Bioinformatical Development of Oligonucleotides for the *d* Chain Gene of the Giant Extracellular Hemoglobin of *Glossoscolex Paulistus*

Érico M. Polo, Rosiane P. Santos, Leonardo M. Moreira, and Gabriel M. Yazbeck

***Abstract*—Aiming the isolation and molecular characterization of the gene for the *d* polypeptide chain of giant extracellular hemoglobin of *Glossoscolex paulistus* (HbGp), we present the design of oligonucleotides to DNA hybridization procedures. From a described amino acid sequence, we performed reverse translation, with an algorithm, which searched for regions with maximum density of non-degenerate codons, focusing minimization of the number of putative oligonucleotide probes. Fourteen different partial oligopeptide sequences six, seven or eight residues long resulted in 64, 256 and 512 possible coding sequences, respectively. Based on the similarity with *d* chain of the hemoglobin from *Lumbricus terrestris* (HbLt), we ranked the resulting sequences. This narrowed the candidate sequences to 2, 20 and 24 possible probes 18, 21 and 24 nucleotides long, respectively, from selecting the highest scoring sequences. We filtered the set of putative probes based on the relative frequency of codon usage from *L. terrestris*. Sequences with the higher scores were chosen, favoring shorter sequences. Thus, five sequences targeting two different regions and possible primer pairs for PCR with the HbGp *d* chain gene were obtained. These oligonucleotides can be employed in hybridization and PCR essays in *G. paulistus* genomic DNA.**

***Index Terms*—Extracellular giant hemoglobin, *Glossoscolex paulitus*, HbGp, *Lumbricus terrestris*, reverse translation.**

I. Introduction

The giant extracellular hemoglobins constitute the summit of structural complexity in hemoproteins that carries molecular oxygen [1]-[3]. These supramolecular systems present a very interesting quaternary structure, involving great number of polypeptide subunits. This great number of polypeptidic chains favor the hydrophobic isolation of the prosthetic groups, *i. e*., the heme groups, which constitute the active site of these hemoproteins, being direct responsible to the bound and carry the oxygen molecule. This isolation is a fundamental prerequiste to avoid significant water accessibility to the heme pocket, which is directly associated to the oxidation of the ferrous ion to ferric ion, which is not a functional coordination center to bound the oxygen ligand [4]-[6].

Great number of important works has been published focused on the giant extracellular hemoglobin of *Lumbricus terrestris* (HbLt), which has been considered relevant reference in this area of research [7]-[9]. We here focus on the giant extracellular hemoglobin of *Glossoscolex paulistus* (HbGp), which presents significant similarities regarding redox, structural and physico-chemical properties in relation to HbLt. *Glossoscolex paulistus* is an annelid found in a considerable quantity in the state of São Paulo, Brazil, mainly in the regions of the municipalities of Rio Claro (SP) and Piracicaba (SP) [10].

Giant extracellular hemoglobin from *Glossoscolex paulistus* (HbGp) is a multi-subunit protein relevant in diverse areas, including biochemistry, biotechnology, medicine and evolution. This giant extracellular hemoglobin presents high resistance to autoxidation, thus being prototype to biosensors and blood substitutes. This hemoglobin constitutes a quaternary arrangement with 180 polypeptide subunits, being 144 polypeptide chains with heme group (globin chains) and 36 polypeptide chains without heme group, which are called "Linker chains", similarly to the HbLt [11]. These "Linker chains" present only structural function, since they cannot carry oxygen molecules because does not have the ferrous ion-porphyrin system to bind this ligand. The "Linker chains" are disposed more predominantly in the central part of the hexagonal quaternary arrangement, being associated, at least partially, to the cohesion of the supramolecular structure.

The spatial arrangement of HbGp comprises two mutually superposed hexagons, sometimes refered to as a "bracelet", due its complex stereochemical disposition, generating hexagonal bilayer that involves a total supramolecular mass of 3.6 MDa [11]. Each vertex of each hexagon has three *abcd* tetramers (constituting a dodecamer each). Considering that the two hexagonal arrangements presents a total value of 12 dodecamers, there is a total of 144 globin chains, since the polypeptide chains *a*, *b*, *c*, and *d* are different types of globin polypeptides [12].

