interactions between adjacent caps. Two contacts are observed, the first involving the a chain with its counterpart from the neighbouring protomer, and a second involving subunit b with subunit c. The latter involves the loop connecting helices F and G of the b chain with helices G and H from the c chain, including a. salt bridge between residues Asp102 of subunit b and Lys108 of subunit c. In summary, there are significant contacts between caps raising the possibility of communication between globin chains of different protomers.

3.8. Interactions between protomers from different layers.

The specific interactions made between protomers from different hexagonal layers occur between those which are effectively superposed (ignoring the small relative rotation of 16^o of the two layers with respect to one another which is characteristic of the type I arrangement). These protomers are related by one of the two families of 2-fold axes present in the point group 622. The interaction involves the globular region of linkers L1 together with the helices of the first domain from linkers L2 and L3 (part of the coiled coil). It is worth mentioning that there are no interactions between the caps so that the two hexagonal layers interact only through the linker subunits and the globin chains in different layers are therefore completely independent. The contact between the β-barrels of the L1 subunits (involving β-strands 1 to 5) results in a buried surface area of 1,102 A^2, and forms at least 16 hydrogen bonds. The coiled-coil structures intersect at an angle of 100^o, and the interface is characterized by a high concentration of charged residues forming ionic interactions, six in total (figure 8).

The last interaction between protomers of different hexagonal layers occurs between those which are related by the second family of 2-fold axes present in the point group 622 (D_6), which describes the symmetry of the particle. These are made between a given protomer and the neighbour of its superposed partner in the other layer. These contacts are made by their respective coiled coils which also pack at an angle close to 100^{o} and involve exclusively the L1 subunits. Due to the symmetry involved residues Thr12, Asn16 and Gln17 of each subunit end up generating four hydrogen bonds at this interface (Figure 8).

From the description above it is clear that the linkers have a key role in stabilizing the full biological particle and particularly in maintaining the two hexagonal discs together. This is emphasized by the fact that no contacts occur between coiled-coils of the same layer. The interactions described above lead to each coiled-coil being sandwiched between two other coiled-coils from adjacent protomers of the opposite layer. This interdigitation creates a network of interactions which appears to lend great stability to the complex and is presumably the principal role of the linkers, depending little or not at all on the globins. This is emphasized by the observation of cap-free hexagonal bilayers, composed solely of linkers, in electron microscopy images (data not shown).

3.9. The structure of d-HbGp subunit and the cap formation.

The structure of the isolated globin *d* chain (*d*-HbGp) was solved independent of the remainder of the structure and presents a global fold very similar to that observed in the full HbGp complex, preserving all the elements of secondary structure. The isolated *d* chains from both HbLt and HbGp are largely monomeric although a small fraction is present in the form of oligomers whose presence is concentration dependent. Initially it was imagined that this small oligomeric fraction could be a trimer (as observed in the full particle), which might act as a nucleating agent for cap formation. However, *d*-HbGp crystallizes with a monomer in the asymmetric unit of space group I222, precluding the possibility of trimers in the crystal structure. On the other hand, one of the symmetry mates leads to the generation of a dimer (*d/d*) employing an equivalent interface to that made between the *a* and *d* subunits of the full particle (Figure 9). This interface principally employs the E and F helices which sandwich the haem group. The measured buried surface area for the *d/d* interface is 646.8 Å^2, somewhat lower than that presented by the native *a/d* interface (914.5 Å^2). Nevertheless, the two interfaces present a similar estimate for the interaction energy, (-11.0 and -13.7 kcal/mol respectively) as calculated by PISA (Krissinel & Henrick, 2007). By contrast, the trimeric structure d_x3, present in the entire particle as part of the dodecameric cap, presents an interface between adjacent subunits of only 331.4 A^2 and an estimated interaction energy of -3.6 kcal/mol. These data, combined with the interfaces found in the crystal lattice of *d*-HbGp show that the existence of a trimeric species is not favoured in solution, suggesting the predominant oligomeric forms are dimers or possibly multiples of dimers.

This is broadly consistent with experimental evidence from MALDI-TOF-MS studies [Oliveira et al, 2007; Carvalho et al., 2011a], as well as, analytical ultracentrifugation data [Carvalho et al, 2011b] which show that the isolated d chain in solution appears as an equilibrium mixture of monomeric and dimeric species, and the contribution of dimers increases at higher protein concentrations.

The fact that the d_x3 trimer does not form spontaneously in solution makes it unlikely that this would be an intermediate during cap formation. This would suggest that the fragile native contacts made between the three *d* subunits around the three-fold axis in the final assembly are only favoured in the presence of the remaining globin chains. Since the cap is a stable structure in the absence of the linkers (Strand *et al.,* 2004) this indicates that the assembly of the cap probably involves the formation of *abcd* tetramers, which subsequently associate. This seems more likely than the formation of an $(abc)_3$ intermediate which would require the subsequent addition of the d-chain trimer. It is known that the globin chains of HbLt have the ability to rebuild the cap after fragmentation in alkaline pH. On the other hand, a hexagonal bilayer of HbGp linkers alone has been observed in electron microscopy experiments (unpublished data), suggesting that this structure could act as a nucleating agent, to which the caps could be added for the formation of the entire complex.

It is of interest to note that the residues which participate in the *a*/*d* interface are largely conserved when comparing the two different isoforms of both the *a* and *d* subunits. However, there is a complementary substitution involving glycines and serines at position 23 of the *a* subunit and position 80 in *d*, suggesting the possibility of a degree of specificity in which *a2* may prefer to pair with *d1* and *a1* with *d2*. At the *d/d* interface of the trimer there is a large accumulation of charged residues and one salt-bridge, between Asp37 of one subunit and Arg34 of its neighbour, appears to be particularly important. Both these residues are substituted for non-charged hydrophilic residues in the *d2* isoform, suggesting that mixture of *d1* and *d2* within a trimer may be disfavoured.

4. Conclusions

Giant haemoglobins are remarkable multimeric complexes which display properties related to fundamental questions in biochemistry such as the basis for specificity at protein-protein interfaces, the stability of macromolecular complexes, cooperativity between subunits and spontaneous assembly. There is still a paucity of structural data for these systems and despite its modest resolution, the structure described here, is still nominally the highest reported to date. We have described the most important subunit-subunit interfaces within the complex, some for the first time. The reason for the existence of isoforms of certain polypeptide chains remains unclear as is the extent to which the assembled particles may be heterogeneous. The difficulty in understanding specificity at interfaces is highlighted by the structure of the isolated *d* chain which uses a promiscuous interface to form a homodimer in place of the trimer observed at the local three-fold axis of the full particle. However, this unexpected result sheds some light on how the complex might be assembled *in vivo*, as it eliminates the possibility of the trimer serving as a nucleus for the formation of the cap. With respect to the stability of the complex, the discovery of an interfacial zinc ion between different protomers seems particularly relevant as it is in agreement with previous reports describing the impact of zinc at low concentrations on the disassembly of the complex. The conservation of the residues which form the binding site suggests this zinc ion to be an important structural component of all such erythrocruorins. On the other hand a previously reported zinc site in HbLt should be treated with caution as its environment seems more compatible with an anion, whose physiological relevance remains to be elucidated. It is to be expected that the inter-subunits contacts, described in detail for HbGp in this paper, will also allow for a better understanding of the main driving forces leading to several experimental observations that, at present, lack an explanation at the molecular level, such as the oligomeric dissociation of the complex under alkaline conditions. Most importantly, there is still very little known about the structural mechanism behind cooperativity and the elevated Hill coefficient which erythrocruorins present. The structure reported here and that of HbLt are good starting points but future efforts must focus on attempts to produce crystals of well-defined R- and T-states of HbGp.

