Evolution of haptoglobin: comparison of complementary DNA encoding *Hpa*^{1S} and *Hpa*^{2FS}

Jill L.Brune, Funmei Yang, Don R.Barnett and Barbara H.Bowman

Department of Cellular and Structural Biology, University of Texas Health Science Center, 7703
Floyd Curl Drive, San Antonio, TX 78284, USA

Received 16 February 1984; Revised and Accepted 1 May 1984

ABSTRACT

Haptoglobin is a transport glycoprotein which removes free hemoglobin from the circulation of vertebrates. In human populations haptoglobin is polymorphic due to three alleles, *Hpa*^{1F}, *Hpa*^{1S} and *Hpa*². The *Hpa*² allele is roughly twice the length of the *Hpa*¹ alleles and is the product of a partial gene duplication possibly resulting from an unequal crossover event in a heterozygous genotype *Hpa*^{1F}/*Hpa*^{1S}. In the study described here we compare the cDNA encoding *Hpa*^{1S} to that encoding *Hpa*^{2FS}. Both have a leader sequence followed by the genotypic α chain sequence, a β sequence and an untranslated sequence in the 3' end. The cDNA encoding *Hpa*^{2FS} is composed of α ^{1F} and α ^{1S} domains differing by four nucleotide replacements. *Hpa*^{1S} cDNA contains the same replacement site mutations found in the α ^{1S} domain of *Hpa*^{2FS}, indicating that this coding region has sustained few, if any, mutations since its incorporation into the *Hpa*^{2FS} gene.

INTRODUCTION

Haptoglobin binds hemoglobin irreversibly and transports it to the liver where the complex is degraded (1,2). In humans this protein has served as a model for point mutation, genetic polymorphism, unequal crossing over, gene duplication and phylogenetic associations with the serine protease family (3,4). Haptoglobin has a tetrachain structure ($\alpha\beta$)₂ with disulfide bonds linking the polypeptide chains. In human populations there are three common genetic haptoglobin types, Hp1, Hp2, and the heterozygous phenotype, Hp2-1, reflecting variations in the Hpa polypeptides (5). Two common electrophoretic Hpa1-chains, α 1F (fast) and α 1S (slow) are characterized by a lysine and a glutamic acid residue, respectively, in the 53rd residue of the α 1-chain (5). Two alleles control their structure, *Hpa*^{1S} and *Hpa*^{1F}. The third common allele, *Hpa*², which is roughly twice the length of the *Hpa*¹ alleles, is the product of a partial gene duplication probably resulting from an unequal crossover event in chromosomes of a heterozygote, *Hpa*^{1F}/*Hpa*^{1S} (5). We recently characterized the cDNA encoding *Hpa*² (4). A leader sequence was followed by the α 2-chain and β chain sequences. An arginine residue was deduced between the α - and β -chain

sequences that serves as a site for limited proteolysis leading to the formation of the α and β polypeptides in mature haptoglobin. This finding strengthened the evolutionary homology previously observed in haptoglobin and the serine proteases (6,7). An additional evolutionary point was revealed in the sequence of cDNA encoding *Hpa*^{2FS}. The α ^{1F} and α ^{1S} domains in the *Hpa*^{2FS} gene were found, as expected, to differ in the codons for the 53rd residue (lysine, AAG, in the α ^{1F} domain) and the 112th residue (glutamic acid, GAG in the α ^{1S} domain). In addition, the discovery of three unexpected differences in the homologous domains of the *Hpa*² gene raised the question of whether the mutations occurred before or after the initial crossover event (4).

The study described here furnishes information on the gene structure of *Hpa*^{1S}, its leader sequence, the nucleotides encoding the α and β chains, the probable site of limited proteolysis and provides information that the cDNA encoding the *Hpa*^{1S} gene contains the same replacement sites found in the α ^{1S} domain of the *Hpa*^{2FS} gene.

MATERIALS AND METHODS

Materials

Restriction endonucleases, polymerases, radiolabelled compounds and other reagents utilized in these experiments have been specified previously (4).