In this context, it is important to notice that the chains *a*, *b* and *c* constitute a trimer (trimer *abc*), through disulfide bounds and the monomer *d* is maintained in the tetramer with only weak interaction, i.e., without covalent bonds, like, for example, disulfide bounds. In this way, the study of the monomer *d* has been considered an important resource of information on the tertiary structure, since the monomer *d* is the unique polypeptide subunit that can be easily isolated from the whole quaternary structure [13].

Manuscript received August 15, 2013; revised November 13, 2013. This work was supported in part by Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) and Companhia Elétrica de Minas Gerais (CEMIG).

Érico Macedo Polo is with the Genetics, Conservation and Evolutionary Biology Program (PPG-GCBEv), Instituto Nacional de Pesquisas da Amazônia (INPA), located at Av. André Araújo, 2936 – Aleixo - Manaus/AM, CEP 69011-970 (e-mail: erico.polo@gmail.com).

Rosiane de Paula Santos, Leonardo Marmo Moreira, and Gabriel Menezes Yazbeck are with the Laboratório de Recursos Genéticos, DEZOO, Universidade Federal de São João Del-Rei, São João Del Rei, MG, CEP 36307-352, Brazil (e-mail: rosianeps2007@yahoo.com.br, leonardomarmo @gmail.com, dna@ufmg.br).

This kind of hemoglobin present high resistance to the autoxidation, which is deeply associated to its complex spatial arrangement [14]. Indeed, these hemoglobins are disposed directly in the blood, without any erythrocytes, differently of the mammalian hemoglobins, and even in these drastic chemical environment conditions present great redox stability. This can be easily observed through of its redox and structural stabilities even when submitted to storage procedures without control of the oxygen presence, i.e., without conditions of inert atmosphere. [15]. Furthermore, these hemoglobins are considered hemoprotein systems with great cooperativity, and in some cases with "supercooperativity" [16], [17].

The affinity of the ferrous ion for molecules of great biological and environmental relevance makes this giant extracellular hemoglobin an interesting model as a "biosensor". Indeed, compounds as nitric oxide (NO), carbon monoxide (CO), cyanide anion (CN-), among others, are examples of ligands that present high chemical stability when coordinated to the ferrous ion [4]. This happens because these compounds are highly π-acid ligands (π-acceptor ligands), which present significant affinity to the ferrous ion (Fe^{II}), which, in turn, present d6 low-spin electronic configuration. Therefore, in the d orbitals with pi symmetry, this cationic center present high electronic density, being more effectively stabilized when bounded with pi-receptor ligands. In this way, this hemoglobin could be the main component of new materials elaborated in order to detect important molecules in several biological and non-biological environments.

Motivated by the lack of availability of the gene sequence for the HbGp *d* chain, we herein approached a reverse translation procedure of a 142 residues long protein sequence in order to generate a set of putative oligonucleotide probes and PCR primers, suitable for molecular essays with genomic DNA from *G. paulistus.*

II. Material and Methods

We developed an original algorithm (written in Perl language, available upon request from author EMP) that used the HbGp *d* chain aminoacid sequence as input. It attributes to each position a value which is given by the number of different alternative codons for that particular aminoacid, using the standard universal genetic code [18]. Then, given a fixed arbitrary size class of aminoacid residues (six, seven or eight), the program would return the regions with the maximum density of non-degenerate or twofold degenerate codons. Through the use of the original algorithm written here, we obtained all possible 18 to 24 bp DNA sequences. These were recorded in FASTA format and used to perform a nucleotide BLAST search [19] in GenBank. Then, the higher scoring sequences were compared and rescored in relation to each other, regarding relative condon usage frequency, estimated from the HbLt *d* chain data. Sequences with the five highest scores from each size class were selected. This procedure favored the maintenance of shorter sequences, every time two sequences overlapped or embedded one another. Oligonucleotide probes and PCR primers were designed from the resulting sequences. Oligonucleotide properties were theoretically derived with the aid of the Oligonucleotide Properties Calculator [20].

