The EMBO Journal vol.5 no.8 pp.1849-1855, 1986

Structural analysis of the X-linked gene encoding human glucose 6-phosphate dehydrogenase

G.Martini, D.Toniolo, T.Vulliamy[1], L.Luzzatto[1], R.Dono, G.Viglietto, G.Paonessa, M.D'Urso and M.G.Persico

International Institute of Genetics and Biophysics, CNR, via Marconi 10, 80125 Naples, Italy, and [1]Department of Haematology, Hammersmith Hospital, London W12 0HS, UK

Communicated by L.Luzzatto

We report the isolation and analysis of human genomic DNA clones spanning about 100 kb of the X chromosome and comprising the entire gene coding for the enzyme glucose 6-phosphate dehydrogenase (G6PD). The G6PD gene is 18 kb long and consists of 13 exons: the protein-coding region is divided into 12 segments ranging in size from 12 to 236 bp; an intron is present in the 5′ untranslated region. Mature G6PD mRNA has a single polyadenylation site in HeLa cells. The major 5′ end of mature G6PD mRNA in several cell lines is located 177 bp upstream of the translation initiating codon; longer mRNA molecules extending further in the 5′ direction could be identified by S1 mapping and by comparing genomic and cDNA sequences. The DNA sequence around the major mRNA start is very GC rich; as to putative transcription regulatory sequences, a non-canonical TATA box and 9 CCGCCC elements are present, but no CAAT element could be identified. The genomic DNA we have isolated includes another ubiquitously transcribed region, provisionally named the *GdX* gene. Although the function of *GdX* is as yet unknown, we have established that this gene is located about 40 kb downstream of G6PD and is transcribed in the same direction. A comparative analysis of the promoter region of G6PD and 10 other housekeeping enzyme genes has confirmed the presence of a number of common features. In particular, in the eight cases in which a 'TATA' box is present, a conserved sequence of 25 bp is seen immediately downstream.

Key words: G6PD/housekeeping genes/eukaryotic promoters/human genes/X chromosome

Introduction

Glucose 6-phosphate dehydrogenase (G6PD) is the first and key regulatory enzyme of the so-called pentose phosphate pathway. The main physiological role of G6PD is to provide NADPH, a compound necessary for a number of detoxification and biosynthetic reactions, including fatty acids synthesis (for reviews on G6PD see Beutler, 1983; Luzzatto and Battistuzzi, 1985).

Because of its biochemical role, and because it is found in all cell types and in all organisms thus far analyzed, G6PD can be regarded as the product of a typical 'housekeeping' gene. At the same time, G6PD is subject in some tissues to physiologically important regulatory phenomena. For instance, it has been shown that subjecting rats to several days of fasting followed by a carbohydrate-rich diet results in a greatly enhanced rate of hepatic lipogenesis which is associated with a 13-fold increase in G6PD activity and in G6PD mRNA in liver cells (Kletzien *et al.*, 1985).

The human X-linked genetic locus encoding G6PD is highly polymorphic; more than 300 different variants are known, nearly 100 of which have polymorphic frequencies in different populations. There is good evidence that this results from malaria selection. Although some G6PD variants are clinically asymptomatic, most polymorphic variants are associated with acute hemolytic anemia triggered by ingestion of various agents and many sporadic variants are associated with a more severe condition, i.e. chronic non-spherocytic hemolytic anemia.

The human G6PD gene is located in the sub-telomeric region of the long arm of the X chromosome (band Xq28), where several other genes have been mapped, constituting the so-called G6PD cluster (Human Gene Mapping, 1985). This region is of special interest in genetic terms because of a relatively high rate of crossing over at or near an adjacent fragile site (Szabo *et al.*, 1984; Purrello *et al.*, 1984; Oberlé *et al.*, 1985; Davies *et al.*, 1985).

By means of synthetic oligonucleotides identified on the grounds of partial protein primary sequence data, we have recently isolated cDNA clones bearing the entire sequence coding for human G6PD (Persico *et al.*, 1986). In the present study the cDNA clones were employed to isolate the G6PD gene and to elucidate its structure.

We had previously isolated a cDNA clone (pGD-6405) able to select G6PD-specific mRNA in a positive selection assay (Persico *et al.*, 1981). Since pGD-6405 identifies RNA molecules different from G6PD mRNA (Persico *et al.*, 1986) we infer the existence of a new gene which we designate provisionally as *GdX*. In this study we show that *GdX* is located 40 kb downstream of the G6PD gene and that the two genes are transcribed in the same direction.

Results

Isolation of the G6PD gene; mapping of exons and introns

The λ recombinant clones shown in Figure 1 were obtained by screening genomic libraries with the inserts of G6PD-specific DNA clones (Persico *et al.*, 1986) or with genomic subclones as described in Materials and methods. A restriction map was constructed and was further confirmed by Southern blot analysis of human genomic DNA (data not shown).

