The EMBO Journal vol.5 no.9 pp.2275-2283, 1986

Molecular cloning and sequencing of the human erythrocyte 2,3-bisphosphoglycerate mutase cDNA: revised amino acid sequence

Virginie Joulin, Jean Peduzzi[1], Paul-Henri Roméo, Raymonde Rosa, Colette Valentin, Anne Dubart, Bruno Lapeyre[2], Yves Blouquit, Marie-Claude Garel, Michel Goossens, Jean Rosa and Michel Cohen-Solal

Unité INSERM U.91, Hôpital Henri Mondor 94010, Créteil Cedex, and [2]Centre de Biochimie et de Génétique Cellulaires du CNRS, 118, route de Narbonne, 31062 Toulouse Cedex, France

[1]Present address: Museum d'Histoire Naturelle, 63, rue Buffon, 75005 Paris, France

Communicated by P.Slonimski

The human erythrocyte 2,3-bisphosphoglycerate mutase (BPGM) is a multifunctional enzyme which controls the metabolism of 2,3-diphosphoglycerate, the main allosteric effector of haemoglobin. Several cDNA banks were constructed from reticulocyte mRNA, either by conventional cloning methods in pBR322 and screening with specific mixed oligonucleotide probes, or in the expression vector λgt 11. The largest cDNA isolated contained 1673 bases [plus the poly(A) tail], which is slightly smaller than the size of the intact mRNA as estimated by Northern blot analysis (~1800 bases). This cDNA encodes for a protein of 258 residues; the protein yielded 34 tryptic peptides which were subsequently isolated by h.p.l.c. Our nucleotide sequence data were entirely confirmed by the amino acid composition of these tryptic peptides and reveal several major differences from the published sequence; the revised amino acid sequence of human BPGM is presented. These findings represent the first step in the study of the expression and regulation of this enzyme as a specific marker of the erythroid cell line.

Key words: cell differentiation/2,3-bisphosphoglycerate mutase/2,3-diphosphoglycerate/nucleotide and amino acid sequences/recombinant DNA

Introduction

In human and in several other animal species, the functional properties of haemoglobin are regulated within red blood cells by organic phosphates. Such regulation in mammals is effected through 2,3-diphosphoglycerate (2,3-DPG) (Benesch and Benesch, 1967; Benesch *et al.*, 1968), a small molecule present at high concentration inside red blood cells, which binds to the deoxy form of haemoglobin (Benesch *et al.*, 1968) and decreases its oxygen affinity. Therefore 2,3-DPG acts as the key mechanism in facilitating the supply of oxygen to the tissues (for review, see Bunn and Forget, 1986).

Rapoport and Luebering (1950) first described the molecular pathways in red blood cells which lead to the synthesis of 2,3-DPG and its degradation. Originally, it was speculated that both reactions were catalysed by two different enzymes, diphosphoglycerate mutase, which synthesizes 2,3-DPG, and diphosphoglycerate phosphatase which catalyses its degradation.

Since then, it has been shown that, in man, the two enzymatic activities were related to the same multifunctional protein (Rosa *et al.*, 1973, 1975; Sasaki *et al.*, 1975; Rose and Dube, 1976) initially called 2,3-diphosphoglycerate mutase or bisphosphoglycerate synthase and later named 2,3-bisphosphoglycerate mutase (BPGM) (EC 2.7.5.4) according to the official nomenclature. In addition, this enzyme displays a third activity identical to that of another glycolytic enzyme, i.e. phosphoglycerate mutase (PGM) (EC 2.7.5.3). In the red cells nevertheless, the major portion of PGM activity is expressed by PGM, a protein genetically distinct from BPGM but structurally related to it (Fothergill-Gilmore, 1986).

The same multifunctional enzyme has been described in the red blood cells of different species (Rosa *et al.*, 1973), including the rabbit (Narita, 1979), horse (Rose and Dube, 1976) and pig (Sasaki *et al.*, 1976).

In contrast with red blood cells, only minute amounts of 2,3-DPG were detected in other tissues, where it acts as a co-factor to PGM. Moreover, attempts to detect any BPGM activity in human tissues other than red blood cells have failed. BPGM appears therefore to be a specific marker of differentiation in this human erythroid cell line.

The levels of 2,3-DPG and of BPGM enzyme activity increase during the maturation of erythroid precursor cells in rabbit marrow (Narita *et al.*, 1980), in murine Friend cells (Narita *et al.*, 1981) and in human K562 cells (Wu *et al.*, 1982), indicating that BPGM is induced during the late stages of differentiation. More recently it was shown by *in vitro* translation in rabbit reticulocyte lysate and specific immunoprecipitation that BPGM mRNA was present in human reticulocytes and foetal liver but was absent in adult liver (Dubart *et al.*, 1984).

In this paper, we report the molecular cloning of human erythrocyte BPGM cDNA and the deduced BPGM amino acid sequence. The sequence of our BPGM cDNA clones, as well as the amino acid compositions of tryptic peptides of the purified protein, revealed several differences with the protein sequence published earlier (Haggarty *et al.*, 1983). This work is the first step in the study of the expression and regulation of this enzyme as a specific marker of the erythroid cell line.

Results

Isolation of human erythrocyte BPGM. Preparation of specific polyclonal antibodies

BPGM was isolated from normal human red blood cells according to the procedure previously described (Rosa *et al.*, 1983). This procedure, which involves several successive chromatographic steps, is probably more tedious than other more rapid procedures (Haggarty *et al.*, 1983), but in our hands it has given a protein of higher purity, with two identical chains of mol. wt 30 000 daltons as estimated by SDS-polyacrylamide gel electrophoresis. Monospecific polyclonal antibodies against purified human BPGM were raised in adult rabbits as previously described (Dubart *et al.*, 1984) and purified by affinity chromatography. These antibodies did not show any cross-reaction with PGM.

