Biochem. J. (1988) 254, 509–517 (Printed in Great Britain)

Nucleotide sequence of the phosphoglycerate kinase gene from the extreme thermophile *Thermus thermophilus*

Comparison of the deduced amino acid sequence with that of the mesophilic yeast phosphoglycerate kinase

Derrick BOWEN,* Jennifer A. LITTLECHILD,* John E. FOTHERGILL,† Herman C. WATSON* and Len HALL*‡

*Department of Biochemistry, University of Bristol, School of Medical Sciences, Bristol BS8 1TD, U.K., and †Department of Biochemistry, University of Aberdeen, Marischal College, Aberdeen AB9 1AS, U.K.

Using oligonucleotide probes derived from amino acid sequencing information, the structural gene for phosphoglycerate kinase from the extreme thermophile, *Thermus thermophilus*, was cloned in *Escherichia coli* and its complete nucleotide sequence determined. The gene consists of an open reading frame corresponding to a protein of 390 amino acid residues (calculated M_r 41791) with an extreme bias for G or C (93.1%) in the codon third base position. Comparison of the deduced amino acid sequence with that of the corresponding mesophilic yeast enzyme indicated a number of significant differences. These are discussed in terms of the unusual codon bias and their possible role in enhanced protein thermal stability.

INTRODUCTION

It is generally believed that only a small amount of extra energy is required in a protein structure to enhance its thermal stability. Several comparisons have been made between proteins for which both mesophilic and thermophilic structures are available, for example, ferredoxins (Perutz & Raidt, 1975) and glyceraldehyde-3-phosphate dehydrogenase (Walker *et al.*, 1980) and, as a result, various mechanisms by which enhanced protein thermal stability may be achieved have been proposed (for recent reviews see Klibanov, 1983; Mozhaev & Martinek, 1984). With the development of modern protein engineering techniques it has now become possible to evaluate such proposals in an experimental system.

As part of a systematic approach towards an understanding of protein thermal stability we embarked on a detailed comparative study of the enzyme phosphoglycerate kinase (PGK; EC 2.7.2.3) from a range of different mesophilic and thermophilic organisms. Here we report on the isolation and subsequent sequencing of the PGK gene obtained from a bacterium, *Thermus thermophilus*, which is capable of growth at 85 °C but fails to proliferate below 47 °C (Fujita *et al.*, 1976). Subsequent work will involve determining the tertiary structure of *T. thermophilus* PGK (see Littlechild *et al.*, 1987) for detailed comparison with that of its mesophilic counterpart. Site-directed mutagenesis will then be used in an effort to introduce into the mesophilic enzyme the properties of the thermophilic protein.

Yeast PGK has been extensively characterized. In common with all PGKs studied to date it is active as a monomer converting 1,3-bisphosphoglycerate to 3-phosphoglycerate with the concomitant phosphorylation of adenosine diphosphate. Its amino acid sequence has been determined from the protein (Perkins *et al.*, 1983) and from the gene (Hitzeman *et al.*, 1982), and the tertiary structure has been solved to high resolution by crystallographic studies (Watson *et al.*, 1982). Site-directed mutants of yeast PGK have also been expressed using yeast shuttle-vector systems (Mas *et al.*, 1987; Wilson *et al.*, 1987). *T. thermophilus* PGK exhibits similar enzymic properties to its mesophilic counterparts; however, it is distinctive in its exceptional stability towards heat. These findings, when coupled with the conservation of amino acid sequence (see Mori *et al.*, 1986), make PGK an attractive model system with which to investigate the molecular basis of thermal stability in proteins.

Whilst *T. thermophilus* PGK has been purified to homogeneity and its amino acid composition determined (Nojima *et al.*, 1979), its primary sequence has not previously been reported. The DNA sequence of the structural gene presented here has now permitted a comparison of the deduced amino acid sequence and composition with that of the mesophilic yeast PGK. Differences observed are discussed in terms of their relevance to protein thermal stability.

MATERIALS AND METHODS

Materials

Restriction endonucleases, the Klenow fragment of DNA polymerase I, bacteriophage T4 polynucleotide kinase, bacteriophage T4 DNA ligase, nick-translation kits and [γ-^{32}P]ATP (3000 Ci/mmol) were purchased from Amersham International. NEN GeneScreen *Plus* hybridization transfer membrane and [α-^{32}P]dATP (> 800 Ci/mmol) were obtained from Du Pont (U.K.)

Abbreviation used: PGK, phosphoglycerate kinase.

‡ To whom correspondence should be addressed.

Ltd., Southampton, U.K. Oligonucleotide probes and M13 universal sequencing primer were synthesized on a Du Pont Coder 300 DNA synthesizer using phosphoramidite chemistry. All other chemicals were of AnalaR quality or the highest grade available.

Purification of *T. thermophilus* PGK and limited sequence analysis

T. thermophilus HB-8 cells were grown in a 60-litre fermenter in a rich medium essentially as described by Oshima & Imahori (1971). Typically 400 g of frozen cell paste was used to prepare 24 mg of purified PGK by a modification of the method of Nojima *et al.* (1979), as previously described (Littlechild *et al.*, 1987). A 500 μg sample of the purified PGK was dialysed against 5% (v/v) formic acid and used for *N*-terminal sequence analysis. A further 500 μg sample was succinylated and digested with trypsin. The resultant peptides were separated by h.p.l.c. on a C_{18} μbondapak column using a linear gradient of acetonitrile/methanol/propan-2-ol (1:1:1, by vol.) and detection at 214 nm. The well-resolved peptides were applied to a gas-phase sequence analyser and the partial sequences obtained are indicated in Fig. 2.