Exon sequences were first localized by hybridization of the ^{32}P-labeled G6PD cDNA clones to restriction digests of cloned genomic DNA from phages λGD-4.1 and λGD-TB (see Figure 1) as well as portions of these two recombinants that were subcloned into the plasmid vector pEMBL8 (Dente *et al.*, 1983). The precise locations of each of the 5′ and 3′ exon-intron boundaries were then defined by sequencing the appropriate region of the cloned genomic DNA. From this analysis, we have determined that the gene is 18 kb long and is divided into 13 exons (Figure 1). The sequences at the 5′ and 3′ boundaries of each intron (Figure 2) are in agreement with the consensus sequence for exon-intron boundaries of other eukaryotic genes (Mount, 1982). All introns begin with the dinucleotide GT and end with AG. Intron number 1 is 550 bp in length and interrupts the 5′ untranslated sequence 116 nucleotides upstream of the initiation methionine codon. All other introns, which are within the protein-

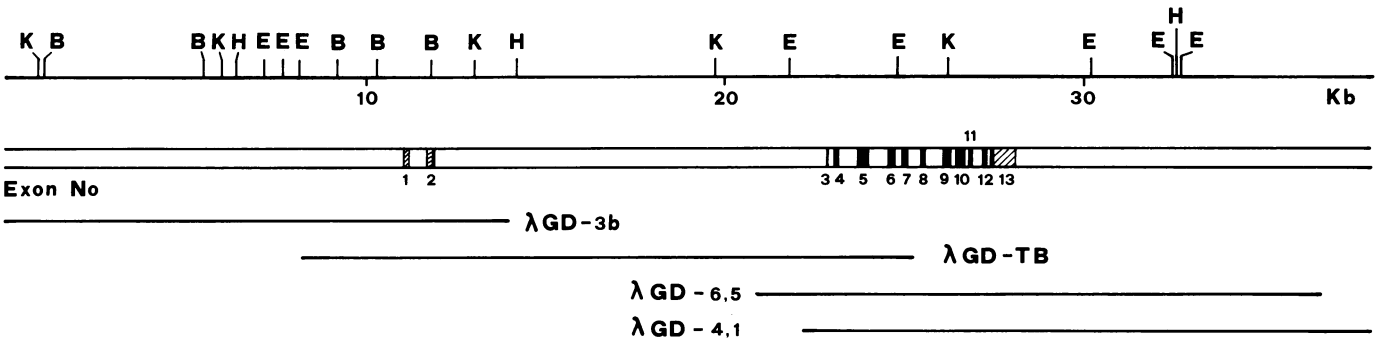

Fig. 1. Physical map of the G6PD gene. The top line shows restriction sites and scale in kb. K = *Kpn*I; B = *Bam*HI; E = *Eco*RI; H = *Hin*dIII. The genomic DNA is shown below the numbered exons, in black for coding region, and in white for non-coding regions. The various recombinant phage clones are shown below.

Exon 1 (60bp)	GACGACGAAGCGCAG gtaaccggcagggcgg	IVS1 (550 bp)	ccttgttaacgagcctttcttccaccag ACAGCGTCATGGCAG
Exon 2 (127bp)	ATG GGT GCA TCG gtgagtatctcctagg	IV S2 (11 Kb)	aaccacacacctgtaccctctgccacag GGT GAC CTG GCC
(12bp coding)	Met Gly Ala Ser		Gly Asp Leu Ala
	3		6
Exon 3 (38bp)	CCC ACC ATC TG gtaagtgtgttccacc	IVS3 (100 bp)	cagctgccctgccctcag G TGG CTG TTC
	Pro Thr Ile Tr		p Trp Leu Phe
	16		18
Exon 4 (109bp)	CCC TTC TTC AAG gtgggtggtgtcaggg	IVS4 (550 bp)	tgtgtgtctgtctgtccgtgtctcccag GCC ACC CCA GAG
	Pro Phe Phe Lys		Ala Thr Pro Glu
	52		55
Exon 5 (218bp)	ATG AGC CAG AT gtaaggcttgccgttg	IVS5 (573 bp)	ggtaacgcaggctccgggctcccagcag A GGC TGG AAC
	Met Ser Gln Il		e Gly Trp Asn
	125		127
Exon 6 (159 bp)	ATG GTG CTG AG gtggggccaagcctgg	IVS6 (180bp)	agttcctccaccttgcccctccctgcag A TTT GCC AAC
	Met Val Leu Ar		g Phe Ala Asn
	178		180
Exon 7 (126bp)	GGG ATC ATC CG gtg	IVS7 (400bp)	ggcgagctctggcctcttccgtccccag G GAC GTG ATG
	Gly Ile Ile Ar		g Asp Val Met
	220		222
Exon 8 (94bp)	CGT GAT GAG AAG gtaggggtgaacccca	IVS8 (450bp)	cccattctctcccttggctttctctcag GTC AAG GTG TTG
	Arg Asp Glu Lys		Val Lys Val Leu
	251		254
Exon 9 (187bp)	AGG TGG GAT G gtaggtgatgccttcg	IVS9 (140bp)	ggtgcgaggcygcccttccgccacgtag GG GTG CCC TTC
	Arg Trp Asp G		ly Val Pro Phe
	314		316
Exon 10 (236bp)	AAC AGA TAC AAG gtgcctpagagaagga	IVS10 (100bp)	catcagcaagacactctctccctcacag AAC GTG AAG CTC
	Asn Arg Tyr Lys		Asn Val Lys Leu
	392		395
Exon 11 (77bp)	TTC GTG CGC AG gtgaggcccaagacctg	IVS11 (300bp)	ctcccnnagccatactatgtcccctcag C GAC GAG CTC
	Phe Val Arg Se		r Asp Glu Leu
	418		420
Exon 12 (93bp)	ATT TAT GGC AG gtgaggaaagggtgggg	IVS12 (97bp)	atgcctctcctccacccgcactctccag C CGA GGC CCC
	Ile Tyr Gly Se		r Arg Gly Pro
	449		451
Exon 13 (695bp)			
(88bp coding)			

Fig. 2. Structural details of the G6PD gene. Each exon−intron junction is illustrated by giving at the extreme left the end sequence of an exon and at the extreme right the beginning of the sequence of the next exon (capital letters). The corresponding amino acids are shown below, numbered in terms of the amino acid sequence. The beginning and ending of each intron are shown in small letters. The sizes of each exon and intron (IVS) are given in parentheses. The length of exon 1 is given as 60 bp based on the main 5′ end of mRNA (see text).

coding portion of the gene, divide the coding region into segments of 12−236 bp. The most 3′ exon contains 88 bp of coding sequence, the terminator codon and all of the 3′ untranslated region (3′UT) which is 607 nucleotides long (see also below).