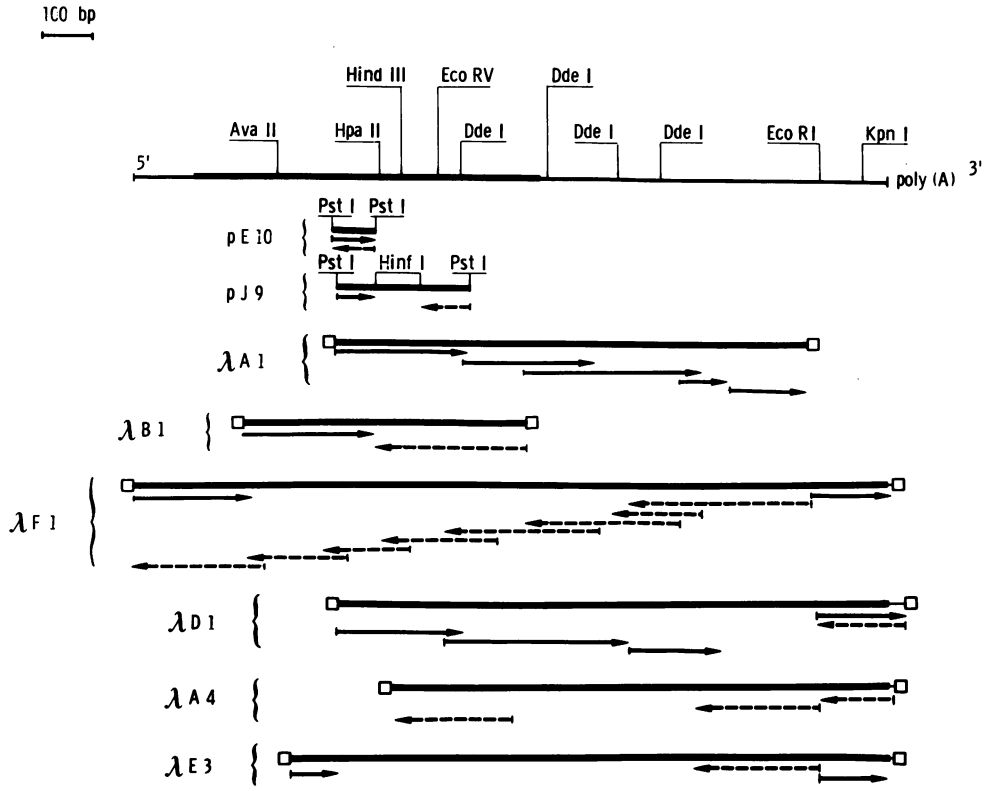

Fig. 1. Strategy for the sequencing of the human BPGM. The top line represents the mRNA, including the 5'- and 3'-untranslated regions and the poly(A) tail. The coding region is indicated as a black box. The partial restriction map was deduced from our sequence data. The open boxes indicate the *Eco*RI linker sequences used for cloning in λgt 11 phage vector. Two short cDNA inserts in pBR322 (pE10 and pJ9) were sequenced by the chemical method of Maxam and Gilbert (1980). All the λgt 11 recombinant clones (λA1, λB1, λF1, λD1, λA4 and λE3) were sequenced after subcloning in phage M13 (Messing and Vieira, 1982) by the method of Sanger *et al.* (1977). Overlapping clones for long fragments were obtained by the method described by Dale *et al.* (1985) with slight modification. The arrows indicate the direction and the extent of the sequence obtained for each individual clone. The solid lines indicate one strand of the DNA corresponding to the coding strand in the mRNA; the dotted lines indicate the other strand.

Isolation and fractionation of human reticulocyte BPGM mRNA

Adult reticulocytes, whose mRNAs contain 10^{-4}% of translatable BPGM mRNA (Dubart *et al.*, 1984) was used as the most convenient source of material.

Blood containing 20% reticulocytes was withdrawn during exchange transfusion from a female patient carrying a sickle cell disease. Previous studies had shown that her red blood cell 2,3-DPG and BPGM activity were high, as is usual in sickle cell disease, and that the electrophoretic mobility of the enzyme was similar to that of a control.

After preparation of total poly(A)$^+$ RNA and its fractionation on a sucrose gradient, the presence of BPGM mRNA was checked by cell-free translation of the contents of each fraction and immunoprecipitation. The fractions containing BPGM mRNA (corresponding to a 12S coefficient) were pooled, ethanol precipitated and used as starting material for cloning. Assay of the radioactivity incorporated in the newly synthesized proteins showed that high mol. wt mRNAs were enriched, and more specifically that the BPGM mRNA was enriched 10–20 times, but that the fractions still contained ~50% globin mRNAs.

Construction of human reticulocyte libraries

Two different cDNA libraries were constructed using the 12S fractions of poly(A)$^+$ reticulocyte RNA.

The first one (bank no. 1) was prepared by conventional methods using GC tailing at the *Pst*I site of plasmid pBR322. Recombinants obtained from bank no. 1 were first screened with [^{32}P]cDNA made by reverse transcription of 9S RNA. This step selected 12 000 non-globin cDNA recombinants which were stored in individual 96-well plates and frozen. Minipreparations of randomly chosen clones indicated that the inserted cDNAs were rather short with an average length of 300 bp, though some were larger (~600 bp). Such a result was probably related to the S1 nuclease step used during the cDNA cloning procedure.

The second cDNA library (bank no. 2) was prepared as described by Okayama and Berg (1982) with slight modifications. After methylation of DNA at potential *Eco*RI sites and addition of *Eco*RI linkers, the resulting cDNAs were fractionated by Agarose A-5m column chromatography. Each fraction was individually ligated with dephosphorylated *Eco*RI cut λgt 11 arms, and packaged. Several banks were obtained (named 2a, 2b, 2c) and stored. Bank no. 2a contained the largest inserts (>1200 bp) and was shown to contain 20% of globin recombinants out of 4.55×10^5 recombinants, while bank no. 2b with shorter inserts (in the range 500–1200 bp) was more abundant (1.5×10^6 recombinants) with ~50% of globin cDNAs. Bank no. 2c was not further characterized. Both banks no. 2a and 2b contain <5% of wild-type λgt 11 phage as assayed by the Lac$^+$ phenotype of recombinant plaques. They were screened without any amplification.

Screening of BPGM cDNAs

Two independent approaches were followed in order to screen the human BPGM cDNA clones.

The first approach was related to the immunological screening of the cDNA library made in λgt 11 phage. Assuming that the concentration of BPGM mRNA was 10^{-4} in the starting material, it was expected to detect at least 10 positive recombinants if there was no enrichment and probably more based on