Preparation of *T. thermophilus* genomic DNA

A frozen paste (0.5 g) of *T. thermophilus* HB-8 cells was thawed and resuspended in 40 ml of 1 mM-EDTA/20 mM-Tris/HCl buffer, pH 8. Following centrifugation (2500 ***g*** for 5 min at 4 °C) the cells were resuspended in 3.2 ml of 50 mM-Tris/HCl buffer, pH 8, containing 25% (w/v) sucrose, and 0.6 ml of a freshly prepared 20 mg of chick egg-white lysozyme/ml solution was added. The suspension was swirled gently on ice for 5 min. Cells were then lysed by the addition of 0.6 ml of 0.5 M-EDTA solution, pH 8 (with NaOH), and 0.5 ml of 20% (w/v) SDS solution. After a further 5 min on ice, 2.5 mg of proteinase K was added and the cell lysate was incubated at 45 °C for 30 min. One-tenth volume of 3 M-sodium acetate solution, pH 5.5 (with acetic acid), was added and the lysate was extracted with an equal volume of phenol/chloroform (1:1, v/v). Total nucleic acid was then precipitated from the aqueous phase with 2 vol. of ethanol at −20 °C, and reprecipitated several times to remove traces of phenol. The nucleic acid was redissolved in 5 ml of 0.3 M-NaCl/30 mM-sodium citrate buffer, pH 7, and incubated at 37 °C for 20 min in the presence of 40 units of ribonuclease A/ml and 200 units of ribonuclease T1/ml. The solution was then extracted with phenol/chloroform (1:1, v/v) and repeatedly ethanol precipitated as described above. Finally, the purified DNA was redissolved in water and stored at −20 °C.

Southern-blot analysis

T. thermophilus genomic DNA was digested with the appropriate restriction endonucleases, and the products were separated by electrophoresis on a 0.8% (w/v) agarose gel. DNA was then blotted by capillary transfer (Southern, 1975) onto an NEN GeneScreen *Plus* membrane and then prehybridized for 6 h in hybridization buffer lacking probe. Oligonucleotide probes, radiolabelled with [γ-^{32}P]ATP (3000 Ci/mmol) and bacteriophage T4 polynucleotide kinase (Wallace *et al.*, 1979), were hybridized at 37 °C for 18 h in 50 mM-Tris/HCl, pH 7.6, containing 0.9 M-NaCl and 1% SDS. Filters were then washed with successive changes of 0.9 M-NaCl/90 mM-sodium citrate buffer, pH 7, at increasing temperatures until an acceptable background level was obtained. Larger double-stranded DNA probes (gel-purified restriction fragments) were ^{32}P-labelled by nick translation (Rigby *et al.*, 1977) and hybridized at 42 °C for 18 h in 50 mM-Tris/HCl, pH 7.6, containing 0.9 M-NaCl, 1% SDS and 50% (v/v) formamide. Filters were subsequently washed as recommended by the membrane manufacturer. Radioactive bands were detected by autoradiography at −70 °C with Kodak X-Omat S X-ray film and Kodak Lanex intensifying screens.

Construction and screening of a genomic library

T. thermophilus DNA (2 μg) was digested with the appropriate restriction endonuclease (*Bam*HI or *Hin*dIII, see the Results section) and fractionated on a preparative 0.7% low-melting temperature agarose gel. A gel slice containing DNA within the required size range was excised and melted at 65 °C for 5 min after addition of an approximately equal volume of 20 mM-Tris/HCl buffer, pH 8, containing 1 mM-EDTA. It was then extracted once with an equal volume of aqueous phenol (equilibrated with the above buffer) and the aqueous phase was repeatedly ethanol precipitated to remove traces of phenol and other impurities.

Approximately 100 ng of the size-selected genomic DNA was ligated with 100 ng of appropriately restricted, alkaline phosphatase-treated, plasmid pAT153 (Twigg & Sherratt, 1980) and used to transform competent *Escherichia coli* HB101 rec A$^-$ cells (Boyer & Roulland-Dussoix, 1969). Recombinant plasmids containing *T. thermophilus* PGK genomic sequences were then identified by colony filter hybridization *in situ* (Grunstein & Hogness, 1975) using a radiolabelled probe.

DNA sequence analysis

Suitable gel-purified restriction fragments were subcloned into M13 mp10 or mp11 vectors (Messing & Vieira, 1982). DNA sequence analysis was carried out using the dideoxy chain termination method of Sanger *et al.* (1977, 1980). Band compression in polyacrylamide/urea sequencing gels was overcome by increasing gel temperature during electrophoresis (achieved by increasing the voltage). This necessitated the clamping of a 2 mm thick aluminium plate to the front glass plate, to dissipate the heat uniformly and prevent band distortion or cracking of the glass plates. Autoradiographs were read manually and sequence overlaps were aligned by eye.

RESULTS AND DISCUSSION

Isolation and sequence determination of the *T. thermophilus* PGK gene

Total genomic DNA libraries prepared using extremely G+C-rich DNA, cloned into standard *E. coli* vectors, tend to be non-random, containing an over-representation of the more A+T-rich sequences and a corresponding under-representation of the more G+C-rich fragments. It is therefore advantageous to enrich for the required sequence by preparing a sub-genomic library using appropriately size-selected restricted DNA. Consequently, duplicate Southern blots of total *T. thermophilus* DNA digested with a series of