3′ end of G6PD mRNA in HeLa cells

In order to locate the 3′ end of G6PD mRNA, S1 analysis was carried out with a genomic DNA fragment extending from an *Xho*I site located in the 3′ untranslated region, 153 bp upstream of the poly(A) tract present in the cDNA (Figure 3a) to a genomic *Eco*RI site about 1.9 kb downstream. A probe was prepared by 3′-end-labeling this fragment at the *Xho*I site. The probe was hybridized with total RNA from HeLa cells, subjected to S1 nuclease digestion and the resistant products were separated on a denaturing gel (Figure 3b). Only fragments of approximately 155 nucleotides in length were protected from S1 digestion, suggesting that a single polyadenylation site is used in HeLa cells to position the 3′ end of G6PD mRNA. An uncommon but not unprecedented poly(A) addition signal ATTAAA (Berget, 1984; McLauchlan *et al.*, 1985) is located 13 bp upstream of the poly(A) tail.

5′ end of G6PD mRNA

The 5′ untranslated region of pGD-T-5B, our longest G6PD-specific cDNA clone, extends 577 bp upstream of the initiator methionine codon (Persico *et al.*, 1986). The same sequence is also present in the genomic clones, interrupted by the first intron 116 nucleotides upstream of the translation start. Initially,

Fig. 3. Analysis of the 3′ end of G6PD mRNA. (**A**) Nucleotide sequence determined from genomic clone λGD-4.1 (see Figure 1). The stop codon and the *Xho*I site utilized for S1 mapping are underlined; a double line indicates the polyadenylation site. The sequence in capital letters is also present in cDNA (Persico *et al.*, 1986). (**B**) S1 mapping with a probe labelled at the *Xho*I site underlined in (**A**) and extending 1.8 kb in the 3′ direction. Lanes are: (M1) pEMBL8, *Taq*I digested; (M2) pBR322, *Msp*I digested; (1) total HeLa RNA; (2) control *E. coli* tRNA; (3) probe only. An arrow indicates the protected fragment with HeLa RNA.

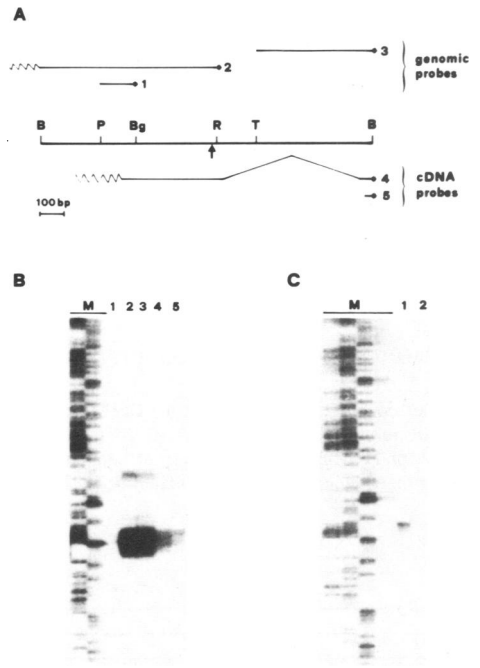

Fig. 4. Analysis of the 5′ end of G6PD mRNA. (**A**) Probes used for analysis are shown as numbered lines above and below a genomic restriction map, segments indicate vector sequences. The arrow against the restriction map indicates the position of the major 5′ end of G6PD mRNA. Only sites relevant for the analysis are shown. (**B**) S1 mapping with probe 4 and control *E. coli* tRNA (1), or total RNA from HeLa (2), hepatoma (3), choriocarcinoma (4), teratoma (5) cell lines. Lanes M display G+A and T+C Maxam and Gilbert sequencing reactions of probe 4. (**C**) Reverse transcriptase analysis with probe 5. Lanes are: (1) poly(A)$^+$ RNA from hepatoma cell line; (2) control *E. coli* tRNA; (M) Maxam and Gilbert sequencing reactions of probe 5.

we attempted to identify the 5′ end of G6PD mRNA in the genomic region further upstream. Three fragments were prepared from genomic subclones, extending in the 5′ direction (probes 1, 2, 3 in Figure 4a). The fragments were 5′ end-labeled with T4 polynucleotide kinase and [^{32}P]ATP and subjected to S1 digestion after hybridization with total RNA from HeLa cells. Only probe 3 showed a protected band of 53 nucleotides in length corresponding to the 3′ junction of intron 1; no S1-resistant material was detected with probe 1 and probe 2 (data not shown). These results as well as inspection of the sequence (see Persico *et al.*, 1986) suggested that the 5′ end of G6PD mRNA in HeLa cells might be within rather than upstream of the 5′ untranslated region of the cDNA clone pGD-T-5B. Therefore, we prepared a new probe from a cDNA fragment of pGD-T-5B extending from a *Bam*HI site at position −67 from the translation start (probe 4 in Figure 4a). The fragment was 5′-end-labeled and used for S1 analysis with total RNA preparations from several sources: HeLa, hepatoma, teratoma and choriocarcinoma cell lines were employed. In each case the major protected fragment was approximately 113 nucleotides long; in addition a minor band of approximately 135 nucleotides in length was observed in all samples (Figure 4b).