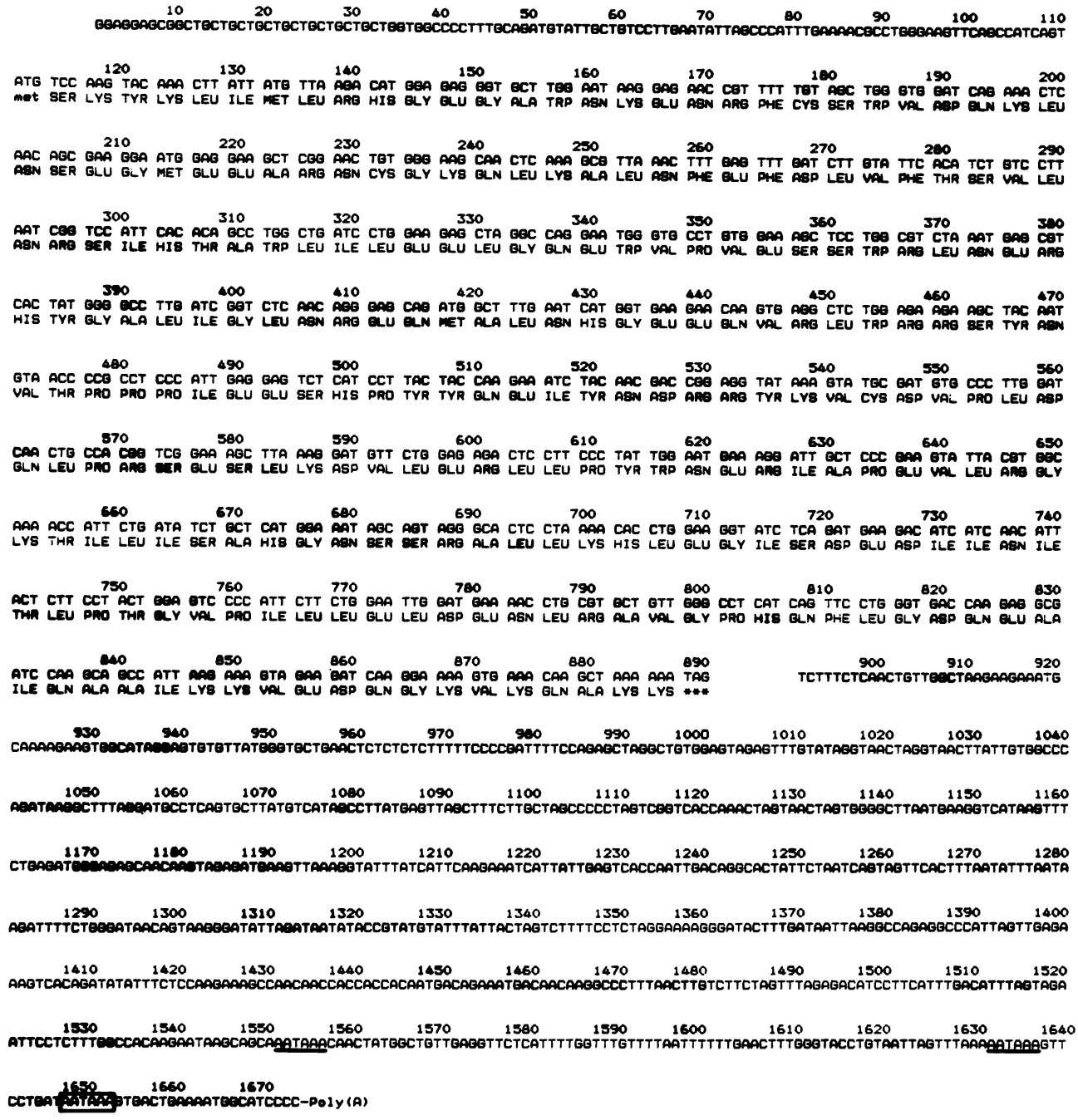

Fig. 2. The DNA sequence of the strand corresponding to the human BPGM mRNA is displayed above the deduced protein sequence; the poly(A) nucleotide tail is not shown. The polyadenylation signal is indicated by an open box. The two potential polyadenylation-like signals found upstream to the more likely authentic signal are underlined.

the fact that the high mol. wt mRNAs were 10 times enriched. A polyclonal rabbit anti-human BPGM antiserum was used to screen 5×10^5 recombinants from bank 2b and 2×10^5 recombinants from bank 2a, but no positive signal was detected.

The second approach for screening was related to oligonucleotide probes. Two probes 16 times degenerated and 14-mer in length were selected from the published amino acid sequence (Haggarty *et al.*, 1983). Probe I: TA(T/C)TA(T/C)CA(A/G)GA(A/G)AT corresponds to amino acids 129−133 Tyr-Tyr-Gln-Glu-Ile and probe II: TA(T/C)TGGAA(T/C)GA(A/G)(A/C)G to amino acids 161−165 Tyr-Trp-Asn-Glu-Arg (according to the sequence of Haggarty *et al.*, 1983).

Screening with oligonucleotide probes was restricted to the cDNA library made using the pBR322 vector (bank no. 1) in order to avoid any possible cross-hybridization with the DNA sequence of the λgt 11 vector. We first screened with probe II because it was thought to hybridize with a region located in the 3′ area of the coding region, and which should be more represented in the short cDNA inserts of this bank. Screening of bank no. 1 gave a high background level for all clones, no distinct signal under the washing conditions used, but several slightly positive clones which were selected and analyzed. Characterization and sequencing of these clones indicated that none of them was related to the human BPGM cDNA. In addition, it was shown that control human α and β globin cDNA plasmids could be detected by probe II under the low stringency washing conditions

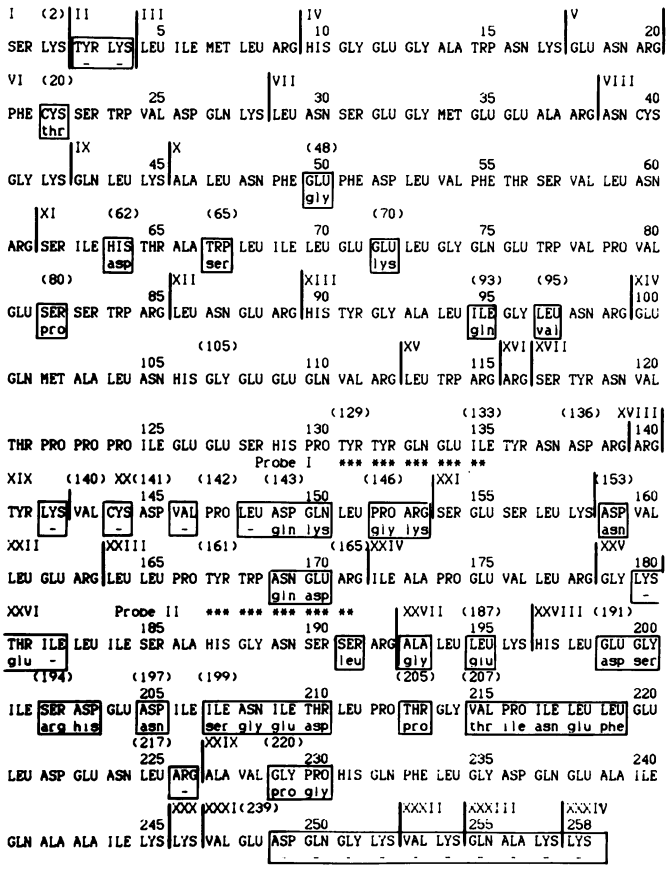

Fig. 3. The upper line corresponds to the amino acid sequence deduced from the cDNA sequence and from the amino acid composition of tryptic peptides of BPGM. Differences observed with Haggarty's sequence (Haggarty *et al.*, 1983) are indicated in lower case letters or as − for missing residues. Numbers in parentheses refer to Haggarty's sequence. The sequences selected for the design of oligonucleotide probes used in the screening of cDNA banks are indicated by *. Tryptic peptides are indicated by Roman numerals (see Table I).

used in these experiments. Conversely, using probe I, six positive clones were selected among the 12 000 non-globin recombinants of bank no. 1, giving a strong positive signal, even after final washing under high stringency conditions (0.2 × SSC at room temperature for 2 × 30 min).

The size of the inserted cDNAs ranges from 100 bp (clone pE10) to 450 bp (clone pJ9). The sequences of E10 and pJ9 were determined (Maxam and Gilbert, 1980) (Figure 1). Clone pE10 corresponds to amino acids 105−136 and clone pJ9 to amino acids 108−201 of the protein sequence of Haggarty *et al.* (1983). The sequence determination of clone pJ9 indicated several amino acid differences upon comparison with the published protein sequence. More precisely, it was found that the sequence of residues 161−165 used to design probe II was Tyr-Trp-Asn-Glu-Arg instead of Tyr-Trp-Gln-Asp-Arg.