Preparation of Probes

The probe used to screen for the *Hpa*¹ cDNA clones was prepared from clone L6-18 isolated by Yang et al. (4). Clone L6-18 contains the cDNA encoding the entire α 2- β haptoglobin sequence. Plasmid DNA from clone L6-18 was digested with *Hpa*II restriction endonuclease and separated by agarose gel electrophoresis. The 550-bp insert fragment was isolated by electroelution. This fragment was then digested with *Dde*I endonuclease to obtain a *Dde*I/*Hpa*II 403-bp fragment. This 403-bp fragment coded for amino acid residues 3-137 of the α 2-haptoglobin chain. The 403-bp fragment was 3'-labelled with [α ³²P] dATP, [α ³²P] dTTP and [α ³²P] dCTP at both *Dde*I and *Hpa*II ends to produce a radiolabelled probe with high specific activity.

Screening of cDNA Clones

The human liver cDNA library was kindly provided by Dr. Stuart H. Orkin, Harvard Medical School, Boston, and has been described (8). The procedures of screening, mapping and sequence determination were the same as described previously for the *Hpa*² complementary DNA (4).

Fig. 1. Comparison of DdeI Restriction Digests of Hpa^2 and Hpa^{1S} Clones. Lanes a and b are two clones, 152c and L6-18, containing Hpa^2 cDNA inserts. Lanes c and d are two clones, 151a and 150a, containing Hpa^{1S} cDNA inserts. Arrows denote internal DdeI fragments characteristic of Hpa^2 (473-bp, which is one of the doublets in lanes a and b; note that the corresponding positions in lanes c and d are single bands) and Hpa^{1S} (297-bp, in lanes c and d).

RESULTS AND DISCUSSION

The library containing human liver cDNA clones was screened with the probe, DdeI/HpaII 403-bp fragment described above. Since this probe contains the sequence coded for amino acid residues 3-137 of the $\alpha 2$ haptoglobin chain, it identified both Hpa^1 and Hpa^2 cDNA clones. The Hpa^1 and Hpa^2 cDNA clones were distinguished according to differences in the

Fig. 2. Structures of the *Hpa*^{1S} and *Hpa*^{2FS} cDNAs and Strategy for Determining the Nucleotide Sequence of *Hpa*^{1S} cDNA. Arrows indicate direction and extent of sequence analysis; solid dots indicate labeling sites. Positions of relevant restriction sites are indicated.

patterns of restriction endonuclease digests. The restriction endonucleases used in this analysis included *Hpa*II, *Dde*I and *Pst*I. The clone L6-18, which was known to contain full length *Hpa*² cDNA, was digested with *Dde*I restriction endonuclease and yielded five internal fragments, 504-bp, 473-bp, 147-bp, 123-bp, and 63-bp (Fig. 1). With the same enzyme, *Hpa*¹ cDNA was expected to yield a 297-bp fragment instead of a 473-bp fragment. Indeed, about half of the positive clones were found to contain the 297-bp fragment (Fig. 1). Sequence analyses have confirmed that these clones contain *Hpa*¹ cDNA. Therefore, the cDNA library was derived from an individual with *Hp*2-1 genotype. One cDNA clone, 150a, contained the entire coding region for the *Hpa*^{1S} sequence. The structures of the *Hpa*¹ and *Hpa*² cDNA are illustrated in Figure 2. This figure also shows the strategy for sequencing the *Hpa*¹ cDNA. The complete sequence of 1222-bp (Fig. 3) was established by analysing each of the two DNA strands or by determining the sequence of the same strand two or more times.

The *Hpa*¹ cDNA contains the same 54-bp leader sequence and same 168-bp noncoding sequence on the 3' end observed in the *Hpa*² cDNA. It is identical with the α ^{1S} domain of the sequence of *Hpa*^{2FS} cDNA characterized by us earlier (4) with two exceptions: residue 2 in the β domain of *Hpa*^{1S} was encoded CTG, rather than CTA, and β residue 6 was CTG, rather than CTA. In

5' CA AGA CCA ACC AAG ⁻¹⁸ ATG AGT GCC CTG GGA GCT GTC ATT GCC CTC CTG CTC TGG GGA CAG CTT TTT GCA -1
~~MET~~ SER ALA LEU GLY ALA VAL ILE ALA LEU LEU LEU TRP GLY GLN LEU PHE ALA

GTG GAC TCA GGC AAT GAC GTC ACG GAT ATC GCA GAT GAC GGC TGC CCG AAG CCC CCC GAG ATT GCA CAT GGC TAT GTG GAG CAC TCG GTT
 1 10 20 30
 VAL ASP SER GLY ASN ASP VAL THR ASP ILE ALA ASP ASP GLY CYS PRO LYS PRO PRO GLU ILE ALA HIS GLY TYR VAL GLU HIS SER VAL
 Hpa^{1S} Chain