In order to confirm these findings we mapped the 5′ end of G6PD mRNA in hepatoma cells by primer extension analysis. A ^{32}P-5′-end-labeled oligonucleotide was prepared by digesting probe 4 with *Hae*III and eluting the 29 nucleotides-long terminal fragment from a urea/polyacrylamide sequencing gel (probe 5, Figure 4a). This fragment was hybridized to poly(A)$^+$ RNA from hepatoma cells and extended with reverse transcriptase. The products of this reaction were displayed on a sequencing gel. A single extended product was observed, having a length similar to that of the major S1-protected fragments (Figure 4c). Primer extended products corresponding to the minor S1-resistant fragments were not detected, most probably due to the lower sensitivity of the primer extension assay. The DNA sequence around the 5′ end of the G6PD gene is shown in Figure 5.

The results reported above locate the major 5′ end of G6PD mRNA at a position within 3 bp around nucleotide −177 from the translation start; they also suggest the presence of a secondary 5′ end at around position −196. The existence of pGD-T-5B in a teratoma cDNA library indicates that, at least in teratoma cells, transcription of G6PD mRNA can start further upstream. However, since S1 analysis of teratoma RNA has failed to reveal a signal corresponding to the complete 5′ untranslated region of pGD-T-5B, we suggest that in these cells mRNA corresponding to pGD-T-5B are a minor species. Because of the very low amount of G6PD mRNA present in teratoma cells, a quantitative assessment of this point will require the use of probes with higher specific activity.

Isolation of clones linking the G6PD gene to pGD-6405 sequences

From a HeLa cell cDNA library we have previously isolated a clone, pGD-6405, able to hybrid-select G6PD-specific mRNA (Persico *et al.*, 1981). The λ recombinant phages λGD-11, λGD-5A and λGD-3C (Figure 6) were obtained by screening genomic libraries with the insert of the cDNA clone pGD-6405 (Persico *et al.*, 1981); repetitive-free fragments were then subcloned and used to enlarge the cloned region by screening again genomic libraries. Re-iteration of this procedure led to the isolation of λ clones spanning between coordinates 61 and 98 (Figure 6). At each step, a detailed restriction analysis was performed to compare new and previous isolates. A recombinant cosmid clone was isolated using the repetitive-free fragment GP3 (see Figure 5) as a probe. This cosmid also hybridized to the G6PD-specific cDNA clones and thus provides a link between the G6PD gene and the genomic region corresponding to pGD-6405. Where phage clones overlap the cosmid clone, restriction analysis with four base-recognizing enzymes gave entirely consistent patterns (not shown).

In order to guard against cloning artefacts, each of the probes

```
                                ccgactcgg acccgcgaac aggcgaaggg ttcccggggg agtggcgcgg cagaaggccc cgcccaagag ccgagggaca gcccagagga -561
ggcgtggcca cgctgccggc ggaagtggag ccctccgcga gcgcgcgagg ccgccgggga ggcggggaaa ccggacagta ggggcggggc gggccggcga tggggatgcg -451
ggagcactac gcggagctgc acccgtgccc gccggaattg gggatgcaga gcagcggcag cgggtatggc aggcaccggc gggccggcct ccagcgcagg tgcccgagag -341
                              (  ••     •• •   •  •••••••)
gcaggggctg gcctgggatg cgcgcgcacc tgccctcgcc ccgccccgcc cgcacgaggg gtggtggccg aggccccgcc ccgcacgcct cgccgaggcg ggtccgctca -231
gcccaggcgc ccgcccccgc ccccgccgat taaatgggcc ggcggggctc agcCCCCGGA AACGGTCGTA CACTTCGGGG CTGCGAGCGC GGAGGGCGAC GACGACGAAG -121
CGCAGgtaaccggcagggcgggggcgcgcagg               IVS1 (550 bp)           tctcctgttcttctgccttgttaacgagcctttcttccaccagACAGC -111
GTCATGGCAG AGCAGGTGGC CCTGAGCCGG ACCCACGTGT GCGGGATCCT GCGGGAAGAG CTTTTCCAGG GCGATGCTTC CATCAGTCGG ATACACACAT ATTCATCATC -1
```

Fig. 5. Nucleotide sequence at the 5′ end of the G6PD gene. Numbering starts at the initiation codon and disregards intron 1. Exon 1 and exon 2 up to the initiation codon are in capital letters; a 3 bp uncertainty has to be considered for the beginning of exon 1. An arrow indicates the site where a sequence identical to cDNA pGD-T-5B begins. A region of homology to the SV40 21 bp repeat is shown by brackets; dots indicate bases in common. GGGCGG elements in direct or reverse orientation are underlined. The non-canonical TATA sequence is boxed.