Isolation of a full length (λE1) cDNA was performed by screening 6 × 10^5 recombinants of the λgt 11 cDNA libraries (banks no. 2a and 2b) with nick-translated pE10 or pJ9 inserts. Six positive recombinants were picked up by the screening of 2 × 10^5 recombinants from bank no. 2a, and seven by the screening of 10^5 recombinants of bank no. 2b. Screening of the 13 positive λgt 11 recombinant clones with polyclonal anti-BPGM antibodies gave only one positive signal after induction with isopropyl β-D-thiogalactopyranoside (IPTG).

Nucleotide sequence of BPGM cDNAs

Figure 1 gives a summary of the sequencing experiments performed. With the exception of the two short inserts isolated from the pBR322 cDNA bank which were sequenced by the chemical method of Maxam and Gilbert (1980), all other data were obtained after subcloning of inserts into phage M13. Sequencing experiments were performed at least in duplicate and the sequence determination of large fragments was performed by the procedure described by Dale *et al.* (1985). During this study, we sequenced several different clones either completely or partially. The data obtained from individual clones were identical in their common parts.

The sequence of the human reticulocyte BPGM cDNA is presented in Figure 2 and extended for 1673 bases. The coding region extends for 780 bases and encodes for 258 amino acids, plus the initiator and termination codons (Figure 2). This result differs from the 239 amino acids in the published sequence of Haggarty *et al.* (1983) and accounts for the insertion of one amino acid in their sequence on the C-terminal side of positions 139, 140, 141, 142, 173, 175 and 217, for the insertion of two amino acids on the C-terminal side of position 2, and for the presence of an extra tail of 10 amino acids at the C terminus. In addition, interchanges were found in several positions (Figure 3). The 3′-untranslated region is 785 bases long [plus the poly(A) tail]; its G + C content is lower (37.4%) than for the coding region (46.6%). It contains a polyadenylation signal located 21 bases upstream from the poly(A) tail, and two additional potential polyadenylation signals in positions 1551 and 1632.

Isolation and amino acid composition of tryptic peptides

Our nucleotide sequence data for the BPGM cDNA and the derived amino acid sequence were verified at the protein level. Around 1 mg of total purified protein was digested with trypsin. Peptides were isolated by h.p.l.c. on a reverse phase column (Figure 4); their amino acid composition is given in Table I. The data are entirely consistent with our nucleotide sequence and correspond to a total of 34 tryptic peptides (Table I).

Northern blot analysis

Total reticulocyte poly(A)$^+$ RNA and the 12S fraction used as starting material for cloning experiments were electrophoresed and hybridized with nick-translated BPGM cDNA. A single band was found with a length of ~1800 bases (Figure 5).

Discussion

2,3-DPG is the main allosteric effector of haemoglobin, and plays a role in determining the affinity of haemoglobin for oxygen by favoring the R→T transition of the high to low affinity forms. Adaptation to variations in oxygen pressure in red blood cells is effected by the adjustment of the level of 2,3-DPG. Several clinical conditions involve abnormalities in the amount of 2,3-DPG, such as sickle cell disease in which its level is increased, favoring the formation of deoxyhaemoglobin S fibers (Briehl and Ewert, 1973).

After having previously demonstrated the presence of BPGM mRNA in human reticulocytes (Dubart *et al.*, 1984), we presently report the cloning of its cDNA and the determination of the sequence of the enzyme. This represents the first step in the study of the specificity and regulation of the expression of this marker during differentiation of the red blood cell line.

According to the low frequency of specific BPGM mRNA molecules found in human reticulocytes (Dubart *et al.*, 1984) one of the main problems anticipated in the cloning of its cDNA was the presence of >90% of globin mRNA in total reticulocyte

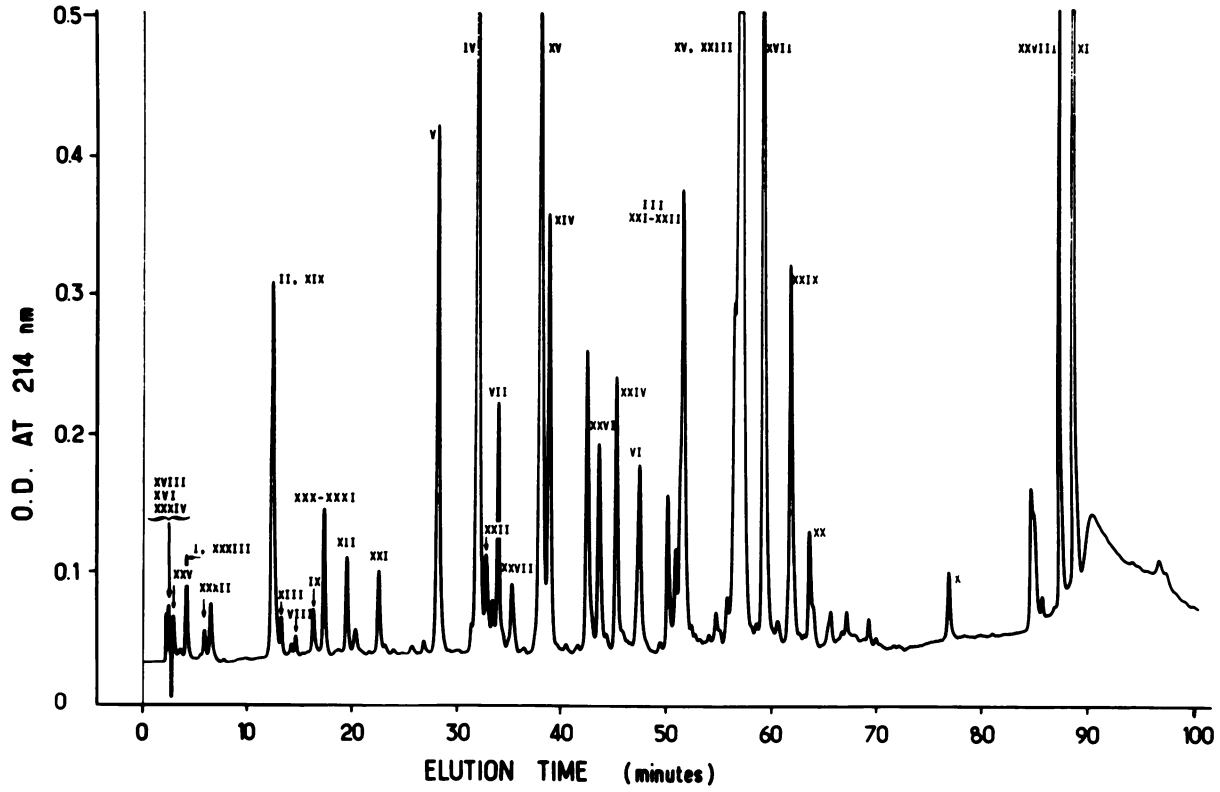

Fig. 4. Elution pattern upon h.p.l.c. fractionation of tryptic peptides of human reticulocyte BPGM. The amino acid composition of individual peptides was determined after acid hydrolysis and the presence of tryptophan evaluated by specific staining (see Table I).

poly(A)$^+$ RNA. The size of BPGM mRNA as determined by sucrose gradient centrifugation was relatively higher than that of globin mRNAs, a finding deriving both from the 2-fold longer coding region predicted by the amino acid sequence (Haggarty *et al.*, 1983) and from the larger untranslated regions found during our study. The gradient fractionation of mRNA was a useful method for initial purification which permitted an ~10-fold enrichment. As the BPGM mRNA is about twice as large as the globin message, the sizing of cDNAs constituted the second crucial step in the cloning procedure. The nature of bank no. 2b, which contained only 20% globin cDNA, indicated that the size fractionation steps at the mRNAs and cDNAs levels were successful.