CGC TAC GAC TGT AAG AAC TAC TAC AAA CTG CGC ACA GAA GGA GAT GGA GTG TAC ACC TTA AAC AAT GAG AAG CAG TGG ATA AAT AAG GCT
 31 40 50 60
 ARG TYR GLN CYS LYS ASN TYR TYR LYS LEU ARG THR GLU GLY ASP GLY VAL TYR THR LEU ASN ASN GLU LYS GLN TRP ILE ASN LYS ALA

GTT GGA GAT AAA CTT CCT GAA TGT GAA GCA GTA TGT GGG AAG CCC AAG AAT CCG GCA AAC CCA GTG CAG CGG ATC CTG GGT GGA CAC CTG
 61 70 80 1
 VAL GLY ASP LYS LEU PRO GLU CYS GLU ALA VAL CYS GLY LYS PRO LYS ASN PRO ALA ASN PRO VAL GLN ARG ILE LEU GLY GLY HIS LEU
 Hp8 Chain

GAT GCC AAA GGC AGC TTT CCC TGG CAG GCT AAG ATG GTT TCC CAC CAT AAT CTC ACC ACA GGT GGC ACG CTG ATC AAT GAA CAA TGG CTG
 10 20 30
 ASP ALA LYS GLY SER PHE PRO TRP GLN ALA LYS MET VAL SER HIS HIS ASN LEU THR THR GLY ALA THR LEU ILE ASN GLU GLN TRP LEU

CTG ACC ACG GCT AAA AAT CTC TTC CTG AAC CAT TCA GAA AAT GCA ACA GCG AAA GAC ATT GCC CCT ACT TTA ACA CTC TAT GTG GGG AAA
 40 50 60
 LEU THR THR ALA LYS ASN LEU PHE LEU ASN HIS SER GLU ASN ALA THR ALA LYS ASP ILE ALA PRO THR LEU THR LEU TYR VAL GLY LYS

AAG CAG CTT GTA GAG ATT GAG AAG GTT GTT CTA CAC CCT AAC TAC TCC CAG GTA GAT ATT GGG CTC ATC AAA CTC AAA CAG AAG GTG TCT
 70 80 90
 LYS GLN LEU VAL GLU ILE GLU LYS VAL VAL LEU HIS PRO ASN TYR SER GLN VAL ASP ILE GLY LEU ILE LYS LEU LYS GLN LYS VAL SER

GTT AAT GAG AGA GTG ATG CCC ATC TGC CTA CCT TCA AAG GAT TAT GCA GAA GTA GGG CGT GTG GGT TAT GTT TCT GGC TGG GGG CGA AAT
 100 110 120
 VAL ASN GLU ARG VAL MET PRO ILE CYS LEU PRO SER LYS ASP TYR ALA GLU VAL GLY ARG VAL GLY TYR VAL SER GLY TRP GLY ARG ASN

GCC AAT TTT AAA TTT ACT GAC CAT CTG AAG TAT GTC ATG CTG CCT GTG GCT GAC CAA GAC CAA TGC ATA AGG CAT TAT GAA GGC AGC ACA
 130 140 150
 ALA ASN PHE LYS PHE THR ASP HIS LEU LYS TYR VAL MET LEU PRO VAL ALA ASP GLN ASP GLN CYS ILE ARG HIS TYR GLU GLY SER THR

GTC CCC GAA AAG AAG ACA CCG AAG AGC CCT GTA GGG GTG CAG CCC ATA CTG AAT GAA CAC ACC TTC TGT GCT GGC ATG TCT AAG TAC CAA
 160 170 180
 VAL PRO GLU LYS LYS THR PRO LYS SER PRO VAL GLY VAL GLN PRO ILE LEU ASN GLU HIS THR PHE CYS ALA GLY MET SER LYS TYR GLN

GAA GAC ACC TGC TAT GGC GAT GCG GGC AGT GCC TTT GCC GTT CAC GAC CTG GAG GAG GAC ACC TGG TAT GCG ACT GGG ATC TTA AGC TTT
 190 200 210
 GLU ASP THR CYS TYR GLY ASP ALA GLY SER ALA PHE ALA VAL HIS ASP LEU GLU GLU ASP THR TRP TYR ALA THR GLY ILE LEU SER PHE