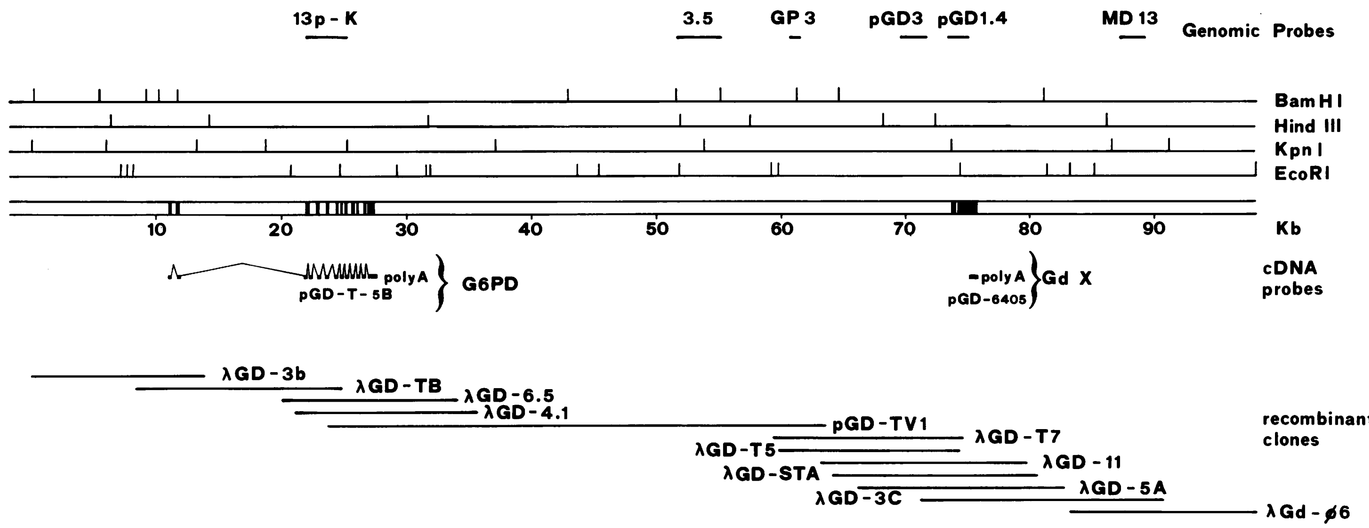

Fig. 6. Extended physical map of the G6PD gene region. Upper lines show sites for four restriction sites. The double line below represents genomic DNA with exons in full black, transcription is from left to right. Diagrams below indicate cDNA clones for the G6PD and *GdX* genes. Lines below indicate recombinant phage and cosmid clones from which the map was constructed. Lines at the top indicate unique sequences used as probes in screening libraries or in Southern blots of human DNA.

shown in Figure 1 was hybridized to Southern blots of total genomic DNA from at least five individuals digested with at least five different restriction enzymes. Only the bands expected from restriction analysis of the clones shown in Figure 5 appeared (Toniolo *et al.*, 1984a and data not shown). These data validate the restriction map depicted in Figure 5. In addition they prove that there is no other G6PD-like gene or pseudogene in the human genome. One of the genomic probes, GP3 (Figure 6) has been previously mapped to Xq28 by analysis of somatic cell hybrids (Martin-De Leon *et al.*, 1985) and by *in situ* hybridization (Purrello *et al.*, 1984). Thus, the entire region shown in Figure 6 is now conclusively assigned to that chromosomal region.

Discussion

We have recently reported the sequence of human G6PD cDNA from which the complete amino acid sequence of the protein has been derived. (Persico *et al.*, 1986). We now report the structure of the human G6PD gene as established from overlapping genomic clones. The G6PD gene consists of 13 exons and 12 introns, covering about 18 kb.

Figure 5 shows the primary sequence of the G6PD gene around its 5′ end: numbering starts from the translation start codon and it disregards the intron present in the 5′ untranslated region. The 5′ terminus of the most abundant species of G6PD mRNA in several cell lines has been identified around position −175 by S1 nuclease and primer extension analysis. From S1 analysis, we also have evidence that transcription of G6PD can start more upstream, because the 5′ terminus of a minor S1-protected RNA species is located at about position −196. Moreover, we have previously reported the sequence of a cDNA clone from teratocarcinoma cells which we now show to contain a sequence identical to genomic DNA, starting at position −577. This suggets that at least rarely transcription can start much further upstream than it would appear from the 5′ terminus of the most abundant G6PD mRNA species. In this discussion we shall refer to the region upstream of position −177 as the G6PD promoter. This region exhibits several noteworthy features.

First, the sequence ATTAAAT, embedded within a highly GC-rich region, is present at position −202 (about 20 nucleotides upstream of the major mRNA 5′ end). A similar sequence, usually referred to as the TATA box (Breathnach and Chambon, 1981), was found at corresponding position in many genes and it was shown by *in vitro* mutagenesis to be important in determining the sites of transcription initiation. The importance of the context in which short regulatory sequences are located is well

Table I. Comparison of promoter regions of several genes coding for housekeeping enzymes