The positive screening of BPGM recombinant clones was based upon the synthesis of mixed oligonucleotide probes predicted by the published protein sequence (Haggarty *et al.*, 1983). Examination of the sequence indicates that there were very few clusters of codons with limited possibilities according to the genetic code. In addition, the sequences of the designed probes were relatively low in G + C content and, as a consequence, the hybridization temperatures used were relatively low. Furthermore, the occurrence of several discrepancies between the published amino acid sequences and the sequence determined during the present investigation prevented the use of some probes and accounted for the failure observed when we used probe II during screening.

Our first attempts to obtain a non-globin cDNA library from human reticulocytes were made using conventional cDNA synthesis and cloning followed by the elimination of globin recombinants by hybridization with 9S cDNA. Although this bank contained relatively short inserts, screening with synthetic oligonucleotide probes was useful in order to isolate a larger specific probe to be used for subsequent screening.

A full length cDNA bank was made by the modified method of Okayama and Berg (1982) for synthesis of cDNAs and insertion in λgt 11. We decided to use such a vector because it allowed us to screen recombinants by the expression of BPGM using specific antibodies as well as by DNA hybridization. Nevertheless no positive clone was obtained in several attempts to directly screen the cDNA banks with specific antibodies. In addition, among the 13 purified positive clones detected by DNA hybridization, only one expressed the BPGM protein in λgt 11. The further characterization of the clones provided explanations for the negative results according to the size or orientation of the BPGM cDNA inserts. The λgt 11 cDNA cloning system and screening with expression using specific antibodies has been shown to be very useful for several purposes (Young and Davis, 1983). Our results illustrate one of the limitations of such methods: i.e. calculations based upon the abundance of BPGM mRNA in reticulocyte indicate that we should have detected several positive clones. Moreover, the expression of the specific protein in one of the 13 authentic BPGM clones which were isolated indicates that the efficiency of the system is lower than could be expected from a theoretical point of view.

Our results on the cDNA sequence of the human protein indicate that it encodes for a protein 258 amino acids in length. Since the sequence previously published involves 239 residues (Haggarty *et al.*, 1983) and in spite of the specificity of our clones, we have checked and entirely confirmed our sequence data at the protein level by the examination of the amino acid composition of tryptic peptides derived from pure human reticulocyte BPGM. The 3′-untranslated region of the cDNA is rather long (785 bases) in comparison with the length of the coding area for the protein (774 bases). The signal AATAAA which is involved in polyadenylation of the mRNA is located 21 bases before the poly(A) tail. Two additional polyadenylation-like signals were found upstream to this more likely authentic signal. The occur-

Table I. Amino acid composition of tryptic peptides from BPGM

Peptide no.	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17
Lys	1.02 (1)	1.0 (1)		0.93 (1)		0.96 (1)		1.05 (1)	1.03 (1)								
His				0.86 (1)							1.08 (1)		0.99 (1)	1.06 (1)			1.10 (1)
Arg			1.04 (1)		0.98 (1)		1.01 (1)			1.14 (1)	1.23 (1)	1.06 (1)	1.02 (1)	1.20 (1)	0.98 (1)	1 (1)	1.06 (1)
Asx				0.93 (1)	0.98 (1)	0.90 (1)	1.1 (1)	0.98 (1)		3.19 (3)		0.91 (1)	1.1 (1)	1.04 (1)			2.98 (3)
Thr										0.94 (1)	1.08 (1)						0.99 (1)
Ser	0.98 (1)					0.90 (1)	0.81 (1)			0.83 (1)	2.57 (3)						1.65 (2)
Glx				1.16 (1)	1.03 (1)	1.1 (1)	2.82 (3)		0.91 (1)	0.83 (1)	4.71 (5)	0.97 (1)		4.77 (5)			3.83 (4)
Pro											0.80 (1)						4.01 (4)
Gly				1.82 (2)			1.07 (1)	1.0 (1)			1.22 (1)		1.95 (2)	1.09 (1)			
Ala				0.95 (1)			1.05 (1)			0.80 (1)	1.06 (1)		1.06 (1)	0.98 (1)			
Val						0.95 (1)				2.04 (2)	1.97 (2)			1.02 (1)			1.2 (1)
Met			0.70 (1)				0.70 (1)							0.69 (1)			
Cys						0.80 (1)		0.60 (1)									
Ile			1.13 (1)								2.0 (2)		1.05 (1)				2.10 (2)
Leu			2.00 (2)				1.03 (1)		1.07 (1)	3.2 (3)	2.98 (3)	1.06 (1)	1.92 (2)	1.01 (1)	1.02 (1)		
Tyr		1.0 (1)											0.91 (1)				3.6 (4)
Phe						0.93 (1)				2.83 (3)							
Trp[b]				+		+					+				+		

Table I. continued

Peptide no.	T18	T19	T20	T21	T22	T23	T24	T25	T26	T27	T28	T29	T30–31	T32	T33	T34	T21–22
Lys		1.0 (1)		1.12 (1)				1.00 (1)		1.05 (1)		1.0 (1)	2.07 (2)	1.1 (1)	1.0 (1)	1.0 (1)	1.23 (1)
His									1.16 (1)		1.16 (1)	1.00 (1)					
Arg	1 (1)		1.11 (1)		0.98 (1)	0.92 (1)	1.08 (1)		0.98 (1)		1.33 (1)						1.02 (1)
Asp			2.11 (2)		0.97 (1)	1.1 (1)			1.07 (1)		4.70 (5)	1.15 (1)	0.99 (1)				1.03 (1)
Thr									1.0 (1)		1.76 (2)						
Ser				1.70 (2)					2.74 (3)		1.15 (1)						1.79 (2)
Glu			0.96 (1)	1.08 (1)	1.03 (1)	1.27 (1)	1.07 (1)				3.80 (4)	3.77 (4)	1.91 (2)		1.1 (1)		2.06 (2)
Pro			1.66 (2)			1.09 (1)	0.86 (1)				1.64 (2)	1.05 (1)					
Gly								1.00 (1)	1.03 (1)		1.99 (2)	2.10 (2)	0.99 (1)				
Ala							1.00 (1)		0.99 (1)	0.90 (1)		3.6 (4)			0.91 (1)		1.03 (1)
Val			1.83 (2)		0.96 (1)		1.00 (1)				1.21 (1)	1.1 (1)	1.06 (1)	0.98 (1)			
Met																	
Cys			0.52 (1)														
Ile							1.00 (1)		1.97 (2)		4.57 (5)	2.00 (2)					
Leu			1.91 (2)	1.00 (1)	1.05 (1)	1.7 (2)	0.97 (1)		1.1 (1)	2.05 (2)	5.7 (6)	1.11 (1)					1.82 (2)
Tyr		1.0 (1)				0.92 (1)											
Phe												1.00 (1)					
Trp[b]						+											

[a]Values in parentheses indicate the amino acid composition expected according to the cDNA sequence.
[b]Tryptophan detected by specific staining according to Easley (1965).
Tryptic peptides were isolated by h.p.l.c. as indicated in Figure 4 and their amino acid composition determined after acid hydrolysis. Results are expressed as moles of residues per mole of peptide. Tryptophan was assayed by specific staining of each isolated peptide.