Acknowledgements: We gratefully acknowledge a CNPq fellowship awarded to J.F.R.B. We are also grateful to Sandra Martha Gomes Dias of the National Laboratory of Biosciences (LNBio) for useful discussions concerning nucleic acid sequencing. This work was financed by FAPESP and CNPq.

We acknowledge Dr. Maria Inês Basso Bernardi (IFSC/USP) for the determination of heavy atoms in the sample, measured using Inductively Coupled Plasma Atomic Emission Analysis and Dr. Annie Héroux for her contributions during diffraction data collection at NSLS beamline X25.

References:

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K., & Terwilliger, T. C. (2002). *Acta Crystallogr. D. Biol. Crystallogr.* **58**, 1948–1954.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.

Bachega, J. F. R., Bleicher, L., Horjales, E. R., Santiago, P. S., Garratt, R. C., & Tabak, M. (2011). *J. Synchrotron Radiat.* **18**, 24–28.

Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R., & Leslie, A. G. W. (2011). *Acta Crystallogr. D. Biol. Crystallogr.* **67**, 271–281.

Bosch Cabral, C., Imasato, H., Rosa, J. C., Laure, H. J., da Silva, C. H. T. de P., Tabak, M., Garratt, R. C., & Greene, L. J. (2002). *Biophys. Chem.* **97**, 139–157.

Cardoso, R. M. F., Silva, C. H. T. P., Ulian de Araújo, A. P., Tanaka, T., Tanaka, M., & Garratt, R. C. (2004). *Acta Crystallogr. D. Biol. Crystallogr.* **60**, 1569–1578.

Carvalho, F. A. O., Carvalho, J. W. P., Alves, F. R., & Tabak, M. (2013). *Int. J. Biol. Macromol.* **59**, 333–341.

Carvalho, F. A. O., Santiago, P. S., Borges, J. C., & Tabak, M. (2009). *Anal. Biochem.* **385**, 257–263.

Carvalho, F. A. O., Carvalho, J. W. P.,Santiago, P. S., Tabak, M. (2011a) Proc. Biochem. 46, 2144-2151.

Carvalho, F. A. O., Santiago, P. S., Borges, J. C., Tabak, M. (2011b) Int. J. Biol. Macromol. 48, 183-193.

Condon, P. J. & Royer, W. E. (1994). *J. Biol. Chem.* **269**, 25259–25267.

Elmer, J. & Palmer, A. F. (2012). *J. Funct. Biomater.* **3**, 49–60.

Emsley, P. & Cowtan, K. (2004). *Acta Crystallogr. D. Biol. Crystallogr.* **60**, 2126–2132.

Flores, J. F., Fisher, C. R., Carney, S. L., Green, B. N., Freytag, J. K., Schaeffer, S. W., & Royer, W. E. (2005). *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2713–2718.

Fushitani, K. & Riggs, A. F. (1991). *J. Biol. Chem.* **266**, 10275–10281.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). *Nat. Biotechnol.* **29**, 644–652.

De Haas, F., Biosset, N., Taveau, J. C., Lambert, O., Vinogradov, S. N., & Lamy, J. N. (1996a). *Biophys. J.* **70**, 1973–1984.

De Haas, F., Taveau, J. C., Boisset, N., Lambert, O., Vinogradov, S. N., & Lamy, J. N. (1996b). *J. Mol. Biol.* **255**, 140–153.

De Haas, F., Zal, F., You, V., Lallier, F., Toulmond, A., & Lamy, J. N. (1996c). *J. Mol. Biol.* **264**, 111–120.

Harbury, P. B., Zhang, T., Kim, P. S., & Alber, T. (1993). *Science*. **262**, 1401–1407.

Harnois, T., Rousselot, M., Rogniaux, H., & Zal, F. (2009). *Artif. Cells. Blood Substit. Immobil. Biotechnol.* **37**, 106–116.

Hirsch, R. E., Jelicks, L. A., Wittenberg, B. A., Kaul, D. K., Shear, H. L., & Harrington, J. P. (1997). *Artif. Cells. Blood Substit. Immobil. Biotechnol.* **25**, 429–444.

Kao, W.-Y., Qin, J., Fushitani, K., Smith, S. S., Gorr, T. A., Riggs, C. K., Knapp, J. E., Chait, B. T., & Riggs, A. F. (2006). *Proteins*. **63**, 174–187.

Kapp, O. H., Vinogradov, S. N., Ohtsuki, M., & Crewe, A. V. (1982). *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **704**, 546–548.

Kleywegt G & Jones T.A (2002) Structure *Homus crystallographicus – quo vadis?* 10, 465-472

Krissinel, E. & Henrick, K. (2007) Inference of Macromolecular Assemblies from Crystalline State *J. Mol. Biol*. 372, 774-797

Leslie, A. G. W. (2006). *Acta Crystallogr. D. Biol. Crystallogr.* **62**, 48–57.

Liochev, S. I., Kuchumov, A. R., Vinogradov, S. N., & Fridovich, I. (1996). *Arch. Biochem. Biophys.* **330**, 281–284.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). *J. Appl. Crystallogr.* **40**, 658–674.

Negrisolo, E., Pallavicini, A., Barbato, R., Dewilde, S., Ghiretti-Magaldi, A., Moens, L., & Lanfranchi, G. (2001). The evolution of extracellular haemoglobins of annelids, vestimentiferans, and pogonophorans.

Numoto, N., Nakagawa, T., Kita, A., Sasayama, Y., Fukumori, Y., & Miki, K. (2005). *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14521–14526.

Numoto, N., Nakagawa, T., Kita, A., Sasayama, Y., Fukumori, Y., & Miki, K. (2008). *Biochemistry*. **47**, 11231–11238.

Ochiai, T., Hoshina, S., & Usuki, I. (1993). *Biochim. Biophys. Acta*. **1203**, 310–314.

Oliveira, M. S., Moreira, L.M., Tabak, M. (2007) Int. J. Biol. Macromol. 40, 429-436.

Ownby, D. W., Zhu, H., Schneider, K., Beavis, R. C., Chait, B. T., Riggs, A. F. (1993) J. Biol. Chem. 268, 13539-13547.

Pardanani, A., Gibson, Q. H., Colotti, G., & Royer, W. E. (1997). *J. Biol. Chem.* **272**, 13171–13179.