III. Results

A total of 14 different oligopeptide stretches six, seven or eight residues long (three, seven and four, respectively) were selected, based on their degeneracy density, resulting in the back translation of 64, 256 and 512 possible underlying coding nucleotide sequences, respectively, binding to three different regions of the analyzed protein (Fig. 1).

DDCSILELLKVKNQWREAFGEGHHRVQFGLELWKR
FFDTHPEVKGLFKGVNGDNIYSPEFAAHAERVLSGL
DMTIGLLDDTNAFKAQVTHLHSQHVERSINPEFYEHF
LGALLHVLPKYLGTKLDQDAWTKCFHTIADGIKG

Fig. 1. The HbGp *d* chain featuring the three selected regions for the screening of possible candidate oligonucleotide probes. Regions may encompass alternative six, seven or eight aminoacids overlapping regions.

GenBank BLAST searches with the 832 candidate probes as queries returned the *L. terrestris* hemoglobin *d* chain gene [GenBank: U55073.1] described by [21]. After ranking the obtained sequences we narrowed down the pool of possible probes to two, 20 and 24 putative sequences 18, 21 and 24 bases long, respectively, chosen only from the highest scoring matches.

These sequences were then sorted with the aid of the relative codon usage bias, estimated from the HbLt *d* chain gene sequence. The first five ranking sequences from each size class group were arbitrarily selected, resulting in a total of nine remaining sequences. We favoured the smaller sequences whenever two sequences had overlapping or embedding sequences, thus narrowing the search down to five oligonucleotide sequence candidates. These five sequences, due to simple substitutions of the transition type can be summarized as three "elected" probes shown in Table I. The five probes pool consisted of two 18 bases sequences that targeted the most downstream region in Fig. 1. The remaining three 21 bases sequences are candidates to the intermediary region in Fig. 1 (where *R=A* or *G* and *Y=C* or *T*). The candidate oligonucleotide probes had individual estimated melting temperature (T_m) ranging from 43,5 to 48,5 ℃ by the most direct simple calculation (T_m=64.9+41(*G*+*C*-16.4)/*L*, where *G*+*C* is the Citosine and Guanine composition of a single strand, and L is its length). Also, the temperatures for a 50 mM salt concentration are given in Table I.

A PCR primer cocktail is proposed comprising the two reverse complement sequences of the shorter 18 bases oligonucleotides as reverse primers, along the direct 5'-3' sequence from the three 21 bases oligonucleotides in Table I as forward primers. These five oligonucleotides allow for six possible primer pairs to be individually screened in *in vitro* essays. Alternatively two primers pair combinations with a single degenerate base are possible (Table I). A virtual PCR simulation with the HbLt cDNA sequence allowing broad primer mismatch (50%) indicates the possible formation of a PCR product with estimated size of around 62 bp, and judging from the 142 aminoacid sequence in Fig 1. the expected amplicons from a cDNA would be around 102 and

99 bp long for the two possible pair.

TABLE I: OLIGONUCLEOTIDE PROBES DETERMINED BY THE PROCESS OF REVERSE TRANSLATION FOLLOWED BY INFORMED SELECTION

Oligonucleotide 5'-3'sequence		Basic T_m (℃)	Salt adjusted T_m (℃)
Elected probes	TGGACCAAATGCTTTCAY	43.5	49.3
	AACCCTGAATTCTACGARCAT	48.5	55.4
	CCTGAATTCTACGAACATTTC	48.5	55.4
PCR primers	Forward: AACCCTGAATTCTACGARCAT *or* CCTGAATTCTACGAACATTTC		
	Reverse: RTGAAAGCATTTGGTCCA		

IV. DISCUSSION

The oligonucleotide probe pool herein proposed is the first set of putative candidates capable of being used in molecular essays involving the use of *G. paulistus* genomic DNA for the screening of the *d* chain gene *in vitro*. The set was strictly developed using an *in silico* approach and the steps used here to narrow down the size of the universe of possibilities could be generalized to be applied to other molecular systems. Even though none of the proposed possibilities might be the real one, this procedure allows for an "educated guess" about the putative true underlying coding sequence in *G. paulistus*. The accuracy of this technique could be accessed in the future by means of computer simulations and application to data sets of known protein to nucleic acid correspondence.