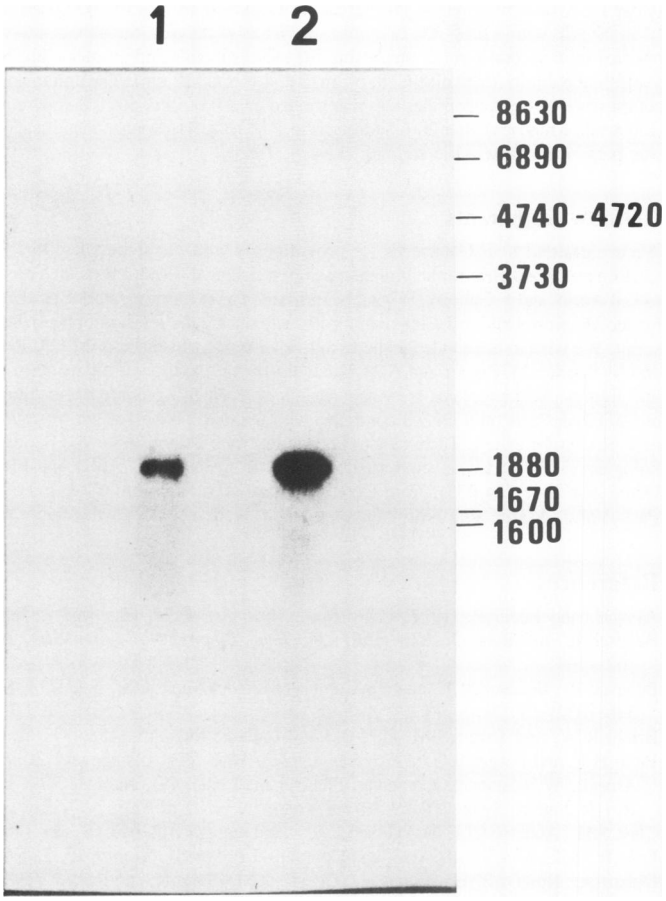

Fig. 5. Northern blot analysis of human BPGM mRNA. The total poly(A)$^+$ RNA prepared from human reticulocyte **(lane 1)** and the 12S fraction **(lane 2)** isolated from a sucrose gradient were electrophoresed on agarose gel, transferred to a nitrocellulose membrane and hybridized with nick-translated λA1-inserted DNA.

rence of several polyadenylation signals could lead to production of mRNAs differing in their 3′ ends. Examination of the lengths of the 3′-untranslated regions in the clones studied did not give any evidence for such a phenomenon, and the only differences observed between clones were related to differences in the length of their poly(A) tails. Northern blot analysis showed that the size of the BPGM mRNA is larger than the longest cDNA sequenced [1673 bases + poly(A) tail], indicating that around 100–150 bases were missing in the 5′ region.

The positive identification of the human BPGM cDNA cloned was related to the expression of the BPGM protein in λgt 11, to the determination of its nucleotide sequence and to the translation into an amino acid sequence. Nevertheless numerous differences were observed between our derived protein sequence and the protein sequence determined earlier (Haggarty *et al.*, 1983). The present sequence corresponds to a protein of mol. wt 29 840 daltons instead of 27 533 daltons, as found by Haggarty and colleagues. Indeed, our results are entirely consistent with the data published earlier indicating a mol. wt of 30 000 daltons (Rose and Whalen, 1973).

The discrepancies between our sequence and that of Haggarty *et al.* are significant since they involve 21% of the residues of the protein. These differences include the insertion of nine residues, the addition of 10 amino acids in the C-terminal region and 36 amino acid replacements. Such marked differences are most unlikely to relate to a polymorphism within the protein, since only one electrophoretic variant was found in two families during the screening of >1500 subjects throughout the world (Chen *et al.*, 1971; R.Rosa, unpublished data); our nucleotide sequence data were strictly identical in five unrelated cDNA clones, and most of the amino acid replacements observed involve multiple base substitutions. Furthermore all our nucleotide data correlate perfectly with the results obtained by analyses at the protein level. One of the most striking distinctions from Haggarty's sequence concerns the lack of the C-terminal tail. In this regard, it is relevant that heterogeneity was not observed in this part of the molecule during the study of the tryptic digests of six different BPGM samples.

Most of the differences observed between the sequence of Haggarty *et al.* and our own are located between residues 140 and 239. Haggarty *et al.* determined the sequence in this region on a large CNBr peptide. This peptide was only partially sequenced by automated degradation and the remainder of the sequence was obtained by several different enzymatic cleavages. Such an approach does not exclude the possible risk of mispositioning small peptides or an eventual misassignment of amides. In view of the lack of precision regarding enzymatic yields and peptide separations, it is impossible to interpret their data further. In addition, alignment of the protein sequences of the human BPGM and yeast PGM (Fothergill and Harkins, 1982) implies the occurrence of gaps in this region. This observation emphasizes the risk involved in the use of data from supposed sequence homologies to determine or to confirm the sequence of related proteins in different species.

The cloning of the BPGM cDNA is a fundamental step in the investigation of its expression during the differentiation of red blood cells and for understanding of its control at the genetic level. At the protein level, new information on its catalytic regulation may be deduced from the amino acid sequence. Comparison of the revised BPGM sequence with that of the yeast PGM (Fothergill and Harkins, 1982) indicates a degree of similarity which is slightly higher than previously described (49% versus 47% homology). This result favors a common origin for both enzymes (Fothergill-Gilmore, 1986).

Materials and methods

BPGM purification and preparation of specific antibodies

BPGM was purified from out-dated normal blood obtained from a blood bank (Rosa *et al.*, 1983). Antibodies against pure BPGM were prepared in rabbits and checked for specificity (Dubart *et al.*, 1984). Non-specific anti-*Escherichia coli* antibodies were removed by binding to BPGM–Sepharose and elution. Affinity-purified antibodies were used at a dilution of 1:50 with respect to the starting antiserum.