Perutz, Max, Unravelling the atomic mechanism of haemoglobin, World Scientific Publshing Co., UK, 1997

Richardson, J.S. & Richardson, D.C. (1988) *Science* **240**, 1648-1652

Ronda, L., Bettati, S., Henry, E. R., Kashav, T., Sanders, J. M., Royer, W. E., & Mozzarelli, A. (2013). *Biochemistry*. **52**, 2108–2117.

Rousselot, M., Delpy, E., Drieu La Rochelle, C., Lagente, V., Pirow, R., Rees, J.-F., Hagege, A., Le Guen, D., Hourdez, S., & Zal, F. (2006). *Biotechnol. J.* **1**, 333–345.

Royer, W.E., Strand, K., van Heel, M & Henrickson, W.A. (2000) Structural hierarchy in erythrocruorin *Proc. Natl. Acad. Sci.* 97, 7107-7111

Royer, W. E., Omartian, M. N., & Knapp, J. E. (2007). *J. Mol. Biol.* **365**, 226–236.

Royer, W. E., Pardanani, A., Gibson, Q. H., Peterson, E. S., & Friedman, J. M. (1996). *Proc. Natl. Acad. Sci. U. S. A.* **93**, 14526–14531.

Royer, W. E., Sharma, H., Strand, K., Knapp, J. E., & Bhyravbhatla, B. (2006). *Structure*. **14**, 1167–1177.

Santiago, P. S., Carvalho, F. A. O., Domingues, M. M., Carvalho, J. W. P., Santos, N. C., & Tabak, M. (2010a). *Langmuir*. **26**, 9794–9801.

Santiago, P. S., Carvalho, J. W. P., Domingues, M. M., Santos, N. C., & Tabak, M. (2010b). *Biophys. Chem.* **152**, 128–138.

Standley, P. R., Mainwaring, M. G., Gotoh, T., & Vinogradov, S. N. (1988). *Biochem. J.* **249**, 915–916.

Strand, K., Knapp, J. E., Bhyravbhatla, B., & Royer, W. E. (2004). *J. Mol. Biol.* **344**, 119–134.

Vinogradov, S. N. (1985). *Comp. Biochem. Physiol. B.* **82**, 1–15.

Weber, R. E. & Vinogradov, S. N. (2001). *Physiol. Rev.* **81**, 569–628.

Xie, Q., Donahue, R. A., Schneider, K., Mirza, U. A., Haller, I., Chait, B. T., & Riggs, A. F. (1997). *Biochim. Biophys. Acta*. **1337**, 241–247.

(1994). *Acta Crystallogr. D. Biol. Crystallogr.* **50**, 760–763.

Table 1

	Hb	monomer
Data Collection		
Space Group	I222	I222
Cell dimensions (Å) *a, b, c.*	272.68; 319.90; 333.18	53.11; 63.15; 78.39
Detector	ADSC Quantum 315	Pilatus 2M
X-ray source	NSLS X29c	DLS,I04-1
Wavelength (Å)	1.00	0.9200
Resolution range (Å)	49.65 - 3.20 (3.37-3.20)	43.97–2.05 (2.11-2.05)
Multiplicity	4.7 (4.2)	4.3 (4.3)
*R*merge (%)*	11.2 (50.0)	6.8 (97.7)
Completeness(%)	99.8 (99.2)	99.5(99.3)
Total reflections	1,105,247 (145,116)	36867(2836)
Unique reflections	237,062 (34,189)	8521(657)
I / σ (I)	11.2 (3.1)	19.5(2.5)
Refinement parameters		
Reflections used for refinement	236866	8511 (823 free)
R (%)**	21.65	20.96
R_{Free}(%)**	23.52	23.39
No. of protein atoms	57081	1182
No. of ligand atoms	2082	54
B-factor in Wilson Plot (Å^2)	60.4	27.60
Cruickshank DPI (Å)	0.21	0.28
Coordinates error (ML)	0.41	0.25
Phase error	24.12	25.21
Ramachandran Plot		
Favored (%)	95.58	97.83
Allowed (%)	3.40	2.17
Outliers (%)	1.01	0.0
All-atom Clashscore	12.78	10.82
RMSD from ideal geometry		
r.m.s. bond lengths (Å)	0.012	0.005
r.m.s. bond angles (°)	0.962	0.967
PDB ID	4U8U	4WCH

Figure Legends

Figure 1. Sequence alignment of the HbGp globin chains. The six different globin sequences from *Glossoscolex paulistus* identified in this study (a1, a2, b, c, d1 and d2) have been aligned with their homologues from *Lumbricus terrestris* (HbLt), *Eisennia andrei* (HbEa) and *Perionyx excavatus* (HbPe) according to the four different groups. The *G.paulistus* haemoglobin shows the presence of two isoforms for both the *a* and *d* chains. Standard nomenclature is used to identify the α-helices (A to H) as cylinders. Functionally important residues are boxed and their significance indicated on the figure. The background colours code for well-conserved residues: basic (red), acidic (pink), hydrophilic (green), glycine (orange), proline (yellow) and hydrophobic (blue).

Figure 2. Sequence alignment of the HbGp linker chains. The five linker sequences identified in this study are shown aligned with homologues from *Lumbricus terrestris* (HbLt) and Ea (HbEa) as well as the marine species *Alvinella pompejana* (HbAp) and *Arenicola marina* (HbAm). Due to the presence of isoforms two distinct sequences were identified for the first and third linker. Those which were most similar to the HbLt sequences have been named L1 and L3 and the remaining isoforms L1b and L3b respectively. Secondary structure elements are indicated as cylinders and arrows. The structure is divided into four domains, Domains 1 and 2 correspond to individual helices and form the coiled-coil stalk of the linker trimer. Domains 3 and 4 are respectively the LDL-A domain and the β-barrel domain. The phase of the coiled-coil is indicated by the letters above the sequence and the hydrophobic residues at positions *a* and *d* are boxed, as are residues of the LDL-A domain which are involved in Ca^{2+} binding. The background colour code is as for Fig. 1.

Figure 3. Selected regions of electron density. (a) the Ca^{2+} binding site within the LDL-A domain is observed in all linker chains, exemplified here by L3, (b) the L3 Zn^{2+} binding site, (c) a second Zn^{2+} site at the interface between L3 and L2 and (d) the glycosylation site on Asn121 of L3.

Figure 4. Organization of the 180 polypeptide chains within the particle. To the left of the figure the full particle is shown viewed along both the six-fold axis and one of the two-fold axes. A single hexagonal layer (centre left) can be dissected into six protomers composed of a dodecameric globin cap together with a hetero-trimeric linker complex. Each globin presents the classical fold based on seven α-helices and each linker is composed of a globular head region together with a tail which points towards the six-fold axis. The three tails form a coiled-coil important for particle stability. The figure defines a colour code for the different chains which will be used throughout this paper (top and bottom right).