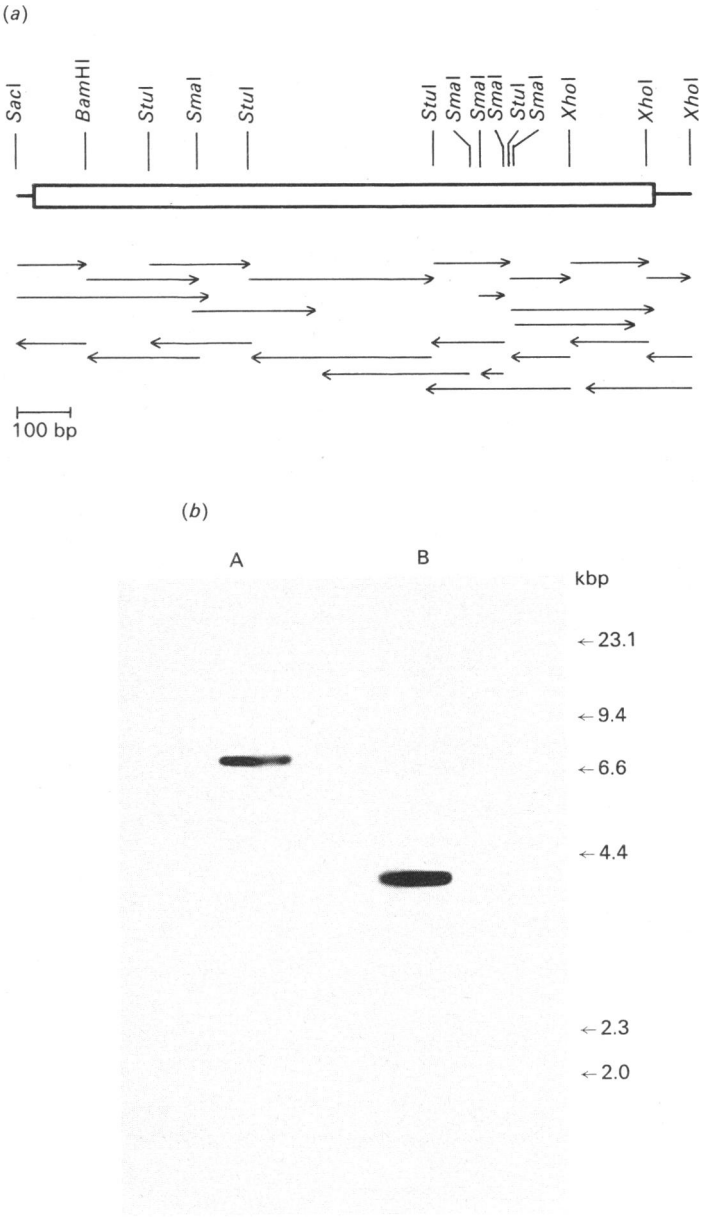

Fig. 1. Genomic organization and sequencing strategy for the *Thermus thermophilus* PGK gene

(*a*) Restriction map of the PGK gene contained within recombinant plasmid pTtPGK-1. The boxed region represents the PGK-coding region. Only those sites for restriction enzymes that were utilized for DNA sequencing are shown. Arrows indicate the direction and extent of individual sequencing runs. Many fragments were sequenced several times, although for clarity only a single arrow is shown for each different restriction fragment. (*b*) Southern-blot analysis of total *T. thermophilus* genomic DNA restricted with *Hin*dIII (lane A) or *Bam*HI (lane B), and probed with a *Thermus* PGK gene fragment (see the Methods section). Arrows indicate the positions of *Hin*dIII-restricted bacteriophage λ DNA size markers.

hexanucleotide-specific restriction enzymes were probed separately with two redundant oligonucleotides. The first oligonucleotide [5′-GT(C/G) GAC TAC AAC GT(C/G) CC-3′] represented just four of the 128 possible sequences coding for amino acid residues 18–24 (Val-Asp-Tyr-Asn-Val-Pro) of *T. thermophilus* PGK and was based on the extreme codon bias previously observed in the isopropylmalate dehydrogenase gene of this organism (Kagawa *et al.*, 1984), the only protein-coding gene from *T. thermophilus* for which sequence data were available. For this gene a G or C residue was found in 89% of the codon third base positions. The second more redundant probe [5′-GT(C/G) GAY TAY AAY GT(C/G) CCN GT-3′], represented 128 of the 512 possible sequences coding for amino acid residues 18–25 of *T. thermophilus* PGK (Val-Asp-Tyr-Asn-Val-Pro-Val). This oligonucleotide was slightly longer than the first probe, but covered essentially the same region of the gene and relied to a much lesser degree on predicted codon usage.