Isolation of mRNA and fractionation

Total RNA was isolated from blood withdrawn for blood exchange transfusion of a patient with sickle cell disease by phenol extraction as described earlier (Goossens and Kan, 1981). Poly(A)$^+$ RNA was selected by two successive chromatographic steps using oligo(dT)-cellulose type 3 from Collaborative Research (Aviv and Leder, 1972). After sucrose gradient centrifugation and translation in a rabbit reticulocyte lysate system (BRL) in the presence of [^{35}S]methionine, followed by immunoprecipitation and analysis of products by SDS–polyacrylamide gel electrophoresis (Dubart *et al.*, 1984; Laemmli, 1970), fractions containing enriched BPGM mRNA were pooled and ethanol precipitated.

cDNA synthesis and cloning in plasmid pBR322 (bank no. 1)

A standard procedure was followed for cDNA cloning and the transformation of *E. coli* strain MC1061 (Wickens *et al.*, 1978; Michelson and Orkin, 1982; Hanahan, 1983). Non-globin clones were selected after hybridization with a [^{32}P]-cDNA probe derived from smaller poly(A)$^+$ RNAs isolated on a sucrose gradient (9S) and stored at −80°C in 96-well microtitration plates (Gergen *et al.*, 1979).

cDNA synthesis and cloning in λgt 11 (bank no. 2)

Full length cDNAs were prepared by the procedure described by Okayama and Berg (1982), with slight modification (Lapeyre and Amalric, 1985). After frac-

tionation by Agarose A-5m column chromatography, the double-stranded cDNAs were ligated with dephosphorylated *Eco*RI λgt 11 arms (Vector Cloning System) at a molar ratio of 1:1, packaged with Packagene preparation (Vector Cloning System) and stored at 4°C under chloroform. The libraries were titred on an *E. coli* Y1088 strain (Young and Davis, 1983). Estimation of globin recombinants was made by hybridization with a labelled globin probe as described below.

Screening of pBR322 library

Oligonucleotide synthesis was performed by the solid-phase phosphotriester method on an Applied Biosystem or a Bio-Research DNA synthesizer.

Oligonucleotides were labelled at the 5′ end using [γ-^{32}P]ATP (3000 Ci/mmol) (Amersham, UK) and T4 polynucleotide kinase (Pharmacia) (Maniatis *et al.*, 1982) and purified by chromatography through a 0.9 × 80 cm Sephadex G25 column. Nick translation of DNA was performed by standard techniques (Maniatis *et al.*, 1982).

Replicas from the ordered colonies in 96-well microtitration plates were grown overnight on LB plates and transferred to Whatman 540 filters (Gergen *et al.*, 1979). The filters were hybridized overnight at 30°C with labelled probe (10^6 c.p.m./filter) in 6 × NET (1 × NET = 150 mM NaCl, 15 mM Tris, 1 mM EDTA, pH 7.5) containing 0.5% Nonidet P-40 (NP-40), and 100 μg/ml sonicated salmon DNA. The papers were then washed sequentially with 6 × SSC (1 × SSC = 150 mM NaCl, 15 mM sodium citrate, pH 7.0) containing 0.1% SDS (four times 30 min each at 4°C) followed by twice for 20 min at room temperature and an additional wash for 20 min in 2 × SSC. Filters were finally dried and exposed at −80°C to Kodak AR5 X-ray films with intensifying screens. More stringent washes were performed in 0.2 × SSC at room temperature twice for 30 min.

Screening of λgt 11 library

Overnight cultures of *E. coli* strain Y1088 (Young and Davis, 1983) were infected with 50 000 plaque-forming units of recombinants from bank 2a or 2b, diluted with 7 ml of melted top agar and plated on 150-mm diameter NZCYM plates (Maniatis *et al.*, 1982). Nitrocellulose sheets (Schleicher and Schull BA85 membrane) were overlaid on plates and hybridized with nick-translated probes (Grunstein and Hogness, 1975). For antibody screening, the λgt 11 library and the 13 isolated positive clones were absorbed to *E. coli* strain Y1090, plated on NZCYM plates and grown at 42°C for 4 h (Young and Davis, 1983).

Isolation of phages and plasmids, insert purifications

Candidate positive signals obtained in the first high-density screen were taken through four to five screening rounds with labelled cDNA probe at progressively lower plaque density. Lysogens were made from isolated λgt 11 positive recombinant clones in *E. coli* strain Y1089 (Young and Davis, 1983). Large-scale preparations were performed after chloroform treatment of lysogen strains followed by precipitation with polyethylene glycol. Phage DNA was obtained after two successive CsCl gradients (Maniatis *et al.*, 1982). DNA was digested with *Eco*RI and the fragments separated by electrophoresis on agarose gels and subcloned into the *Eco*RI site of phage M13 mp8 for sequencing purposes (Messing and Vieira, 1982).

Northern blot

About 10 μg of human reticulocyte poly(A)$^+$ RNA were electrophoresed on 1.2% agarose gel in the presence of formaldehyde (Lehrach *et al.*, 1977). Conditions for transfer to nitrocellulose, hybridization and washing of filters were as described by Thomas (1980).

DNA sequence analysis

DNA fragments were labelled at their 3′ ends with terminal deoxytransferase and [α-^{32}P]ddATP (3000 Ci/mmol) (Amersham, UK). After digestion with a second restriction enzyme or strand separation, single end-labelled fragments were sequenced by the chemical modification method with slight modification (Maxam and Gilbert, 1980).

The chain termination procedure (Sanger *et al.*, 1977) was performed after subcloning fragments at the *Eco*RI site of M13 mp8 (Messing and Vieira, 1983) using [α-^{35}S]dATP (400 Ci/mmol) (Amersham, UK). Sequence determination of long fragments was performed using the Cyclone System (IBI, New Haven, CT, USA) with slight modification (Dale *et al.*, 1985).

Sequence data were stored and processed with the DB system programs of Staden (1980). Translation of the nucleotide sequence to amino acid sequence and sequence comparisons were performed using programs described by Staden (1977).

Isolation of tryptic peptides of BPGM and amino acid composition

Purified human BPGM was digested with TPCK-treated trypsin (Worthington, Freehold, NJ, USA) with a ratio trypsin/BPGM of 4% (w/w) (Schroeder *et al.*, 1979). Tryptic peptides were isolated by h.p.l.c. on a Macherey and Nagel Nucleosil C18 column (300 Å, 5 μm, i.d. 4.6 mm, 13 cm) using a Beckman system 343 with a Kratos detector set at 214 nm (Kratos Analytical Institute, Ramsey, NJ, USA). Solvents were made according to Schroeder *et al.* (1979) with the following modifications: solvent A, 0.1% trifluoroacetic acid (TFA), solvent B, 0.1% TFA, 60% acetonitrile. The gradient was linear from 0 to 100% solvent B, in 60 min with a flow-rate of 1 ml/min. When necessary, the peptides were purified by a second h.p.l.c. step using the same solvents and gradient, but on a different column (Brownlee C18, 5 μm, i.d. 4.6 mm, 25 cm). The amino acid composition of tryptic peptides was determined on a Biotronik 6001.E apparatus (Biotronik, Munchen) after 20 h hydrolysis in 6 M HCl. Tryptophan identification was performed by specific staining (Easley, 1965).