Figure 5. The dodecameric cap. The upper part of the figure shows the similarity between the cap observed in the land-living species (HbGp and HbLt) and half of the hollow spherical structure seen in *Oligobrachia mashikoi* ($\mathrm{Hb_LOm}$) and *Riftia pachyptila* ($\mathrm{Hb_LRp}$). In the latter a second dodecameric cap completes the 24-mer, whilst in HbGp and HbLt the trimeric complex of linkers interacts in an approximately analogous fashion. In both cases a hollow cavity is generated.

Figure 6. Details of the coiled-coil. The coiled-coil of the hetero-trimer of linkers (left) is shown in detail (centre). The *a* and *d* positions of the heptad repeats are shown explicitly. Details of the d_1 and d_4 layers, which include the presence of hydrophilic residues, are shown explicitly on the right.

Figure 7. Protomer-protomer interactions within a hexagonal layer. Interactions between the β-barrel domains of linker chains from adjacent protomers contribute to the stability of the hexagonal disc. Hydrogen bonding between anti-parallel β-strands coming from L2 of one protomer and L3 of another together with a previously undescribed metal binding site (here interpreted as Zn^{2+}) contribute to the interface.

Figure 8. Interactions made between hexagonal layers (a) and (b) show overviews of the full particle with part of the molecular surface removed. (c) interactions between layers are made principally by the β-barrel domains of L1 and via the coiled-coils. (d) details of the interactions

between L2/L3 coiled coils made between "superimposed" protomers and (e) those made by L1 coiled-coils between non-superimposed protomers.

Figure 9. Structure of the isolated *d* chain. The cap, as observed in the full particle is shown bottom left with the isolated trimer of *d* chains above it. Bottom right shown the interface between the *d* and *a* chains observed in the cap and above it (top right) the analogous interactions formed by a *d/d* interface observed in the structure of the isolated *d* chain.

Figure S1 Resolving sequence ambiguity. Omit maps are shown for a series of positions where there is residue variation between the *a1* and *a2* isoforms. In all cases the maps can be seen to be more consistent with the *a2* sequence. Grey density corresponds to the 2Fo-Fc map and the green peaks are positive peaks in the difference map.

Figure S2 Organization of the dodecameric globin cap *a* and *d* chains and *b* and *c* chains associate via similar interfaces. These dimers pair to form tetramers, three of which associate to form the dodecamer. The simplified representation in the form of spheres follows the same colour code as the ribbons diagram. Bottom left is shown the electron density for an inter-tetrameric disulphide bridge (boxed).

Figure S3 Ribbons representation of the linker chains. The structures from HbGp are shown coloured and superimposed upon their homologs from HbLt (in grey). Overall the folds are very

similar with minimal deviations of the long α-helix (domain 1). L3 in HbGp has additional zinc and glycosylation sites not reported previously.

Figure S4 The cavity within the protomer. Defects at inter-domain and inter-subunit interfaces lead to the existence of 5 different types of channel giving access to the central cavity which is lined by the concave face of the cap and the heads of the linkers. The five different types of channel are shown surrounding a protomer (centre) from which part of the structure has been removed to reveal the cavity. Colour coding for the seven different chains is as for other figures.

Tables

Table 1 Details of the crystallographic refinement. Standard crystallographic parameters are given for both the full particle and the isolated *d* chain.

Table S1. Comparison of predicted physicochemical properties for the 11 HbGp sequences identified in this study. For both molecular mass and pI, where appropriate, the values for *Lumbricus terrestris* haemoglobin are given in parentheses. In terms of predicted pI, the isoforms *d2* and L1b show the most divergent values for the groups to which they belong. The signal peptide sequences were predicted using the SignalP server and show an elevated content of hydrophobic residues as anticipated. The "." indicates the predicted cleavage site.