Gene	Species	CCGCCC or GGGCGG	CAAT (position)[a]	TATA	Multiple 5′ mRNA ends	Sequence alignment downstream of the TATA box[b]	Percent agreement 25 bp after TATAA	References
Consensus						TATAAA$^{T}_{G}$GCGGCCGCCGCGGC$^{G}_{T}$CGG$^{C}_{T}$G$^{G}_{C}$C$^{G}_{C}$		
DHFR	Human	3	No	−27	Yes?[e]	C C AG A AG G G	72	Chen *et al.*, 1984
GAPDH	Chicken	7	−74[c]	−27	No	Ĝ − − A -- Ĝ	72	Stone *et al.*, 1985
ALAS	Chicken	4	−139, −105[d]	−30	No	G G A - -	80	Maguire *et al.*, 1986
SOD	Human	4	−128[c], −69[c]	−29	No	TA T - G A A G	68	Levanon *et al.*, 1985
TIM	Chicken	2	No	−29	No	A C	92	Straus and Gilbert, 1985
G6PD	Human	9	No	−30	Yes	AT -G G G A C	76	This study
TK	Chicken	5	No	−27	Yes	- C C CT GATT -	64	Kwoh and Engler, 1984
ADA	Human	5	No	−27	No ?[f]	- GA -TG - C --	64	Valerio *et al.*, 1985
HPRT	Mouse	3	No	No	Yes[g]	No TATA		Melton *et al.*, 1986
PGK	Human	2	No	No	Yes	No TATA		Singer-Sam *et al.*, 1984
HMGCR	Hamster	3	−43[d]	No	Yes	No TATA		Reynolds *et al.*, 1984

[a]From major 5′ mRNA end; [b]in each line, the letters represent changes required to the consensus above, ^ indicates insertion of one base and − indicates a gap of one base; [c]CATT; [d]CACT; [e]minor start sites could reflect methodological artifacts; [f]possibility of additional upstream initiation sites not ruled out; [g]recent data shows the situation to be equivalent in humans (Kim *et al.*, 1986).

illustrated by the G6PD gene. Indeed, analogy with other genes suggests that the very same sequence ATTAAAT fulfils the role of a TATA box at the 5′ end of the gene and the role of a polyadenylation signal at the 3′ end of the gene (compare Figures 3a and 5).

Second, the GGGCGG hexanucleotide and/or its complement CCGCCC is an important component of several viral and cellular promoters and it interacts *in vitro* with a host cell-encoded transcription factor, Sp1 (for a review, see Kadonaga *et al.*, 1986). A comparison of several promoters has led to the identification of the consensus Sp1-binding decanucleotide $^{G}_{T}$GGG-CGG$^{GGC}_{AAT}$ Kadonaga *et al.*, 1986). The core CCGCCC sequence is present six times and its inverted complement GGGCGG is present three times in the region around the 5′ end of G6PD mRNA. In three cases two elements overlap each other and have the same orientation; four elements perfectly match the consensus decanucleotide. All the sequence elements but one are located in the region upstream of the ATTAAAT sequence, their positions ranging from 12 to 400 bp upstream of ATTAAAT. The remaining GGGCGG sequence is found 70 bp downstream of the major mRNA 5′ end and it is located within the first intron, 10 bp downstream of the 5′ intron−exon junction.

Third, no CAAT element, frequently present at −70 to −90 bp (Esfradiatis *et al.*, 1980) in various eukaryotic genes and known to affect the level of transcription, is found in either direct or opposite orientation in the very GC-rich region extending 220 bp upstream of the main 5′ end of G6PD mRNA. We note, however, the existence of an ATTG at position −411.

Fourth, a region of more extended homology with the SV40 21 bp repeat (Benoist and Chambon, 1981) is observed at position −314, i.e. 112 bp upstream of ATTAAAT: 13 nucleotides can be matched with no insertion or deletion.

The structure we have described has all the features of the functional gene coding for human G6PD. The DNA region we have isolated, however, extends some 65 kb downstream from the polyadenylation site of G6PD mRNA. Towards the 3′ end of this region is located the genomic sequence corresponding to cDNA pGD-6405, oriented in the same direction of transcription as G6PD (Figure 6). This clone, isolated from a size-selected HeLa cell-cDNA library, was able to hybrid-select from human fibroblast poly(A)$^+$ RNA, a fraction of RNA able to direct *in vitro* the synthesis of several polypeptides including G6PD (Persico *et al.*, 1981). Thus, pGD-6405 is related to G6PD both by physical linkage (Figure 6) and by some sequence homology. Sequencing the 608 bp insert of pGD-6405 revealed features typical of an mRNA 3′ untranslated region (Toniolo *et al.*, 1984b) with no striking homology with G6PD cDNA (Persico *et al.*, 1986). Longer cDNA clones have been mapped in the vicinity of the genomic region corresponding to pGD-6405 (Toniolo *et al.*, unpublished). Thus, there are two major mRNA species corresponding to the genomic region depicted in Figure 6. The major G6PD-specific mRNA is shown on the left, and it does not terminate with the pGD 6405 sequence as we had previously stated (Toniolo *et al.*, 1984a, b; Battistuzzi *et al.*, 1985). For convenience we shall refer to the genomic region comprising pGD-6405 as the *GdX* gene, although we must emphasize that

we have not yet proven the existence of a *GdX* protein product. The data reported above can be explained in a variety of ways, most of which can be visualized as falling into one or the other of two groups of models. The first group postulates a single primary transcript, which would have to be about 85 kb long. In this case the result of positive selection would be explained by the ability of pGD-6405 to select hypothetical splicing-intermediate RNA molecules having the G6PD coding sequence covalently linked to pGD-6405 sequences. In contrast, the second group of models postulates two separate primary transcripts. In this case, pGD-6405 DNA selects is own mRNA which in turn selects G6PD mRNA through hydrogen bonding with a common short sequence. Transcripts corresponding to pGD-6405 have been detected in up to 10 different cell types (data not shown) suggesting that *GdX* is a housekeeping gene.