Also, the empirical use of the proposed probes could be used to validate the results found here. The application of the designed probes, under a set of varying hybridization stringency conditions (i.e. different salt concentrations and annealing temperatures) will allow for the further refinement and scrutinizing among the pool. The two 18 bases long probes target anchor in the last selected region and the three larger 21 bases probes will target the middle region. This allowed for the proposal of PCR primers to be designed from the oligonucleotide pool. An *in silico* PCR simulation with one herein proposed primers pair and the the HbLt, allowing for 50% of degeneracy predicted the formation PCR product, but caution must be taken since the *L. terrestris* sequence used here for comparison [22] was derived from a cDNA originated from reverse transcription from the *d chain* mRNA, which lacks intron sequences. Genomic essays using this primers will most likely encompass a larger amplicon, due to the presence of possible introns, usually two, interrupting three exon regions in *L. terrestris* globins [22].

The isolation of this particular globin chain from *G. paulistus* is of great relevance for biotechnology and can provide an *in vivo* recombinant system for bacterial expression of the *d* chain peptide in large scales, thus facilitating the access to this material that can be further explored to its potential application as a biosensor.

Other very auspicious applications of giant extracellular hemoglobins is focused on the elaboration of "blood substitute", which are also denominated "artificial blood" [23] and the relation between the presence of hemoproteins and the generation and/or control of free radical generation. In fact, the problem of blood transfusion constitutes a great academic challenge in all world, since high number of persons dye all years due to the absence of enough quantity of suitable blood to a rapid transfusion. Regarding the studies focused free radical generation, this approach has received great attention of relevant research groups due to the association between the hemoproteins and the free radicals generations, which are responsible for several biological processes, including physiological and pathological chemical mechanisms. Furthermore, it is interesting to mention the studies upon hybrid hemoproteins [24], [25] and reconstitution of hemoproteins [26], [27] has furnished a representative advancement in the understanding of the structure-activity relationship of very complex metalloproteins.

Several relevant research areas associated to the possible applications of giant hemoglobins, such as biomedical engineering, materials engineering, supramolecular chemistry, between others, require a more detailed characterization of these proteins, aiming a consistent understanding upon its structure-function relationship. In this way, advancements in the characterization in terms of molecular biology can be a very interesting alternative to obtain more detailed physico-chemical and biochemical information of this complex supramolecular system.

V. CONCLUSION

The present work constitutes a relevant and original contribution to the understanding of the molecular bases of the monomer *d* of HbGp. This approach, involving an emphasis related to genetics and molecular biology, is an important prerequisite to the isolation and full characterization of its coding gene. This will facilitate the development of expression systems capable of yielding large quantities of this polypeptide subunit and thus helping in the elucidation of several physico-chemical properties of the chain *d*. In this way, aspects involving the evolution of the giant extracellular hemoglobins could be more effectively studied. We believe that this present work can be a relevant contribution to obtain a more consistent analysis of spectroscopic data inherent to giant extracellular hemoglobins, which have been obtained from several research groups. These novel findings would be decisive in the development of various potential biotechnological applications of HbGp. New evaluations regarding monomer *d* of HbGp has been made in our group and soon will be published.

REFERENCES

[1] J. N. Lamy, B. Green, A. Toulmond, J. Walls, R. Weber, and S. Vinogradov, "Giant hexagonal bilayer hemoglobins," *Chemical Reviews*, vol. 96, pp. 3113-3124, 1996.