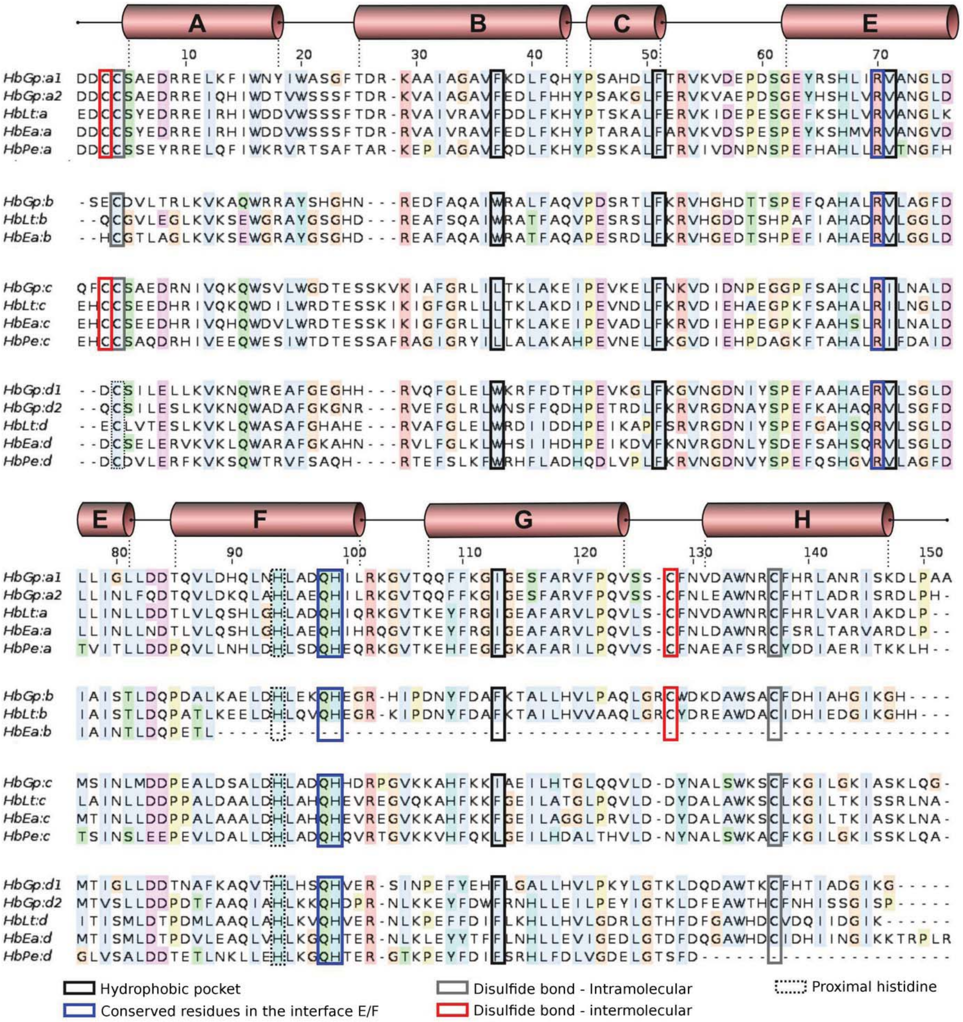

A
B
C
E
10
20
30
40
50
60
70
HbGp:a1 DDCCSAEDRRELKFIWNYIWASGFTDR-KAAIAGAVFKDLFQHYPSAHDLFTRVKVDEPDSGEYRSHLIRVANGLD
HbGp:a2 DDCCSAEDRREIQHIWDTVWSSSFTDR-KVAIAGAVFEDLFHHYPSAKGLFERVKVAEPDSGEYHSHLVRVANGLD
HbLt:a EDCCSYEDRREIRHIWDDVWSSSFTDR-RVAIVRAVFDDLFKHYPTSKALFERVKIDEPESGEFKSHLVRVANGLK
HbEa:a DDCCSYEDRREIRHIWDDVWSSSFTDR-RVAIVRAVFEDLFKHYPTARALFARVKVDSPESPEYKSHMVRVANGVD
HbPe:a DDCCSSEYRRELQFIWKRVRTSAFTAR-KEPIAGAVFQDLFKHYPSSKALFTRVIVDNPNSPEFHAHLLRVTNGFH
HbGp:b -SECDVLTRLKVKAQWRRAYSHGHN---REDFAQAIWRALFAQVPDSRTLFKRVHGHDTTSPEFQAHALRVLAGFD
HbLt:b --QCGVLEGLKVKSEWGRAYGSGHD---REAFSQAIWRATFAQVPESRSLFKRVHGDDTSHPAFIAHADRVLGGLD
HbEa:b --HCGTLAGLKVKSEWGRAYGSGHD---REAFAQAIWRATFAQAPESRDLFKRVHGEDTSHPEFIAHAERVLGGLD
HbGp:c QFCCSAEDRNIVQKQWSVLWGDTESSKVKIAFGRLILTKLAKEIPEVKELFNKVDIDNPEGGPFSAHCLRILNALD
HbLt:c EHCCSEEDHRIVQKQWDILWRDTESSKIKIGFGRLLLTKLAKDIPEVNDLFKRVDIEHAEGPKFSAHALRILNGLD
HbEa:c EHCCSEEDHRIVQHQWDVLWRDTESSKIKIGFGRLLLTKLAKEIPEVADLFKRVDIEHPEGPKFAAHSLRILNALD
HbPe:c EHCCSAQDRHIVEEQWESIWTDTESSAFRAGIGRYILLALAKAHPEVNELFKGVDIEHPDAGKFTAHALRIFDAID
HbGp:d1 --DCSILELLKVKNQWREAFGEGHH---RVQFGLELWKRFFDTHPEVKGLFKGVNGDNIYSPEFAAHAERVLSGLD
HbGp:d2 --QCSILESLKVKNQWADAFGKGNR---RVEFGLRLWNSFFQDHPETRDLFKRVRGDNAYSPEFKAHAQRVLSGFD
HbLt:d --ECLVTESLKVKLQWASAFGHAHE---RVAFGLELWRDIIDDHPEIKAPFSRVRGDNIYSPEFGAHSQRVLSGLD
HbEa:d --DCSELERVKVKLQWARAFGKAHN---RVLFGLKLWHSIIHDHPEIKDVFKNVRGDNLYSDEFSAHSQRVLSGLD
HbPe:d --DCDVLERFKVKSQWTRVFSAQH----RTEFSLKFWRHFLADHQDLVPLFKRVNGDNVYSPEFQSHGVRVLAGFD
E
F
G
H
80
90
100
110
120
130
140
150
HbGp:a1 LLIGLLDDTQVLDHQLNHLADQHILRKGVTQQFFKGIGESFARVFPQVSS-CFNVDAWNRCFHRLANRISKDLPAA
HbGp:a2 LLINLFQDTQVLDKQLAHLAEQHILRKGVTQQFFKGIGESFARVFPQVSS-CFNLEAWNRCFHTLADRISRDLPH-
HbLt:a LLINLLDDTLVLQSHLGHLADQHIQRKGVTKEYFRGIGEAFARVLPQVLS-CFNVDAWNRCFHRLVARIAKDLP--
HbEa:a LLINLLNDTLVLQSHLGHLAEQHIHRQGVTKEYFRGIGEAFARVLPQVLS-CFNLDAWNRCFSRLTARVARDLP--
HbPe:a TVITLLDDPQVLLNHLDHLSDQHEQRKGVTKEHFEGFGKAFARILPQVVS-CFNAEAFSRCYDDIAERITKKLH--
HbGp:b IAISTLDQPDALKAELDHLEKQHEGR-HIPDNYFDAFKTALLHVLPAQLGRCWDKDAWSACFDHIAHGIKGH----
HbLt:b IAISTLDQPATLKEELDHLQVQHEGR-KIPDNYFDAFKTAILHVVAAQLGRCYDREAWDACIDHIEDGIKGHH---
HbEa:b IAINTLDQPETL--
HbGp:c MSINLMDDPEALDSALDHLADQHDRPGVKKAHFKKIAEILHTGLQQVLD-DYNALSWKSCFKGILGKIASKLQG-
HbLt:c LAINLLDDPPALDAALDHLAHQHEVREGVQKAHFKKFGEILATGLPQVLD-DYDALAWKSCLKGILTKISSRLNA-
HbEa:c MTINLLDDPPALAAALDHLAHQHEVREGVKKAHFKKFGEILAGGLPRVLD-DYDALAWKSCLKGILTKIASKLNA-
HbPe:c TSINSLEEPEVLDALLDHLADQHQVRTGVKKVHFQKLGEILHDALTHVLD-NYNALSWKACFKGILGKISSKLQA-
HbGp:d1 MTIGLLDDTNAFKAQVTHLHSQHVER-SINPEFYEHFLGALLHVLPKYLGTKLDQDAWTKCFHTIADGIKG-----
HbGp:d2 MTVSLLDDPDTFAAQIAHLKKQHDPR-NLKKEYFDWFRNHLLEILPEYIGTKLDFEAWTHCFNHISSGISP-----
HbLt:d ITISMLDTPDMLAAQLAHLKVQHVER-NLKPEFFDIFLKHLLHVLGDRLGTHFDFGAWHDCVDQIIDGIK------
HbEa:d MTISMLDTPDVLEAQLVHLKGQHTER-NLKLEYYTFFLNHLLEVIGEDLGTDFDQGAWHDCIDHIINGIKKTRPLR
HbPe:d GLVSALDDTETLNKLLEHLKGQHTER-GTKPEYFDIFSRHLFDLVGDELGTSFD----------------------
Hydrophobic pocket
Conserved residues in the interface E/F
Disulfide bond - Intramolecular
Disulfide bond - intermolecular
Proximal histidine