Neither of these oligonucleotides behaved in the predicted manner when used to probe Southern blots of restricted *T. thermophilus* DNA. Both produced positive signals, but these were extremely weak and were lost when the blots were washed at temperatures well below

```
GAGCTCGTCCTGAAGAAGGGGGTCTAG ATG CGG ACC CTT TTG GAC CTG GAC CCC AAG GGC AAG CGG GTC CTG
                            Met Arg Thr Leu Leu Asp Leu Asp Pro Lys Gly Lys Arg Val Leu
                                                5                  10                  15

GTG CGG GTG GAC TAC AAC GTC CCC GTC CAA GAC GGG AAG GTC CAG GAC GAG ACC CGG ATC CTG GAA
Val Arg Val Asp Tyr Asn Val Pro Val Gln Asp Gly Lys Val Gln Asp Glu Thr Arg Ile Leu Glu
                    20                  25                  30                  35

AGC CTC CCC ACC CTC CGC CAC CTC CTC GCC GGG GGG GCT TCC CTC GTC CTC CTC TCC CAC CTG GGC
Ser Leu Pro Thr Leu Arg His Leu Leu Ala Gly Gly Ala Ser Leu Val Leu Leu Ser His Leu Gly
            40                  45                  50                  55

CGC CCC AAG GGC CCG GAC CCC AAG TAC TCC CTG GCC CCG GTG GGG GAG GCC TTG AGG GCC CAC CTC
Arg Pro Lys Gly Pro Asp Pro Lys Tyr Ser Leu Ala Pro Val Gly Glu Ala Leu Arg Ala His Leu
 60                  65                  70                  75                  80

CCA GAG GCC CGC TTC GCC CCC TTC CCT CCG GGC TCG GAG GAG GCG AGG CGG GAG GCG GAG GCC CTG
Pro Glu Ala Arg Phe Ala Pro Phe Pro Pro Gly Ser Glu Glu Ala Arg Arg Glu Ala Glu Ala Leu
                85                  90                  95                 100

AGG CCC GGG GAG GTC CTC CTC CTG GAG AAC GTC CGC TTT GAG CCG GGA GAG GAG AAG AAC GAC CCC
Arg Pro Gly Glu Val Leu Leu Leu Glu Asn Val Arg Phe Glu Pro Gly Glu Glu Lys Asn Asp Pro
    105                 110                 115                 120                 125

GAG CTT TCC GCC CGC TAC GCC AGG CTC GGG GAG GCC TTC GTC CTG GAC GCC TTC GGG AGC GCC CAC
Glu Leu Ser Ala Arg Tyr Ala Arg Leu Gly Glu Ala Phe Val Leu Asp Ala Phe Gly Ser Ala His
                130                 135                 140                 145

CGG GCC CAC GCC AGC GTG GTG GGG GTG GCG AGG CTC CTC CCC GCC TAC GCC GGC TTC CTC ATG GAG
Arg Ala His Ala Ser Val Val Gly Val Ala Arg Leu Leu Pro Ala Tyr Ala Gly Phe Leu Met Glu
        150                 155                 160                 165

AAG GAG GTG AGG GCC CTT TCC CGC CTC CTC AAG GAC CCG GAA AGG CCC TAC GCC GTG GTG CTG GGC
Lys Glu Val Arg Ala Leu Ser Arg Leu Leu Lys Asp Pro Glu Arg Pro Tyr Ala Val Val Leu Gly
170                 175                 180                 185                 190

GGG GCC AAG GTC TCG GAC AAG ATC GGG GTG ATT GAG AGC CTC CTT CCC CGC ATA GAC CGC CTC CTC
Gly Ala Lys Val Ser Asp Lys Ile Gly Val Ile Glu Ser Leu Leu Pro Arg Ile Asp Arg Leu Leu
            195                 200                 205                 210

ATT GGC GGG GCC ATG GCC TTC ACC TTC CTC AAG GCC CTA GGG GGA GAG GTG GGG AGG AGC CTG GTG
Ile Gly Gly Ala Met Ala Phe Thr Phe Leu Lys Ala Leu Gly Gly Glu Val Gly Arg Ser Leu Val
    215                 220                 225                 230                 235

GAG GAG GAC CGG CTG GAC CTG GCC AAG GAC CTC TTG GGG CGG GCC GAG GCC TTG GGG GTC AGG GTC
Glu Glu Asp Arg Leu Asp Leu Ala Lys Asp Leu Leu Gly Arg Ala Glu Ala Leu Gly Val Arg Val
                240                 245                 250                 255

TAC CTC CCC GAA GAC GTG GTG GCG GCG GAG CGC ATA GAG GCG GGG GTG GAG ACC CGG GTC TTC CCG
Tyr Leu Pro Glu Asp Val Val Ala Ala Glu Arg Ile Glu Ala Gly Val Glu Thr Arg Val Phe Pro
        260                 265                 270                 275

GCC CGG GCC ATC CCC GTC CCC TAC ATG GGC CTG GAC ATC GGC CCC AAG ACC CGG GAG GCC TTC GCC
Ala Arg Ala Ile Pro Val Pro Tyr Met Gly Leu Asp Ile Gly Pro Lys Thr Arg Glu Ala Phe Ala
280                 285                 290                 295                 300

CGG GCC CTG GAA GGG GCG AGG ACG GTC TTC TGG AAC GGG CCC ATG GGG GTC TTT GAG GTG CCT CCC
Arg Ala Leu Glu Gly Ala Arg Thr Val Phe Trp Asn Gly Pro Met Gly Val Phe Glu Val Pro Pro
            305                 310                 315                 320

TTT GAC GAG GGG ACC TTG GCC GTG GGG CAG GCC ATC GCC GCC CTC GAG GGC GCC TTC ACC GTG GTG
Phe Asp Glu Gly Thr Leu Ala Val Gly Gln Ala Ile Ala Ala Leu Glu Gly Ala Phe Thr Val Val
    325                 330                 335                 340                 345

GGC GGG GGC GAC TCG GTG GCC GCG GTG AAC CGC CTG GGC CTT AAA GAG CGC TTC GGC CAC GTC TCC
Gly Gly Gly Asp Ser Val Ala Ala Val Asn Arg Leu Gly Leu Lys Glu Arg Phe Gly His Val Ser
                350                 355                 360                 365

ACC GGG GGC GGG GCG AGC CTG GAG TTC CTG GAA AAG GGC ACC CTG CCC GGC CTC GAG GTC CTG GAA
Thr Gly Gly Gly Ala Ser Leu Glu Phe Leu Glu Lys Gly Thr Leu Pro Gly Leu Glu Val Leu Glu
        370                 375                 380                 385

GGC TAA GGCGCGGGCGTTTAGAATGGGCCTGGGGCGTGGAGCCCCAGGCCATGGCTTACGGCAAGGCCCACCTCGAG
Gly ***
390
```

Fig. 2. Complete DNA sequence of the *Thermus thermophilus* PGK gene and immediate flanking sequences

The partial amino acid sequences of tryptic peptides derived from purified *T. thermophilus* PGK are underlined.

the expected melting point. Control experiments using the same blots and an oligonucleotide probe of similar length and redundancy, but corresponding to a region of the isopropylmalate dehydrogenase gene, produced strong hybridization signals, even under stringent wash conditions. Nevertheless, although there were both qualitative and quantitative differences in the results obtained with the two different PGK probes, both identified a 4.3 kb *Bam*HI fragment and a 7 kb *Hin*dIII fragment.

A sub-genomic library was therefore prepared using 4.3 kb size-selected *Bam*HI-cleaved, *T. thermophilus* DNA, cloned into the unique *Bam*HI site of the plasmid pAT153. Hybridization *in situ* of duplicate nitrocellulose filters containing 200 transformants, with each of the oligonucleotide probes, produced two very strong positive clones with the least redundant probe. The other probe yielded a very high background, thereby masking the positive clones, presumably because of the much greater redundancy of this mixed oligonucleotide.

Restriction analysis of the two positive recombinant clones indicated that they contained identical inserts.

Table 1. Codon usage in the *Thermus thermophilus* PGK gene

Optimal (o) and non-optimal (x) codons observed in *E. coli* are indicated. Term, termination code.