A comparison of the promoter regions of G6PD and 10 other housekeeping enzyme genes reveals several common features (see Table I), some of which have been previously noticed (Melton *et al.*, 1986). First, the GGGCGG hexanucleotide and/or its complement CCGCCC are invariably present in multiple copies. Second, with respect to the TATA box, three different situations are seen. A canonical TATA box may be present, or it may be absent from the expected location, or it may be present at the correct location but in a variant form. With a normal TATA box only one 5′ end of mRNA is seen. However, with a variant or absent TATA box multiple 5′ ends are usually seen, confirming that this element plays a crucial role in phasing firmly the site at which transcription starts. Third, if we align the TATA box or its equivalent for the eight genes where it exists, a 'consensus' sequence for the DNA region between it and the main cap site emerges. Each of these genes has at least 68% homology with the consensus sequence, which we have not found in specialized differentiation genes. It is possible therefore that this element, together with the others that have been described, may be of some importance in preventing these genes from ever being completly switched off, a characteristic feature which can be taken as the operational definition of housekeeping genes.

Materials and methods

Isolation and characterization of genomic clones

The isolation of cDNA clones has been previously reported (Persico *et al.*, 1981 and 1986). The following human genomic libraries have been utilized: partial *Mbo*I digest of genomic DNA cloned in Charon 28 (Hieter *et al.*, 1980); partial *Alu*I/*Hae*III digest of total genomic DNA cloned in Charon 4A (Lawn *et al.*, 1978); partial *Mbo*I digest of cell-sorted X chromosome cloned in Charon 30 (Kunkel *et al.*, 1985); partial *Eco*RI digest of total genomic DNA from a female individual cloned in Charon 4A as described (Maniatis *et al.*, 1982); partial *Mbo*I digest of total genomic DNA cloned in the cosmid pTM (Grosveld *et al.*, 1982). Screening of these libraries was performed with nick-translated probes as described (Maniatis *et al.*, 1982). For the construction of subclones, restriction fragments were isolated from acrylamide gels and ligated into convenient sites of plasmid pEMBL8 (Dente *et al.*, 1983). DNA was sequenced by the procedure of Maxam and Gilbert (1977).

S1 and reverse transcriptase analysis

Total RNA frpm HeLa, choriocarcinoma (JEG) (Kolher and Bridon, 1971), teratocarcinoma (Ntera2) (Andrews *et al.*, 1984), hepatoma (HepG2) (Aden *et al.*, 1979) cell lines was isolated by cell lysis in 4 M guanidine thiocyanate and sedimentation through 5.7 M CsCl (Chirgwin *et al.*, 1979). Poly(A)$^+$ RNA from hepatoma (HepG2) was selected by chromatography on oligo(dT) cellulose (Aviv and Leder, 1972).

S1 analysis was carried out according to Favaloro *et al.* (1980), with gel-purified restriction fragments labelled at their 3′ or 5′ ends as described (Maniatis *et al.*, 1982). Hybridization temperatures were 52 and 59°C for 3′ and 5′ analysis, respectively. Digestions were performed for 3 h at 15°C in the presence of 400 U/ml of S1.

Reverse transcriptase analysis was carried out as described (Reynolds *et al.*, 1984) with minor modifications.

Acknowledgements

We thank Drs P.Leder, F.G.Grosveld and S.A.Latt for providing genomic libraries; Dr A.Simeone for providing the teratocarcinoma RNA; Drs P.Verde, F.Blasi and E.Boncinelli for critical reading of the manuscript and helpful discussions; Ms. Maria Terracciano, Carmela Salzano and Concetta Sole for excellent technical assistance. This work was carried out with the financial support of the project 'Ingegneria genetica e basi molecolari delle malattie ereditarie' of CNR, Italy, and of a program grant by MRC, Great Britain.

References

Aden,P.A., Fogel,A., Plotkin,S., Damjanov,I. and Knowles,B.B. (1979) *Nature*, **282**, 615–616.

Andrews,P., Lamjanov,I., Simon,D., Banting,G.S., Carlin,C., Dracopoli,L. and Fogh,J. (1984) *Lab. Invest.*, **50**, 147–162.

Aviv,H. and Leder,P. (1972) *Proc. Natl. Acad. Sci. USA*, **69**, 1408–1412.

Battistuzzi,G., D'Urso,M., Toniolo,D., Persico,G.M. and Luzzatto,L. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1465–1469.

Benoist,C. and Chambon,P. (1981) *Nature*, **290**, 304–310.

Beutler,E. (1983) In Stanbury,J.B., *et al.* (eds), *The Metabolic Basis of Inherited Disease*. 5th ed., McGraw-Hill, New York, 1629–1653.

Berget,S.M. (1984) *Nature*, **309**, 179–182.

Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349–383.

Chen,M.J., Shimada,T., Moulton,A.D., Cline,A., Humphries,K., Maizel,J. and Nienhuis,A.W. (1984) *J. Biol. Chem.*, **259**, 3933–3943.

Chirgwin,J.M., Przybyla,A.E., MacDonald,R.J. and Rutter,W.J. (1979) *Biochemistry*, **18**, 5294–5304.

Davies,K.E., Mattei,M.G., Mattei,J.F., Veenema,H., McGlade,S., Harper,K., Tommerup,N., Nielsen,K.B., Mikkelsen,M., Beighton,P., Drayna,D., White,R. and Pembrey,M.E. (1985) *Hum. Genet.*, **70**, 249–255.

Dente,L., Cesareni,G. and Cortese,R. (1983) *Nucleic Acids Res.*, **11**, 1645–1655.