GAT AAG AGC TGT GCT GTG GCT GAG TAT GGT GTG TAT GTG AAG GTG ACT TCC ATC CAG GAC TGG GTT CAG AAG ACC ATA GCT GAG AAC TAA
 220 230 240
 ASP LYS SER CYS ALA VAL ALA GLU TYR GLY VAL TYR VAL LYS VAL THR SER ILE GLN ASP TRP VAL GLN LYS THR ILE ALA GLU ASN STOP

TGC AAG GCT GGC CGG AAG CCC TTG CCT GAA AGC AAG ATT TCA GCC TGG AAG AGG GCA AAG TGG ACG GGA GTG GAC AGG AGT GGA TGC GAT

AAG ATG TGG TTT GAA GCT GAT GGG TGC CAG CCC TGC ATT GCT GAG TCA ATC AAT AAA GAG CTT TCT TTT GAC CCA AAA AAAA 3'

Fig. 3. The Nucleotide Sequence of the cDNA and Deduced Amino Acid Sequence of *Hpa^{1S}*. The amino acid sequence is numbered according to the system described in Ref. 3.

TABLE
SITE REPLACEMENTS IN HOMOLOGOUS DOMAINS
OF *Hpa*^{2FS} AND *hpa*^{1S} GENES

Residue No. in Domains	Leader -15	α F/S				β		Ref.
		47/106	51/110	52/111	53/112	2	6	
F Domain of <i>Hpa</i> ^{2FS}	Leu	Val	Asn	Asp	Lys			
	CTG	CTA	AAT	GAT*	AAG	--	--	(4)
	TTG	GTA	AAT	GAT	AAG	--	--	(9)
	TTG	GTA	AAT	GAT	AAG	--	--	(10)
S Domain of <i>Hpa</i> ^{2FS}				Asn	Glu	Leu	Leu	
	--	GTG	AAC	AAT	GAG	CTA	CTA*	(4)
	--	CTG	AAC	AAT	GAG	CTG	TTG	(9)
	--	GTG	AAC	AAT	GAG	CTG	CTT	(10)
<i>Hpa</i> ^{1S}	CTG	GTG	AAC	AAT	GAG	CTG	CTG	

*Examination of sequence data revealed α -52 and -89 were GAT and GTT and β -6 was CTA, rather than those 3 codons reported in Ref. 4.

reports of two other *Hpa*^{2FS} cDNAs (9,10) the codon for β -2 was CTG, whereas β -6 has been reported TTG (9) and CTT (10).

In human evolution the nucleotide sequence of the two domains of *Hpa*^{2FS} resulting from the intragenic duplication of the *Hpa*¹ allele have remained contiguous on chromosome 16 (11). Thus the nucleotide divergence in the *Hpa*^{1F} and *Hpa*^{1S} domains of the *Hpa*^{2FS} gene can be observed. The Table is a compilation of site replacements reported in *Hpa*^{2FS} cDNAs from three individuals (4,9,10) and in the *Hpa*^{1S} cDNA reported here. There are only minor variations among the four nucleotide sequences (4,9,10). These occur at residues -15 of the α chain, β -2 and β -6. When the two homologous domains, α ^{1F} and α ^{1S} of *Hpa*^{2FS} are aligned and compared there appears to be one small nucleotide sequence containing four replacement site mutations: 47:106, 51:110, 52:111 and 53:112. Out of the four site replacements observed two result in different but synonymous codons, while the site replacement at α 53/112 is responsible for the lys→glu substitution characteristic of α ^{1F}/ α ^{1S} phenotypes and the site replacement at 52/111 is responsible for asp→asn. The sequence data of *Hpa*^{1S} cDNA reveals that the *Hpa*^{1S} gene contains the identical nucleotide sequence observed in the *Hpa*^{1S} domain of *Hpa*^{2FS}. It was unknown from the *Hpa*^{2FS} characterizations whether the mutations producing three of the site replacements occurred before or

after the crossover event producing the Hpa^{2FS} gene. The Hpa^{1S} cDNA sequence indicates that the Hpa^{1S} domain of the Hpa^{2FS} has sustained few, if any, mutations since its incorporation into the Hpa^{2FS} gene. After this work was submitted for publication, van der Straten et al. (12) reported similar findings after comparing cDNAs coding for α^{2FS} and α^{1S} .