First position	Second position U			C			A			G			Third position
U	Phe	x	3	Ser		0	Tyr	x	0	Cys		0	U
	Phe	o	13	Ser		6	Tyr	o	7	Cys		0	C
	Leu	x	0	Ser		0	Term		1	Term		0	A
	Leu	x	5	Ser		3	Term		0	Trp		1	G
C	Leu	x	5	Pro	x	2	His		0	Arg	o	0	U
	Leu	x	24	Pro	x	18	His		6	Arg	o	11	C
	Leu	x	1	Pro	o	1	Gln	x	1	Arg	x	0	A
	Leu	o	19	Pro	o	6	Gln	o	2	Arg	x	12	G
A	Ile	x	2	Thr	o	0	Asn	x	0	Ser		0	U
	Ile	o	5	Thr	o	10	Asn	o	5	Ser		6	C
	Ile	x	2	Thr	x	0	Lys	o	1	Arg	x	0	A
	Met		5	Thr	x	1	Lys	x	14	Arg	x	10	G
G	Val	o	0	Ala	o	1	Asp		0	Gly	o	0	U
	Val	x	18	Ala	x	35	Asp		18	Gly	o	18	C
	Val	o	0	Ala	o	0	Glu	o	6	Gly	x	2	A
	Val	o	20	Ala	o	10	Glu	x	31	Gly	x	24	G

Preliminary dideoxy sequence analysis confirmed that they both contained the first 35 amino acid residues of *T. thermophilus* PGK at one end of the insert, the *Bam*HI site interrupting the coding sequence at residue 35. The G+C-rich DNA sequence around amino acid residue 22 exhibited a high degree of compression on standard polyacrylamide/urea sequencing gels. This suggests a high level of secondary structure in this part of the gene and may explain the weak hybridization signals obtained on Southern blots when using an oligonucleotide probe from this region. This assumption is supported by the finding that the least (4-fold) redundant oligonucleotide probe could not be used as a primer for DNA sequencing even though it contained the correct sequence.

The above restriction analysis indicated the presence of a *Stu*I site about 600 bp upstream of the *Bam*HI site within the PGK coding sequence. This 600 bp *Bam*HI–*Stu*I fragment, containing the PGK *N*-terminal coding region, was used as a probe for Southern-blot analysis of restricted *T. thermophilus* genomic DNA (Fig. 1*b*). These and other results suggested that the entire PGK gene should reside near the centre of a 7 kb *Hin*dIII fragment and that this organism possesses a single PGK gene. Therefore, a second sub-genomic library was constructed using 7 kb size-selected *Hin*dIII-restricted *Thermus* DNA, cloned into the unique *Hin*dIII site of plasmid pAT153. Hybridization *in situ* of 400 of the resulting transformants, using the 600 bp *Bam*HI–*Stu*I fragment as a probe, yielded 13 positive clones. These clones all contained identical inserts as determined by preliminary restriction analysis, and possessed the expected unique *Bam*HI site approximately 2.5 kb from one end of the insert. All subsequent work was carried out on one of these clones, designated pTtPGK-1.

Further restriction analysis of the 7 kb *Hin*dIII insert of pTtPGK-1 located the PGK gene within a 1.62 kb *Sac*I fragment. The gene was mapped in detail and subjected to dideoxy sequence analysis using the strategy depicted in Fig. 1*a*. The entire PGK coding region was sequenced on both DNA strands. Initially, many parts of the polyacrylamide/urea sequencing gels suffered from band compression due to the high G+C content of *Thermus* DNA, but all of these regions were subsequently resolved by increasing the gel temperature during electrophoresis (see the Methods section). There were no unresolved gel reading artifacts due to pile-ups or compression. This is rather unusual in view of the problems experienced by several other workers when sequencing DNA of high G+C content.

The complete DNA sequence of the *T. thermophilus* PGK coding region, together with the immediate 5′- and 3′-flanking regions, is shown in Fig. 2. The first 27 amino acid residues after the deduced initiating methionine were confirmed by limited *N*-terminal protein sequencing (the first residue, preceding the arginine, could not be identified by protein sequencing). Additional peptide-sequence data, determined by gas phase sequence analysis of tryptic peptides (see the Methods section), are indicated in Fig. 2. The open reading frame extends for 1170 bp (excluding the termination codon), thereby encoding a protein of 390 amino acid residues, with a calculated M_r of 41791. This value agrees well with previous estimates of 44000 and 44600 (Nojima *et al.*, 1979). A recognizable Shine–Dalgarno sequence (Shine & Dalgarno, 1974) is present seven nucleotides upstream of the initiating methionine codon. Whilst this represents a possible ribosome-binding site, it should be viewed with some caution since the nucleotide sequence of *T. thermophilus* 16 S rRNA is not known. However, a similar Shine–Dalgarno sequence, homologous to that associated with *E. coli* ribosome-binding sites, is also present at a similar location upstream of the *T. thermophilus* isopropylmalate dehydrogenase coding region (Kagawa *et al.*, 1984).

The DNA sequence immediately preceding the *Thermus* PGK gene contains an open reading frame which, when compared with data in the NBRF protein sequence database, was found to exhibit almost total homology with the *C*-terminal sequence of *Thermus aquaticus* glyceraldehyde-3-phosphate dehydrogenase