Esfradiatis,A., Posakony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spiritz,R.A., De Riel,J.K., Forget,B.G., Slighton,L., Blechl,A.E., Smithies,O., Barralle,F.E., Shoulders,C.C. and Proudfoot,N.J. (1980) *Cell*, **21**, 653–668.

Favaloro,J,. Freisman,R. and Kamen,R. (1980) *Methods Enzymol.*, **65**, 718–748.

Grosveld,F.G., Lund,T., Murray,A.L., Dahl,H.H.M. and Flavell,R.A. (1982) *Nucleic Acids Res.*, **10**, 6715–6732.

Hieter,P.A., Max,E.E., Seidman,J.G., Maizel,J.V.,Jr. and Leder,P. (1980) *Cell*, **22**, 197–207.

Human Gene Mapping 8 (1985) *Cytogenet. Cell Genet.*, **21**, Nos. 1–4.

Kadonaga,J.T., Jones,K.A. and Tjian,R. (1986) *TIBS*, **11**, 20.

Kim,S.E., Moores,J.C., David,D., Respess,J.G., Jolly,D.J. and Friedman,T. (1986) *Nucleic Acids Res.*, **14**, 3103–3118.

Kletzien,R., Prostko,C.R., Stumpo,D.J., McClung,J.K. and Dreher,K.L. (1985) *J. Biol. Chem.*, **260**, 5621–5624.

Kolher,P.O. and Bridon,W.E. (1971) *J. Clin. Endocrinol. Metab.*, **32**, 683–690.

Kunkel,L.M., Lalande,M., Monaco,A.P., Flint,A., Middlesworth,W. and Latt,S.A. (1985) *Gene*, **33**, 251–258.

Kwoh,H. and Engler,J.A. (1984) *Nucleic Acids Res.*, **12**, 3959–3971.

Lawn,R.M., Fritsch,E.F., Parker,R.C., Blake,G. and Maniatis,T. (1978) *Cell*, **15**, 1157–1174.

Levanon,D., Lieman-Hurwitz,J., Dafni,N., Wigderson,M., Sherman,L., Bernstein,Y., Laver-Rudich,Z., Danciger,E., Stein,O. and Groner,Y. (1985) *EMBO J.*, **4**, 77–84.

Luzzatto,L. and Battistuzzi,G. (1985) *Adv. Hum. Genet.*, **14**, 217–329.

Maguire,D.J., Day,A.R., Borthwick,I.A., Srivastava,G., Wigley,P.L., May,B.K. and Elliot,W.H. (1986) *Nucleic Acids Res.*, **14**, 1379–1391.

Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory Press, NY.

Martin-De Leon,P.A., Wolf,S.F., Persico,G., Toniolo,D., Martini,G. and Migeon,B.R. (1985) *Cytogenet. Cell. Genet.*, **39**, 87–92.

Maxam,A.M. and Gilbert,W. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 560–566.

McLaugchlan,J., Gaffney,D., Whitton,J.L. and Clements,J.B. (1985) *Nucleic Acids Res.*, **13**, 1347–1368.

Melton,D.W., McEwan,C., McKie,A.B. and Reid,A.M. (1986) *Cell*, **44**, 319–328.

Mount,S.M. (1982) *Nucleic Acids Res.*, **10**, 459–472.

Oberlé,I., Drayna,D., Camerino,G., White,R. and Mandel,J.L. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2824–2828.

Persico,M.G., Toniolo,D., Nobile,C., D'Urso,M. and Luzzato,L. (1981) *Nature*, **294**, 778–780.

Persico,M.G., Viglietto,G., Martini,G., Toniolo,D., Paonessa,G., Moscatelli,C., Dono,R., Vulliamy,T., Luzzatto,L. and D'Urso,M. (1986) *Nucleic Acids Res.*, **14**, 2511–2522.

Purrello,M., Nussbaum,R., Rinaldi,A., Filippi,G., Traccis,S., Latte,B. and Siniscalco,M. (1984) *Hum. Genet.*, **65**, 295–299.

Reynolds,G.A., Basu,S.K., Osborne,T.F., Chin,D.J., Gil,G., Brown,M.S., Goldstein,J.L. and Luskey,K.L. (1984) *Cell*, **38**, 275–285.
Singer-Sam,J., Keith,D.H., Tani,K., Simmer,R.I., Shively,L., Lindsay,S., Yoshida,A. and Riggs,A.D. (1984) *Gene*, **32**, 409–417.
Stone,E.M., Rothblum,K.N., Alevy,M.C., Kuo,T.M. and Schwartz,R.J. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1628–1632.
Straus,D. and Gilbert,W. (1985) *Mol. Cell. Biol.*, **5**, 3497–3506.
Szabo,P., Purrello,M., Rocchi,M., Archidiacono,N., Alhadeff,B., Filippi,G., Toniolo,D., Martini,G., Luzzatto,L. and Siniscalco,M. (1984) *Proc.Natl. Acad. Sci. USA*, **81**, 7855–7859.
Toniolo,D., D'Urso,M., Martini,G., Persico,M., Tufano,V., Battistuzzi,G. and Luzzatto,L. (1984a) *EMBO J.*, **3**, 1987–1995.
Toniolo,D., Persico,M.G., Battistuzzi,G. and Luzzatto,L. (1984b) *Mol. Biol. Med.*, **2**, 89–103.
Valerio,D., Duyvesteyn,M.G.C., Dekker,B.M.M., Weeda,G., Berkvens,T.M., van der Voorn,L., Ormondt,H. and van der Erb,A.J. (1985) *EMBO J.*, **4**, 437–443.

Received on 5 May 1986