Other plasma proteins that are composed of homologous domains resulting from intragenic duplication are albumin (13), ceruloplasmin (14) and transferrin (15); however, the domains of the Hpa^{2FS} gene display the least evolutionary divergence having nucleotide sequences 98% identical (4), a dramatic contrast to the homologous domains of transferrin, which are 50% identical (16) and of albumin, which are 40-55% identical (13).

The partial gene duplication producing Hpa^{2FS} arose in man and must have occurred soon after the divergence of man (3). Hpa^{2FS} occurs in all human populations studied, although the allele frequency differs widely from population to population, and its distribution has been associated with selective advantage, although the basis for selection is unknown (17,18).

ACKNOWLEDGEMENTS

We thank Dr. S.H. Orkin, Harvard Medical School, Boston, for the human liver cDNA library, Judith Bergeron for technical assistance and Mrs. Betty Russell for preparation of the manuscript. This work was supported in part by NIH grant HD16584 and a Grant-in-Aid from the American Heart Association with funds contributed in part by Texas Affiliate.

REFERENCES

1. Jayle, M.-F. and Moretti, J. (1962) *Prog. Hematol.* 3, 342-359.
2. Polonovski, M. and Jayle, M.-F. (1939) *Bull. Soc. Chim. Biol. Paris* 21, 66-91.
3. Bowman, B.H. and Kurosky, A. (1982) in Harris, H. and Hirschhorn, K., *Advances in Human Genetics*, Vol. 12, Plenum Publ. Corp., N.Y., pp. 189-261.
4. Yang, F., Brune, J.L., Baldwin, W.D., Barnett, D.R. and Bowman, B.H. (1983) *Proc. Natl. Acad. Sci. USA* 80, 5875-5879.
5. Smithies, O., Connell, G.E. and Dixon, G.H. (1962) *Nature* 196, 232-236.
6. Barnett, D.R., Lee, T.-H. and Bowman, B.H. (1970) *Nature* 225, 938-939.
7. Kurosky, A., Barnett, D.R., Lee, T.-H., Touchstone, B., Hay, R.E., Arnott, M.S., Bowman, B.H. and Fitch, W.M. (1980) *Proc. Natl. Acad. Sci. USA* 77, 3388-3392.
8. Prochownik, E.V., Markham, A.F. and Orkin, S.H. (1983) *Jour. Biol. Chem.* 258, 8389-8394.
9. van der Straten, A., Herzog, A., Jacobs, P., Cabezon, T. and Bollen, A. (1983) *EMBO Jour.* 2, 1003-1007.
10. Rauei, G., Bensi, G., Colantuoni, V., Romano, V., Santoro, C., Costanzo, F. and Coreses, R. (1983) *Nucleic Acids Res.* 11, 5811-5819.

11. McGill, J.R., Yang, F., Baldwin, W.D., Brune, J.K., Barnett, D.R., Bowman, B.H. and Moore, C.M. (1984) *Cytogenet. and Cell Genet.* (in print).
12. van der Straten, A., Herzog, A., Caberzón, T. and Bollen, A. (1984) *FEBS* 168, 103-107.
13. Sargent, T.D., Yang, M. and Bonner, J. (1981) *Proc. Natl. Acad. Sci. USA* 78, 243-246.
14. Takahashi, N., Bauman, R.A., Ortel, T.L., Dwulet, F.E., Wang, C.-C. and Putnam, F.W. (1983) *Proc. Natl. Acad. Sci. USA* 80, 115-119.
15. MacGillivray, R.T.A., Mendes, E., Sinha, S.K., Sutton, M.R., Lineback-Zins, J. and Brew, K. (1982) *Proc. Natl. Acad. Sci. USA* 79, 2504-2508.
16. Yang, F., Lum, J.B., McGill, J.R., Moore, C.M., Naylor, S.L., van Bragt, P.H., Baldwin, W.D. and Bowman, B.H. (1984) *Proc. Natl. Acad. Sci. USA* (in print).
17. Kirk, R.L. (1968) The Haptoglobin Groups in Man, in *Human Genetics Monographs*, Vol. 4, Backman, J. and Hauge, M. Eds.), Karger, Basel.
18. Giblett, E.R. (1969) Haptoglobin. in *Genetic Markers in Human Blood*, Davis, F.A. Ed., Philadelphia, Pennsylvania.