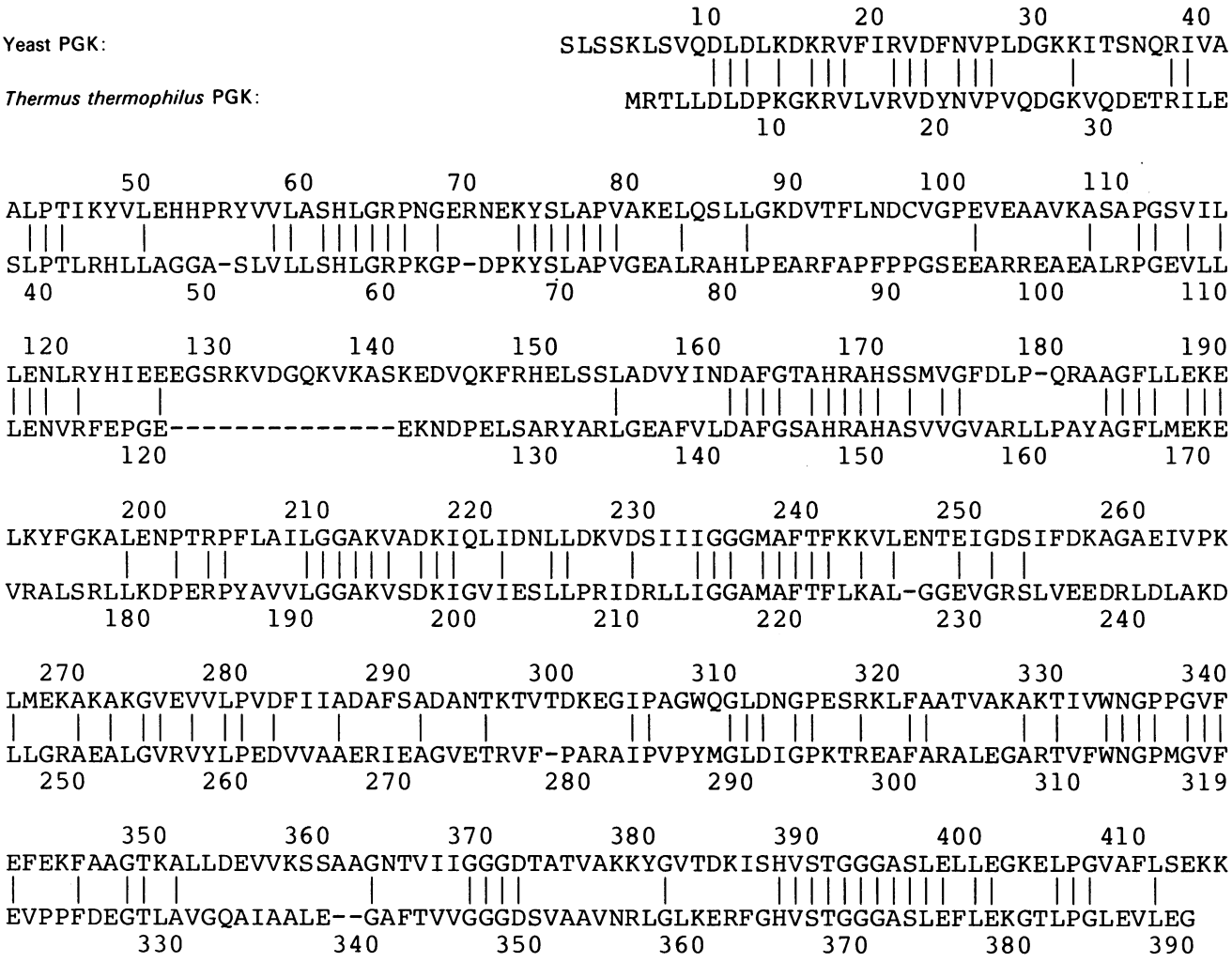

Fig. 3. Comparison of the deduced amino acid sequence of *Thermus thermophilus* PGK with that of yeast

Dashed lines represent gaps introduced to maximize homology. Vertical lines indicate residues conserved in these two species.

(Hocking & Harris, 1980). In fact the termination codon (TAG) of the assumed *T. thermophilus* glyceraldehyde-3-phosphate dehydrogenase gene immediately precedes the initiating methionine codon of the PGK gene (see Fig. 2), without any intervening spacer DNA.

The nucleotide sequence extending 200 bp downstream of the PGK termination codon does not show any distinctive features resembling *E. coli* rho-dependent transcription terminators (Rosenberg & Court, 1979). Such structures have been found associated with genes from the moderate thermophile *Bacillus stearothermophilus* (Hoshino *et al.*, 1985). The 3′-flanking region contains two open reading frames, both of which extend beyond the present limit of sequence analysis. However, in each case the deduced codon usage is not consistent with that observed in the PGK and isopropylmalate dehydrogenase genes from this organism, suggesting that neither of these open reading frames represent part of a functional protein.

Codon usage within the *T. thermophilus* PGK gene

T. thermophilus PGK codon usage differs markedly from that generally observed in *E. coli* (Table 1). The *Thermus* gene shows an extreme bias for G or C in the codon third base position (93.1%) thereby helping to maintain the overall high G+C content of the gene (70.9%). The G+C contents of the first and second base positions were 74.4% and 45.3% respectively. A similar bias for codons ending with G or C (89.4%) was observed in the isopropylmalate dehydrogenase gene from this organism (Kagawa *et al.*, 1984).

Analysis of the deduced amino acid sequence of *T. thermophilus* PGK

Alignment of the deduced amino acid sequence of *T. thermophilus* PGK with that of the yeast enzyme, after the insertion of a small number of gaps to maximize alignment (Fig. 3), indicates an overall homology (identity) of 40.6%. Most of the differences represent conservative amino acid substitutions.

The deduced amino acid composition of *T. thermophilus* PGK agrees well with that obtained from compositional analyses (Nojima *et al.*, 1979) of the purified protein, with the exception of the sulphur-containing amino acids. On the basis of DNA sequence analysis, cysteine is totally absent from *Thermus* PGK, a common property among thermophilic proteins (Mozhaev & Martinek, 1984).

Table 2. Substitution of residues between *Thermus thermophilus* and yeast PGK sequences

Thermus thermophilus and yeast PGK sequences were aligned as shown in Fig. 3. Values along the diagonal indicate the conserved residues. All other values indicate the type of replacement. For example, four of the six histidine residues found in the *Thermus* sequence are conserved in yeast PGK, the other two being replaced by leucine and tyrosine residues.

		Thermus thermophilus																				
		Gly	Ala	Val	Leu	Ile	Phe	Met	Cys	Pro	Tyr	Thr	Ser	Trp	Gln	Asn	Asp	Glu	His	Arg	Lys	Absent	Total
Yeast	Gly	26	2							2			2				1			1	1	2	37
	Ala	3	16	5	3						1		2				2	5		4		2	43
	Val	1	7	12	7	2	2				1	1					1	1		1		2	38
	Leu	1	2	5	24		1	1		2	1								1	2		1	41
	Ile		1	9	6	5	1															1	23
	Phe		1	5	3	1	7			2													19
	Met			1	1			1														1	4
	Cys									1													1
	Pro		1		1			1		12								1			1		17
	Tyr		1		1		2				1		1						1				7
	Thr	1	2	1			2					6	2		1			1			1	1	18
	Ser	1	3		2							2	6				1	5		2		4	26
	Trp										1			1									2
	Gln	1			2			1		1		1								1		1	8
	Asn	1	1		1	1	1						1			3	2	2			1		14
	Asp	2	2							3					2	1	9	5		1		1	26
	Glu	2	3							3		1					1	8		4	3	4	29
	His	2	1															1	4				8
	Arg		1										1							8		3	13
	Lys	3	2		3			1		2						1	1	8		9	8	4	42
	Absent									1													
	Total	44	46	38	54	9	16	5	0	27	7	11	15	1	3	5	18	37	6	33	15		

The net charge on *T. thermophilus* PGK at physiological pH is −7, compared with a value of 0 for the yeast enzyme. The higher acidity of *T. thermophilus* PGK is not associated with an increased content of acidic amino acid residues, but is due to a decrease in the content of basic residues (Table 2). A common property of thermophilic proteins compared with their mesophilic counterparts appears to be an increase in acidity (Crabb *et al.*, 1981).