[2] S. N. Vinogradov, "The stoichiometry of the four linker subunits of Lumbricus terretris hemoglobin suggests an aymetric distribution," *Micron*, vol. 35, pp. 213-222, 1996.

[3] A. Krebs, P. Zipper, and S. N. Vinogradov, "Lack of size and shape alteration of oxygenated and deoxygenated *Lumbricus terrestris* hemoglobin?" *Biochimica et Biophysica Acta*, vol. 1297, pp. 115-118, 1996.

[4] A. L. Poli, L. M. Moreira, A. A. Hidalgo, and H. Imasato, "Autoxidation studies of extracellular hemoglobin of *Glossoscolex paulistus* at pH 9: Cyanide and Hydroxyl effect," *Biophysical Chemistry*, vol. 114, pp. 253-260, 2005.

[5] A. L. Poli, L. M. Moreira, M. Tabak, and H. Imasato, "SDS (sodium dodecyl sulfate) effect on the autoxidation of the Glossoscolex paulistus giant extracellular hemoglobin: Kinetic studies at pH 7.0 and 9.0," *Colloids and Surfaces B*, vol. 52, pp. 96-104, 2006.
[6] L. M. Moreira, A. L. Poli *et al*., "Pentacoordinate and hexacoordinate ferric hemes in acid medium: EPR, UV-Vis and CD studies of the giant extracellular hemoglobin of Glossoscolex Paulistus," *Biophysical Chemistry*, vol. 124, pp. 62-72, 2006.
[7] S. C. Dorman, J. P. Harrington, M. S. Martin, and T. V. Johnson, "Determination of the formal reduction potential of *Lumbricus terrestris* hemoglobin using thin layer spectroeletrochemistry," *Journal of Inorganic Biochemistry*, vol. 98, pp. 185-188, 2004.
[8] O. H. Kapp, G. Polidori, M. G. Mainwaring, A. V. Crewe, and S. N. Vinogradov, "The reassociation of *lumbricus terrestris* hemoglobin dissociated at alkaline pH*," *The Journal of Biological Chemistry*, vol. 259, pp. 628-639, 1984.
[9] P. D. Martin, A. R. Kuchumov, and B. R. Green, "Mass spectrometric composition and molecular mass of *lumbricus terrestris* hemoglobin: A refined model of its quaternary structure," *Journal of Molecular Biology*, vol. 225, pp. 154-169, 1996.
[10] G. Righi, "Topography of circulatory system of glossoscolex paulistus michaelsen 1926 (oligochaeta glossoscolecidae)," *Monitore Zoology Italian*, vol. 6, pp. 19-36, 1972.
[11] A. R. Kuchumov, J. C. Taveau, J. N. Lamy, J. S. Wall, R. E. Weber, and S. N. Vinogradov, "The role of Linkers in the reassembly of the 3.6 MDa hexagonal bilayer hemoglobin from *Lumbricus terrestris*," *Journal of Molecular Biology*, vol. 289, pp. 1361-1374, 1998.
[12] W. E. Royer, K. Strand, M. Van Hell, and W. A. Hendrickson, "Structural hierarchy in erythrocruorin, the giant respiratory assemblage of annelids," in *Proc. the National Academy of Sciences of the United States of America*, vol. 97, 2000, pp. 7107-7111.
[13] C. B. Cabral, H. Imasato, J. C. Rosa, H. J. Laure, C. H. T. P. Silva, M. Tabak, R. C. Garrett, and L. J. Greene, "Fluorescence properties of tryptophan residues in the monomeric d-chain of Glossoscolex paulistus hemoglobin: An interpretation based on a comparative molecular model," *Biophysical Chemistry*, vol. 97, pp. 139-157, 2002.
[14] A. L. Poli, L. M. Moreira, and H. Imasato, "Autoxidation of giant extracellular hemoglobin of Glossoscolex paulistus: Molecular mechanism and oligomeric implications," *Spectrochimica Acta A*, vol. 82, pp. 306-315, 2011.
[15] L. M. Moreira, A. L. Poli *et al*., "Ferric species equilibrium of the giant extracellular hemoglobin of Glossoscolex Paulistus in alkaline medium: HALS hemichrome as a percursor of pentacoordinate species," *International Journal of Biological Macromolecules*, vol. 42, pp. 103-110, 2008.