Figure 1

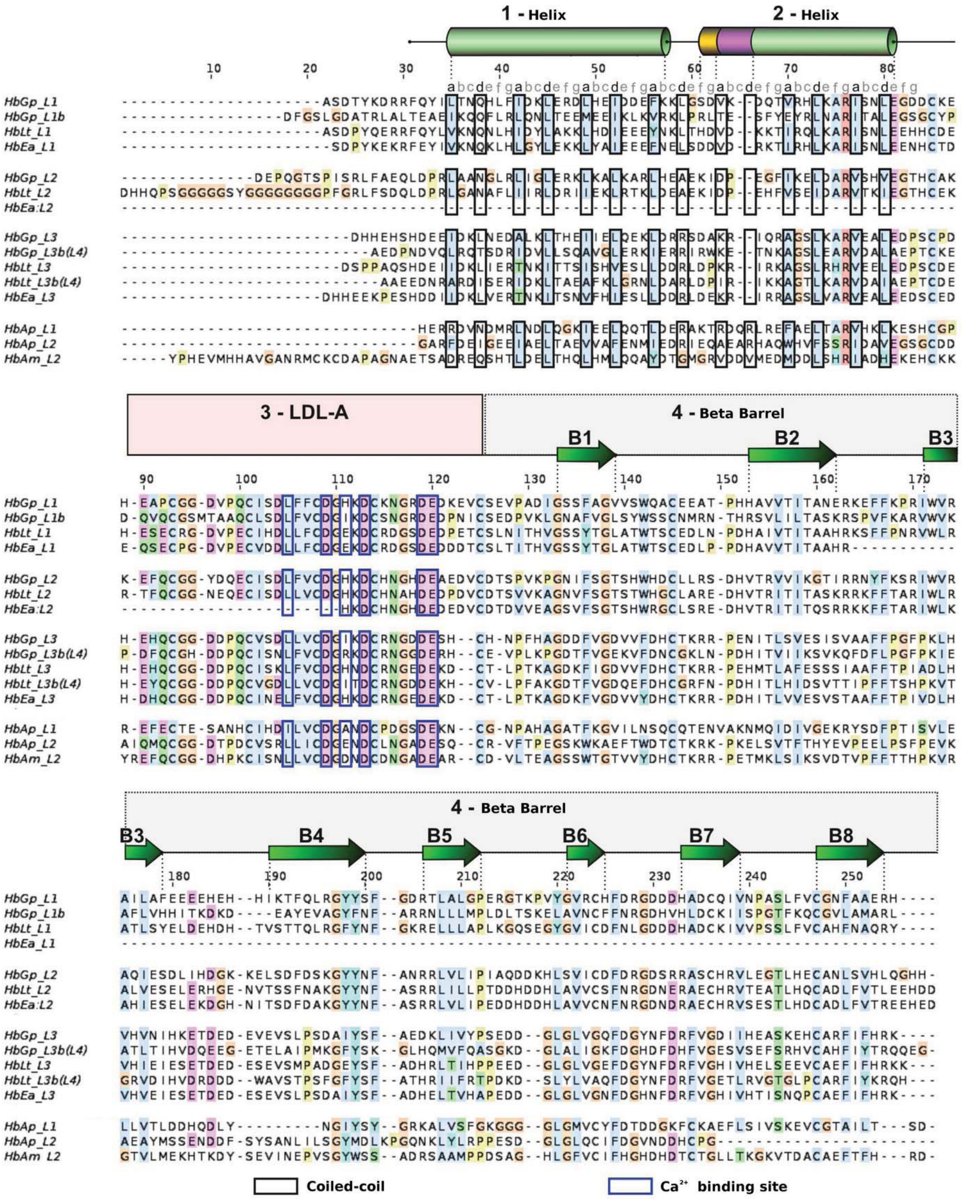

1 - Helix
2 - Helix
3 - LDL-A
4 - Beta Barrel
B1
B2
B3
B4
B5
B6
B7
B8
HbGp_L1
HbGp_L1b
HbLt_L1
HbEa_L1
HbGp_L2
HbLt_L2
HbEa:L2
HbGp_L3
HbGp_L3b(L4)
HbLt_L3
HbLt_L3b(L4)
HbEa_L3
HbAp_L1
HbAp_L2
HbAm_L2
Coiled-coil
Ca2+ binding site

Figure 2

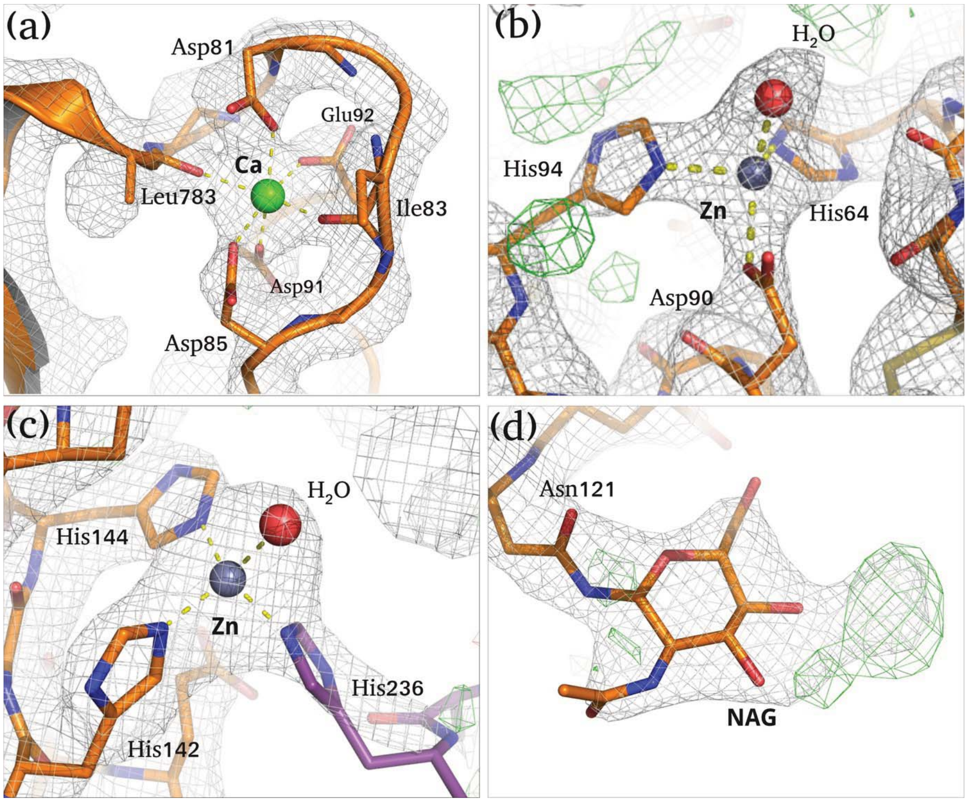

(a)
Asp81
Glu92
Ca
Leu783
Ile83
Asp91
Asp85
(b)
H_2O
His94
Zn
His64
Asp90
(c)
H_2O
His144
Zn
His236
His142
(d)
Asn121
NAG