An increase in the proportion of aliphatic amino acids has previously been observed in thermophilic proteins (Ikai, 1980). The hydrophobic amino acid content of *T. thermophilus* PGK consists of a significantly larger proportion of aliphatic residues (86%) as compared with the yeast enzyme (73%). The aliphatic index (for explanation see Ikai, 1980) of *T. thermophilus* PGK is 12% higher than that of yeast PGK. This difference is significant, especially since glycine is excluded from the analysis, and indicates an appreciable reduction in the number of aromatic amino acid residues as is shown in Table 2.

A preference for the bulkier aliphatic residues has been observed in some heat-stable proteins and has been postulated to enhance thermal stability by increasing the compactness of the folded molecule (Mozhaev & Martinek, 1984; Crabb *et al.*, 1981). No such preference is observed in *T. thermophilus* PGK relative to the yeast enzyme. In fact the thermophilic enzyme contains a higher glycine content than its mesophilic counterpart. It could be argued that this may be a consequence of the high G+C content of *Thermus* DNA, which would favour the increased occurrence of glycine codons (GGN). This argument would not, however, explain why the alanine (GCN) content is not increased in the thermophilic enzyme. Instead it would seem that the high glycine and also the high proline (CCN) content of *T. thermophilus* PGK cannot be explained simply as a consequence of the high G+C content. It is possible that glycine residues take up peptide conformations in the *Thermus* structure which are not allowed for other amino acids. It can be argued that glycine and proline might be located predominantly in the loop regions of the thermophilic, but not the mesophilic, enzyme. If this is the case then a higher energy input would be needed to alter their peptide conformations than for loop regions without such residues.

In comparison with yeast PGK, the *Thermus* protein shows a decrease in the content of isoleucine and an increase in leucine (Table 2). Isoleucine is predominantly substituted by valine and to a lesser extent by leucine. Both these replacements involve a single base change and would be favoured by the high G+C content of the *T. thermophilus* genome.

There is a marked preference for arginine relative to lysine residues in the thermophilic protein. Over half of the arginine residues are either conserved in the two proteins, or replace lysine in the yeast enzyme. This feature of the *Thermus* protein may be imposed by the high G+C content of the gene, although a similar preference for arginine instead of lysine has also been observed in proteins from more moderate thermophiles possessing a less biased base composition (Merkler *et al.*, 1981). Two possible roles for arginine residues in the enhancement of protein thermal stability have previously been proposed. The first of these involves salt bridges, which will be stronger with arginine than lysine (Klibanov, 1983). It is a fact, however, that formation of an ionic bond with an ion in the surrounding medium is energetically more favourable than with a side-chain residue. Secondly, it has been suggested that the guanidinium group of arginine permits better screening of the hydrocarbon part from its thermodynamically unfavourable contact with water. To some extent this has been borne out by chemical modification studies using *O*-methylisourea (see Mozhaev & Martinek, 1984).

Although the number of acidic amino acid residues in *T. thermophilus* and yeast PGK is similar, there is a preference for glutamate in the thermophilic protein, which cannot be explained by the high G+C content of *Thermus* DNA. Approximately 20% of the yeast PGK aspartate residues are replaced by glutamate in the *T. thermophilus* protein whereas the reciprocal substitution occurs only once.

The content of the polar amino acids serine and threonine is considerably lower in the thermophilic as compared with the mesophilic protein. Mozhaev & Martinek (1984) have suggested that these amino acids may be replaced by more hydrophobic residues in the interior of thermophilic proteins. The alignment of the *T. thermophilus* and yeast PGK sequences using the yeast structural information (Watson *et al.*, 1982) would not appear to support this observation. Indeed most of the serine and threonine residues which are substituted by hydrophobic residues are located on the surface of the yeast structure.

The most obvious difference between the amino acid compositions of *T. thermophilus* and yeast PGK is in the number of asparagine and glutamine residues, both of which are considerably reduced in the thermophilic enzyme. Although asparagine might be expected to occur less frequently in DNA with a high G+C content, this alone may not adequately explain the much reduced asparagine content of *Thermus* PGK for the following reasons: (i) despite the high G+C content of *Thermus* DNA, there are more tyrosine (TAY) and considerably more lysine (AAR) residues used than asparagine (AAY); (ii) whilst the conservative replacement of asparagine by glutamine (CAR) would increase the G+C content of this codon, no such substitutions are observed between the *Thermus* and yeast enzymes, indicating a tendency to preclude these amino acids; (iii) asparagine content is not significantly reduced in the proteins of streptomyces, mesophilic organisms possessing DNA with a G+C content similar to that of *T. thermophilus* (72%).

Thus the marked decrease in asparagine and glutamine content in *T. thermophilus* PGK may play a role in enhanced thermal stability. It has been suggested (Ahern *et al.*, 1987) that the susceptibility of asparagine (and to a much lesser extent glutamine) to deamination at elevated temperatures, especially within certain local protein environments, may be a major reason for the observed preclusion of these amino acids from thermophilic proteins.