[16] A. F. Riggs, "Self-association, cooperativity and supercooperativity of oxygen binding by hemoglobins," *The Journal of Experimental Biology*, vol. 201, pp. 1073-1084, 1998.
[17] W. E. Royer, J. E. Knapp, K. Strand, and H. A. Heaslet, "Cooperative hemoglobins: Conserved fold, diverse quaternary assemblies and allosteric mechanisms," *Trends in Biochemical Sciences*, vol. 26, pp. 297-304, 2001.
[18] F. H. C. Crick, "The origin of the genetic code," *Journal of Molecular Biology*, vol. 38, pp. 367–379, 1976.
[19] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "1990Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389-402, Sep. 1, 1997.
[20] W. A. Kibbe, "OligoCalc: An online oligonucleotide properties calculator," *Nucleic Acids Res.*, vol. 35, May 25, 2007.
[21] Q. Xie, R. A. Jr. Donahue, K. Scheineder, U. A. Mirza, I. Haller, B. T. Chait, and A. F. Roggs, "Structure of chain d of the gigantic hemoglobin of the earthworm," *Biochimica et Biophysica Acta*, 1997.
[22] S. M. Jhiang, J. R. Garey, and A. F. Riggs, "Exon-intron organiation in genes of earthworm and vertebrate globins," *Science*, vol. 240, 1988.
[23] R. E. Hirsh, L. A. Jelicks, B. Wittenberg, D. K. Kaul, H. L. Shear, and J. P. Harrington, "A first evaluation of the natural high molecular weight polymeric *Lumbricus terrestris* hemoglobin as an oxygen carrier," *Artificial Cells, Blood Substitutes and Immobilization Biotechnology*, vol. 25, pp. 429-444, 1997.
[24] K. Ishimori and I. Morishima, "Ruthenium-iron hybrid hemoglobins as a model for partially liganded hemoglobin: NMR studies of their tertiary and quaternary structures," *Biochemistry*, vol. 27, pp. 4060-4066, 1988.
[25] K. Ishimori, A. Tsuneshige, K. Imai, and I. Morishima, "Ruthenium-iron hybrid hemoglobins as a model for partially liganded hemoglobin: Oxygen equilibrium curves and resonance raman spectra," *Biochemistry*, vol. 89, pp. 8603-8609, 1989.
[26] R. W. Larsen, D. J. Nunez, J. Macleod, A. K. Shiemke, S. M. Musser, H. H. Nguyen, M. R. Ondrias, and S. I. Chan, "Spectroscopic characterization of heme a reconstituted myoglobin," *Journal of Inorganic Biochemistry*, vol. 48, pp. 21-31, 1992.
[27] S. Neya, M. Nakamura, K. Imai, H. Hori, and N. Funasaki, "Functional comparison of the myoglobins reconstituted with symmetric deuterohemes," *Biochimica et Biophysica Acta*, vol. 1296, pp. 245-249, 1996.

Érico Macedo Polo was born in 1981. Érico Macedo Polo attended to the Computer Sciences degree program at Pontifícia Universidade Católica de Minas Gerais (PUC-MG), but dropped out after 2 years, in order to pursuit his keen interest in Biology. In 2007, he entered the Biological Sciences undergraduate program at the Universidade Federal de São João Del-Rei (UFSJ), where he employed old and new programming skills in order to solve biological problems. In 2012 Érico was awarded the 1st place prize by the Brazilian Genetics Society (SBG) as the best work on Animal Genetics, Evolution and Breeding (scientific initiation category) from the *58º Congresso Brasileiro de Genética***.** Recently, he has been admited in the Genetics, Conservation and Evolutionary Biology master degree program at the National Institute for Amazonian Research (INPA), in Manaus, AM, Brazil where he is beginning to do research on birds phylogeography.