Figure 3

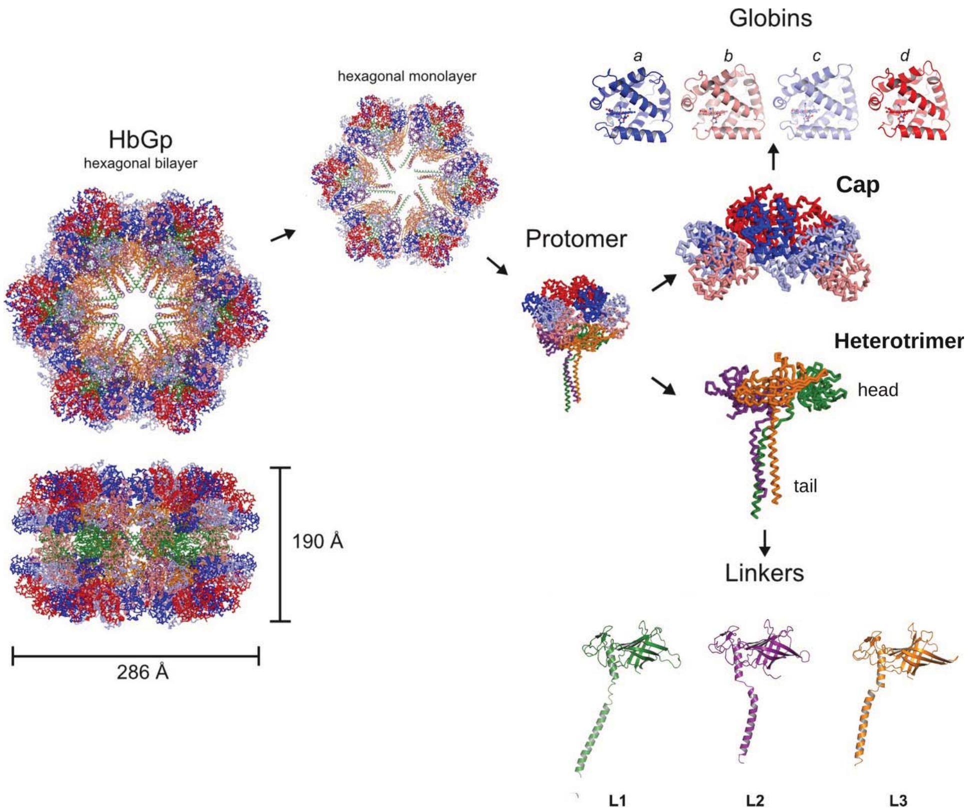

Globins
a
b
c
d
hexagonal monolayer
HbGp
hexagonal bilayer
Cap
Protomer
Heterotrimer
head
tail
190 Å
286 Å
Linkers
L1
L2
L3

Figure 4

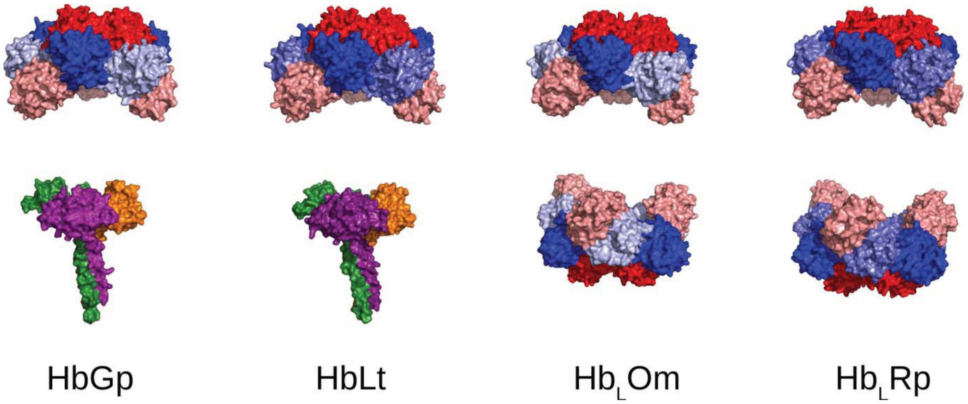


Figure 5

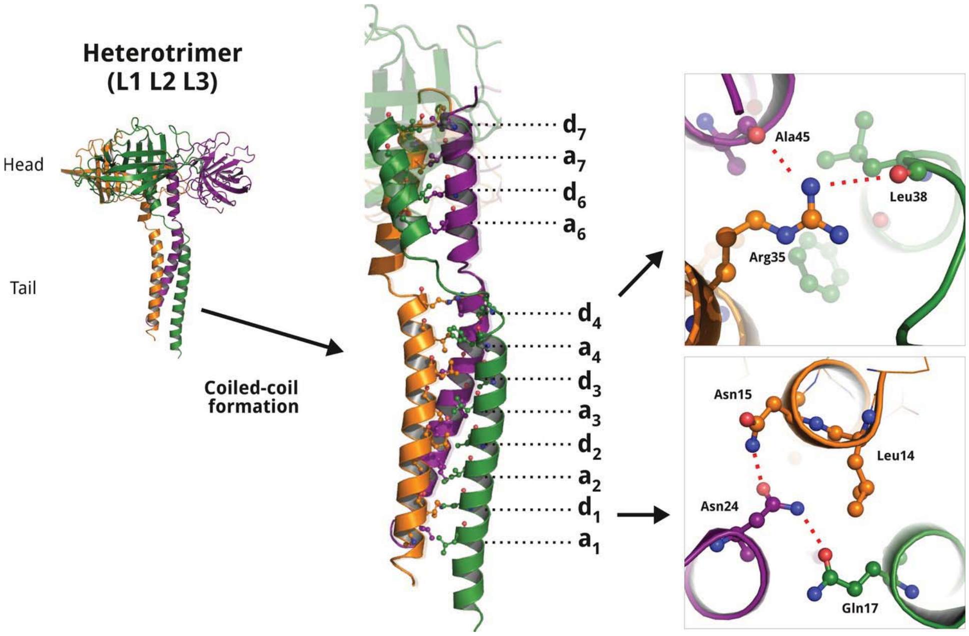

Heterotrimer
(L1 L2 L3)
Head
Tail
Coiled-coil
formation
d7
a7
d6
a6
d4
a4
d3
a3
d2
a2
d1
a1
Ala45
Leu38
Arg35
Asn15
Leu14
Asn24
Gln17