Acknowledgements

We are grateful to F.Galacteros for providing us with blood samples, to J.P. Cartron and F.Amalric for helpful discussion, to J.Igolen (Institut Pasteur, Paris), to L.d'Auriol and F.Galibert (Hôpital Saint-Louis, Paris) for the synthesis of oligonucleotide probes and to J.Chapman for his help during the revision of the manuscript. We wish to acknowledge the excellent technical assistance of M.C.Calvin and M.Segear, M.Tassier and A.M.Dulac for the preparation of the manuscript and figures. Sequence data treatment was performed using computer facilities at CITI 2 in Paris on a DPS 8 computer with the help of the French Ministère de la Recherche et de la Technologie (Programme mobilisateur Essor des Biotechnologies). This work was supported by grants from INSERM (U-91 and PRC Génie génétique 124 035), from CNRS (UA 040607) and from the Université Paris-Val de Marne and the Fondation pour la Recherche Médicale Française.

References

Aviv,H. and Leder,P. (1972) *Proc. Natl. Acad. Sci. USA*, **69**, 1408–1412.
Benesch,R. and Benesch,R.E. (1967) *Biochem. Biophys. Res. Commun.*, **26**, 162–167.
Benesch,R., Benesch,R.E. and Yu,C.I. (1968) *Proc. Natl. Acad. Sci. USA*, **59**, 526–632.
Briehl,R.W. and Ewert,S.M. (1973) *J. Mol. Biol.*, **80**, 445–458.
Bunn,H.F. and Forget,B.G. (1986) *Hemoglobin: Molecular, Genetic and Clinical Aspects*. W.B.Saunders Company, Philadelphia/London/Toronto/Mexico City/Rio de Janeiro/Sydney/Tokyo/Hong Kong.
Chen,S.H., Anderson,J.E. and Giblett,E.R. (1971) *Biochem. Genet.*, **5**, 481–486.
Dale,R.M.K., McClure,B.A.M. and Houchins,J.P. (1985) *Plasmid*, **13**, 31–40.
Dubart,A., Roméo,P.-H., Tsapis,A., Goossens,M., Rosa,R. and Rosa,J. (1984) *Biochem. Biophys. Res. Commun.*, **120**, 441–447.
Easley,C.W. (1965) *Biochim. Biophys. Acta*, **107**, 386–388.
Fothergill-Gilmore,L.A. (1986) *Trends Biochem. Sci.*, **11**, 47–51.
Fothergill,L.A. and Harkins,R.N. (1982) *Proc. R. Soc. Lond. Ser B*, **215**, 19–44.
Gergen,S.P., Stern,R.H. and Wensink,P.C. (1979) *Nucleic Acids Res.*, **7**, 2115–2136.
Goossens,M. and Kan,Y.W. (1981) *Methods Enzymol.*, **76**, 805–807.
Grunstein,M. and Hogness,D.S. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 3962–3965.
Haggarty,N.W., Dunbar,B. and Fothergill,L.A. (1983) *EMBO J.*, **2**, 1213–1220.
Hanahan,D. (1983) *J. Mol. Biol.*, **166**, 557–580.
Laemmli,U.K. (1970) *Nature*, **227**, 680–685.
Lapeyre,B. and Amalric,F. (1985) *Gene*, **37**, 215–220.
Lehrach,H., Diamond,D., Wozney,J.M. and Boedtker,H. (1977) *Biochemistry*, **16**, 4743–4751.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, NY.
Maxam,A. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499–560.
Messing,J. and Vieira,J. (1982) *Gene*, **19**, 269–276.
Michelson,A.M. and Orkin,S.H. (1982) *J. Biol. Chem.*, **257**, 14773–14782.
Narita,H., Utsumi,S., Ikura,K., Sasaki,R. and Chiba,H. (1979) *Int. J. Biochem.*, **10**, 25–38.
Narita,H., Ikura,K., Yanagawa,S., Sasaki,R., Chiba,H., Saimyoji,H. and Kumagai,N. (1980) *J. Biol. Chem.*, **255**, 5230–5235.
Narita,H., Yanagawa,S., Sasaki,R. and Chiba,H. (1981) *Biochem. Biophys. Res. Commun.*, **103**, 90–96.
Okayama,H. and Berg,P. (1982) *Mol. Cell. Biol.*, **2**, 161–170.
Rapoport,S. and Luebering,J. (1950) *J. Biol. Chem.*, **183**, 507–516.
Rosa,R., Gaillardon,J. and Rosa,J. (1973) *Biochem. Biophys. Res. Commun.*, **51**, 536–542.
Rosa,R., Audit,I. and Rosa,J. (1975) *Biochimie*, **57**, 1059–1063.
Rosa,R., Préhu,M.O., Albrecht-Ellmer,K. and Calvin,M.C. (1983) *Biochim. Biophys. Acta*, **742**, 243–249.
Rose,Z.B. and Whalen,R.G. (1973) *J. Biol. Chem.*, **248**, 1513–1519.
Rose,Z.B. and Dube,S. (1976) *Arch. Biochem. Biophys.*, **177**, 284–292.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Sasaki,R., Ikura,K., Sugimoto,E. and Chiba,H. (1975) *Eur. J. Biochem.*, **50**, 581–593.

Sasaki,R., Ikura,K., Narita,H. and Chiba,H. (1976) *Agric. Biol. Chem.*, **40**, 2213–2221.
Schroeder,W.A., Shelton,J.B., Shelton,J.R. and Powars,D. (1979) *Hemoglobin*, **3**, 145–159.
Staden,R. (1977) *Nucleic Acids Res.*, **4**, 4037–4051.
Staden,R. (1980) *Nucleic Acids Res.*, **8**, 3673–3694.
Thomas,P.S. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 5201–5205.
Wickens,M.P., Buell,G.N. and Schimke,R.T. (1978) *J. Biol. Chem.*, **253**, 2483–2495.
Wu,Y., Dean,A. and Schechter,A.N. (1982) *Blood*, **60**, 59a.
Young,R.A. and Davis,R.W. (1983) *Science*, **222**, 778–782.

Received on 3 June 1986

Note added in proof

Since this paper was submitted for publication, we heard that similar work has been performed by Sasaki's group, in Kyoto, on the rabbit 2,3-bisphosphoglycerate mutase mRNA [S.I.Yanagawa, K.Hitomi, R.Sasaki and H.Chiba (1986) *Gene*, in press]. Their data indicate that the rabbit mRNA is slightly smaller than the human mRNA (1700 nucleotides versus 1800). Both mRNAs encode for a protein of 258 residues. Divergence at the nucleotide level in the coding region is 10.7% which accounts for only 8 amino acid differences. The differences observed between our amino acid sequence data and those of Haggarty (Haggarty *et al.*, 1983) are located in the same positions in the rabbit protein. Homology in the 3′ untranslated regions is lower than in the coding region; 71% of the bases were found to be identical after computer alignments, which included several gaps in the sequences in order to give the best fits.