Whilst some of the differences between the sequences and the amino acid compositions of *T. thermophilus* and yeast PGK may simply be a consequence of the high G+C content of the *Thermus* genome, this alone cannot adequately explain all of the changes. Furthermore, these changes are not only observed in a *Thermus* versus yeast comparison. We have also carried out a number of other

PGK comparisons including *Thermus* versus *Aspergillus nidulans* (Clements & Roberts, 1986), human (Michelson *et al.*, 1983), horse (Hardy *et al.*, 1981; Merrett, 1981), mouse (Mori *et al.*, 1986) and *Trypanosoma brucei* (Osinga *et al.*, 1985), and in all cases our observations with the *Thermus* versus yeast comparison are borne out by these other comparisons. Some of the differences noted in these studies have been observed in other mesophile/thermophile comparisons, and may thus support existing theories about ways in which increased protein thermal stability may be effected. Features such as the marked increase in the glycine and proline content of the thermophilic protein have not been observed in similar comparative studies with proteins from intermediate thermophiles. Work is currently in progress to determine the structure of *T. thermophilus* PGK, which will then be compared in detail with the known yeast structure (Watson *et al.*, 1982). Only when deductions made from these extended studies have been tested by engineering the necessary changes into the mesophilic enzyme will we begin to understand the phenomenon of protein thermal stability.

This work was funded by the SERC/Industry Co-ordinated Programme in Protein Engineering with financial support from the SERC, ICI, Celltech, Glaxo and Sturge.

REFERENCES

Ahern, T. J., Casal, J. I., Petsko, G. A. & Klibanov, A. M. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 675–679
Boyer, H. W. & Roulland-Dussoix, D. (1969) J. Mol. Biol. **41**, 459–472
Clements, J. M. & Roberts, C. F. (1986) Gene **44**, 97–105
Crabb, J. W., Murdock, A. L., Suzuki, T., Hamilton, J. W., McLinden, J. H. & Amelunxen, R. E. (1981) J. Bacteriol. **145**, 503–512
Fujita, S. C., Oshima, T. & Imahori, K. (1976) Eur. J. Biochem. **64**, 57–68
Grunstein, M. & Hogness, D. S. (1975) Proc. Natl. Acad. Sci. U.S.A. **72**, 3961–3965
Hardy, G. W., Darbre, A. & Merret, M. (1981) J. Biol. Chem. **256**, 10284–10292
Hitzeman, R. A., Hagie, F. E., Hayflick, J. S., Chen, C. Y., Seeburg, P. H. & Derynck, R. (1982) Nucleic Acids Res. **10**, 7791–7808
Hocking, J. D. & Harris, J. I. (1980) Eur. J. Biochem. **108**, 567–579
Hoshino, T., Ikeda, T., Tomizuka, N. & Furukawa, K. (1985) Gene **37**, 131–138
Ikai, A. (1980) J. Biochem. (Tokyo) **88**, 1895–1898
Kagawa, Y., Nojima, H., Nukiwa, N., Ishizuka, M., Nakajima, T., Yashuhara, T., Tanaka, T. & Oshima, T. (1984) J. Biol. Chem. **259**, 2956–2960
Klibanov, A. M. (1983) Adv. Appl. Microbiol. **29**, 1–28
Littlechild, J. A., Davies, G. J., Gamblin, S. J. & Watson, H. C. (1987) FEBS Lett. **225**, 123–126
Mas, M. T., Resplandor, Z. E. & Riggs, A. D. (1987) Biochemistry **26**, 5369–5377
Merkler, D. J., Kingfarrington, G. & Wedler, F. C. (1981) Int. J. Pept. Protein Res. **18**, 430–442
Merrett, M. (1981) J. Biol. Chem. **256**, 10293–10305
Messing, J. & Vieira, J. (1982) Gene **19**, 269–276
Michelson, A. M., Markham, A. F. & Orkin, S. H. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 472–476
Mori, N., Singer-Sam, J. & Riggs, A. D. (1986) FEBS Lett. **204**, 313–317
Mozhaev, V. V. & Martinek, K. (1984) Enzyme Microb. Technol. **6**, 50–59
Nojima, H., Oshima, T. & Noda, H. (1979) J. Biochem. (Tokyo) **85**, 1509–1517
Oshima, T. & Imahori, K. (1971) J. Gen. Appl. Microbiol. **17**, 513–517
Osinga, K. A., Swinkels, B. W., Gibson, W. C., Borst, P., Veeneman, G. H., Van Bloom, J. H., Michels, P. A. M. & Opperdoes, F. R. (1985) EMBO J. **4**, 3811–3817
Perkins, R. E., Conroy, S. C., Dunbar, B., Fothergill, L. A., Tuite, M. F., Dobson, M. J., Kingsman, S. M. & Kingsman, A. J. (1983) Biochem. J. **211**, 199–218
Perutz, M. F. & Raidt, H. (1975) Nature (London) **255**, 256–259
Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) J. Mol. Biol. **113**, 237–251
Rosenberg, M. & Court, D. (1979) Annu. Rev. Genet. **13**, 319–353
Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467
Sanger, F., Coulson, A. R., Barrel, B. G., Smith, A. J. H. & Roe, B. A. (1980) J. Mol. Biol. **143**, 161–178
Shine, J. & Dalgarno, L. (1974) Proc. Natl. Acad. Sci. U.S.A. **71**, 1342–1346
Southern, E. M. (1975) J. Mol. Biol. **98**, 503–517
Twigg, A. J. & Sherratt, D. (1980) Nature (London) **283**, 216–218
Walker, J. E., Wonacott, A. J. & Harris, J. I. (1980) Eur. J. Biochem. **108**, 581–586
Wallace, B., Shaffer, J., Murphy, R. F., Bonner, J., Hirose, T. & Itakura, K. (1979) Nucleic Acids Res. **6**, 3543–3557
Watson, H. C., Walker, N. P. C., Shaw, P. J., Bryant, T. N., Wendell, P. L., Fothergill, L. A., Perkins, R. E., Conroy, S. C., Dobson, M. J., Tuite, M. F., Kingsman, A. J. & Kingsman, S. M. (1982) EMBO J. **12**, 1635–1640
Wilson, C. A. B., Hardman, N., Fothergill-Gilmore, L. A., Gamblin, S. J. & Watson, H. C. (1987) Biochem. J. **241**, 609–614

Received 19 October 1987/17 March 1988; accepted 23 May 1988