Conservation of high efficiency promoter sequences in *Saccharomyces cerevisiae*

M.J.Dobson, M.F.Tuite, N.A.Roberts, A.J.Kingsman and S.M.Kingsman
Department of Biochemistry, South Parks Road, Oxford OX1 3QU, UK

R.E.Perkins, S.C.Conroy, B.Dunbar and L.A.Fothergill
Department of Biochemistry, University of Aberdeen, Marischal College, Aberdeen AB9 1AS, UK

Received 2 March 1982; Revised and Accepted 5 April 1982

ABSTRACT

The position of the yeast phosphoglycerate kinase (PGK) gene has been mapped on a 2₀95kb Hind III fragment. We have determined the nucleotide sequence of the 5' flanking region and compared this sequence with those from 16 other yeast genes₀ PGK, like all other yeast genes has an adenine residue at position -3. It has two possible TATA boxes at positions -114 and -152 and a CAAT box at -129. In addition we have defined a structure at position -68 to -39 that is common to all yeast genes that encode an abundant RNA. This structure is a CT-rich block followed, about 10 nucleotides later, by the sequence CAAG.

INTRODUCTION

The glycolytic enzyme genes of Saccharomyces cerevisiae (yeast) encode some of the most abundant mRNA and protein species in the cell with each gene contributing up to 5% of total poly-A mRNA and protein (1). The expression of these unlinked genes is coordinately regulated by the carbon source. When cells are grown on a fermentable carbon source such as glucose enzyme levels are one hundred fold higher than when they are grown on a non-fermentable source (2,3) and there is good evidence that this regulation occurs at the level of transcription (1). A detailed analysis of the 5' regions of these genes should define the structure of a high efficiency eukaryotic promoter and provide insight into the requirements for co-ordinate control₀

Several glycolytic enzyme genes from yeast have been isolated on recombinant DNA molecules by a variety of techniques (4,5,6,7). Two multigene families coding for enolase (ENO) and glyceraldehyde- 3-phosphate dehydrogenase (GAP) have been analysed and compared in some detail (8) and show considerable

0305-1048/82/1008-2625$ 2.00/0

homology in their 5' regions. In this paper we describe the localisation of the sole phosphoglycerate kinase (*PGK*) gene in the yeast genome on a cloned restriction fragment, and we present a strategy for creating deletions ending in the 5' region of this gene. The deletions are exploited to determine the nucleotide sequence of the 5' region and this sequence is compared with those of other eukaryotic genes.

MATERIALS AND METHODS

Bacterial and yeast strains:

E.coli strain AKEC28=C6000 *thrC* *leuB6* *thyA* *trpC1117* *hsdRk* *hsdMk*. *Saccharomyces* *cerevisiae* strain LL20=*leu2-3* *leu2-112* *his3-11* *his3-15*. (30)

Amino acid sequence analysis:

PGK was cleaved with CNBr, and the resulting four fragments were separated by gel filtration (10). Fragment CN2 corresponds to residues 271-419 (the C-terminus), and we further cleaved at its two tryptophan residues with a heptofluorobutyric acid and CNBr mixture (11), or with proteolytic enzymes. Sequencing was done manually by the dansyl-Edman method or automatically with a Beckman 890C liquid-phase sequencer as described previously (10). The dansyl-amino acids were identified by two-dimensional thin-layer chromatography, and the phenylthiohydantoin derivatives of the amino acid residues from the sequencer were identified quantitatively by reversed-phase high-pressure liquid chromatography.

Enzymes, electrophoresis and fragment purification:

Restriction endonucleases were purchased from Bethesda Research Laboratories (BRL). Restriction fragments smaller than 0.5kb were sized on 5% polyacrylamide gels as previously described (12). Fragments were purified from agarose by the method of Tabak and Flavell (13).