Figure 6

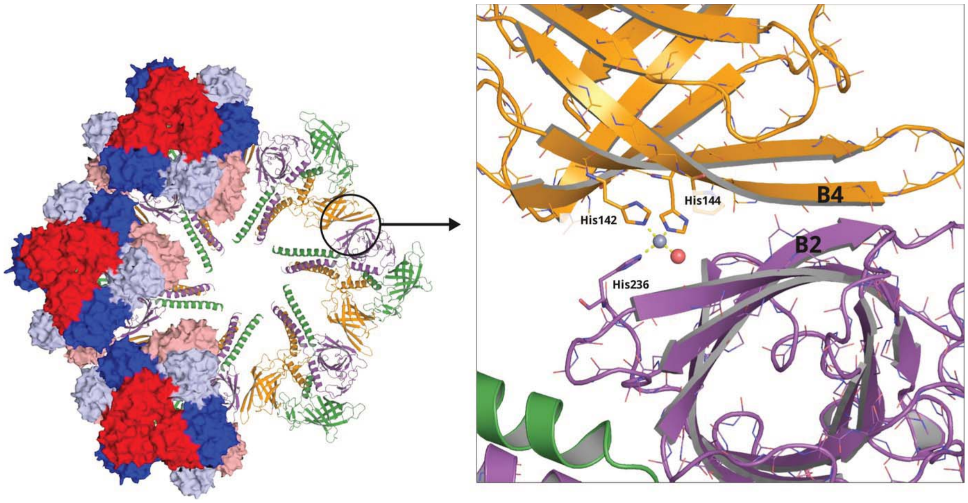


Figure 7

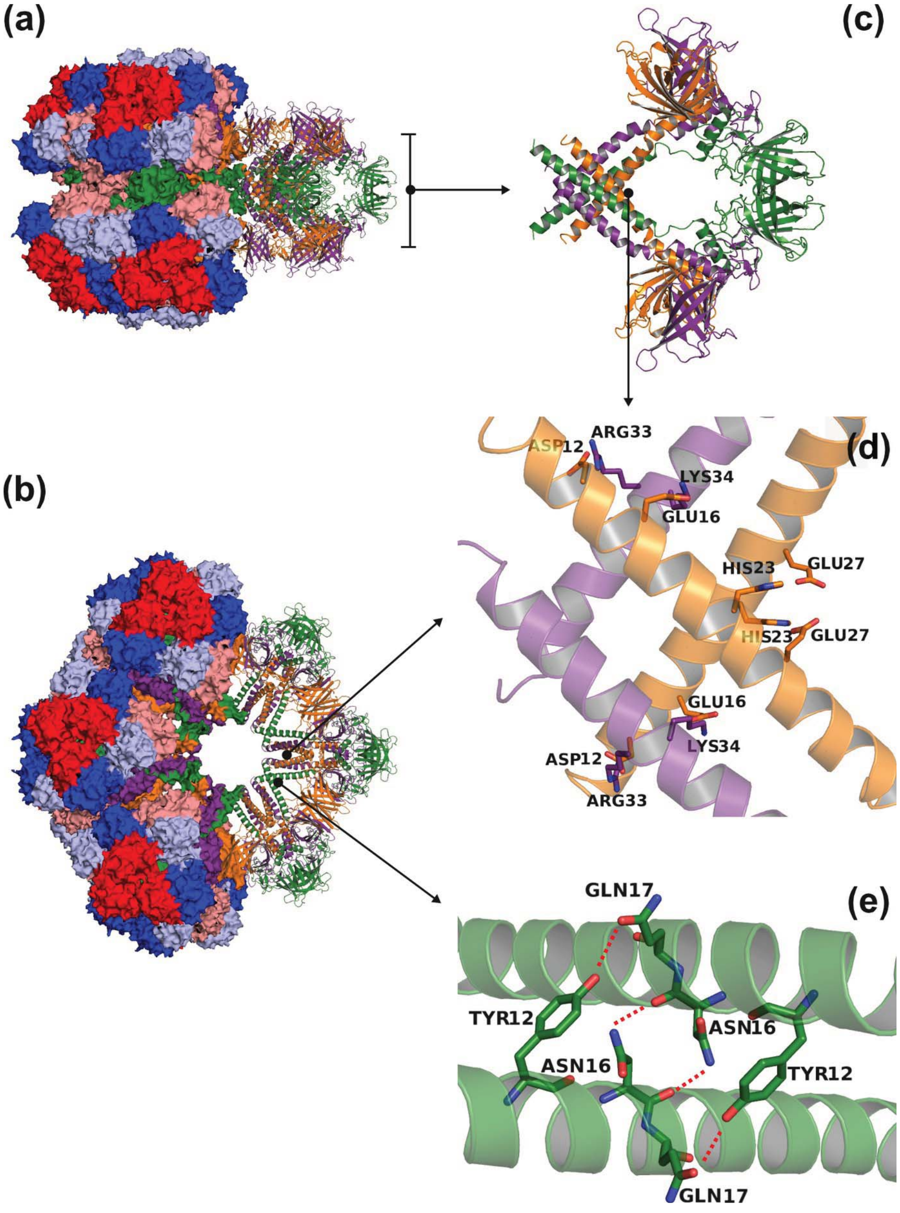
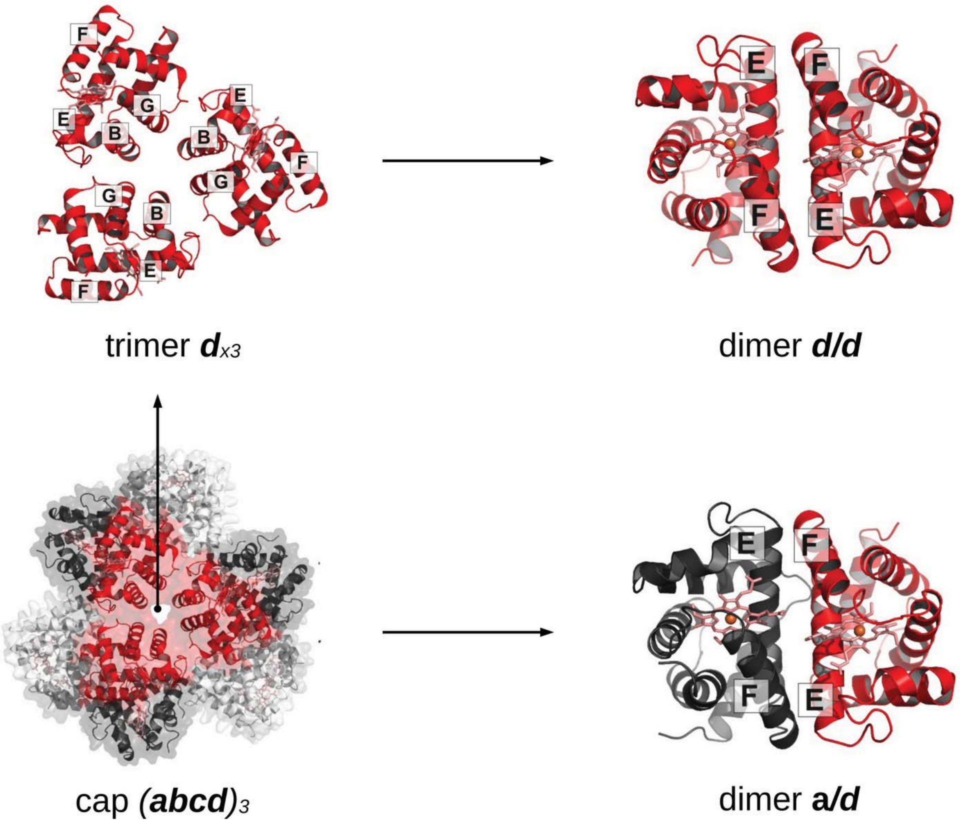

(a)
(c)
(b)
(d)
ASP12
ARG33
LYS34
GLU16
HIS23
GLU27
HIS23
GLU27
GLU16
LYS34
ASP12
ARG33
(e)
GLN17
TYR12
ASN16
ASN16
TYR12
GLN17

Figure 8



trimer ***d****x3*

dimer ***d/d***

cap *(**abcd**)3*

dimer **a/*d***

Figure 9