RNA isolation and transcript mapping:

Yeast RNA was isolated from 2 x 10^{10} yeast cells (LL20), harvested at 2 x 10^{7} cells per ml., by vortexing with glass beads (40 mesh) in 1 ml. of LETS buffer (0.1M Tris-HCl pH7.5, 0.1M LiCl, 1mM EDTA) for 2 minutes. A further 3 ml. of LETS buffer was added, the extract was made 0.1% in SDS and 0.1% in diethyl pyrocarbonate and extracted at room temperature 4-6 times with an equal volume of phenol:chloroform (50:50) equilibrated against LETS buffer. After ethanol precipitation from 0.2M sodium acetate the pellet was resuspended in 0.1M NaCl, 1mM $MgCl_2$ and treated with 5μg/ml. deoxyribonuclease I (Worthington) which had been treated with agarose: 5'-(p-aminophenyl phosphoryl)-uridine-2'(3')-phosphate (Miles-Yeda Ltd) to remove ribonuclease activity. SDS was added to 0.5% and EDTA to 25mM and the solution treated with 200μg/ml. protease K (Merck) for 30 minutes at 37°C and then extracted 3-4 times with phenol:chloroform (50:50) equilibrated against 0.1M NaCl, 25mM EDTA. Total yeast RNA was precipitated with ethanol from 0.2 M sodium acetate, washed twice with 3 M sodium acetate and twice in ethanol:0.4 M sodium acetate (2.5:1), air dried and stored at -70°C.

Protection of purified restriction fragments against S1 nuclease digestion was performed according to the procedures of Berk and Sharp (14).

Southern transfers, hybridisation and in vitro labelling:

DNA fragments fractionated on agarose gels were transferred to nitrocellulose sheets by a modification of Southern's procedure (15) as previously described (16). Restriction fragments were labelled using (^{32}p)-TTP (Amersham) by nick translation (17). Hybridisation was carried out in 0.3 M NaCl, 0.03 M sodium citrate, 0.02% PVP, 0.02% BSA, 0.02% ficoll at 65°C for 48 hours.

Delection formation and linker ligations:

Plasmids were cleaved with an appropriate restriction enzyme and then digested with BAL 31 nuclease (BRL) at a concentration of 0.05 units/0.05μg DNA/ml in BAL buffer (20 mM Tris-HCl pH8.1, 600 M NaCl, 12 mM $CaCl_2$ 1mM EDTA) at 15°C. Under these conditions 90bp are removed per end per minute.

Bam HI linkers were purchased from Collaborative Research Ltd and phosphorylated according to the manufacturers recommendations. Blunt end ligations were carried out at 20°C for 6 hours in 20mM Tris-HCl, 7.5 mM MgCl, 0.1 mM EDTA, 1 mM ATP, 1 mM DTT, 1 mM spermidine with 400 units of T4 DNA ligase (New England Biolabs).

DNA sequencing:

The dideoxy-chain termination method of Sanger et al (20) was used. Eco RI - Bam HI fragments were subcloned into phage M13mp701. This phage was constructed by David Bentley (Department of Pathology, University of Oxford) and contains unique Bam HI and Eco RI sites with the Bam HI site closest to the 'universal' primer sequence (18) i.e. chain elongation is from the Bam HI site towards the Eco RI site. All reagents were purchased from BRL.

RESULTS

The yeast PGK gene exists on a 2.9kb Hind III fragment in a a yeast-E.coli vector, pMA3, which comprises plasmid pBR322 with a 3.3kb double Eco RI fragment containing the yeast *LEU2* gene and the 2μplasmid origin of replication (M.J. Dobson, S.M. Kingsman and A.J. Kingsman unpublished data). A partial restriction map of this molecule is shown in figure 1a. The *PGK* Hind III fragment was isolated from a Hind III fragment collection inserted into λ762 (22) by plaque hybridisation with P-labelled cDNA prepared from yeast poly-A RNA. The fragment is identical to the '3.1kb' fragment described by Hitzeman *et al.*(21) in plasmid pB1 and in hybrid selection translation experiments the fragment was shown to encode a protein of identical mobility to pure PGK in SDS-PAGE (M.F. Tuite and S.M. Kingsman unpublished data). A restriction map of the 2.95kb fragment is shown in figure 1b.

Amino acid sequence

The amino acid sequence of residues 270-400 of yeast PGK is shown in figure 2b. The sequence was determined by manual and automated Edman degradation. The amino acid sequence data allowed us to match restriction sites on the 2.95kb Hind III fragment with groups of two or three amino acids in the protein sequence. Figure 2a shows the relevant restriction sites and those sites are marked on the amino acid sequence in figure 2b. The positions of the four sites on the restriction map and the protein sequence are congruent allowing us to orientate the gene with respect to the sites on the 2.95kb Hind III fragment. Given that the molecular weight of PGK is 40Kd (419 amino acid residues) and assuming that there are no large introns we can also predict the positions of the 5' and

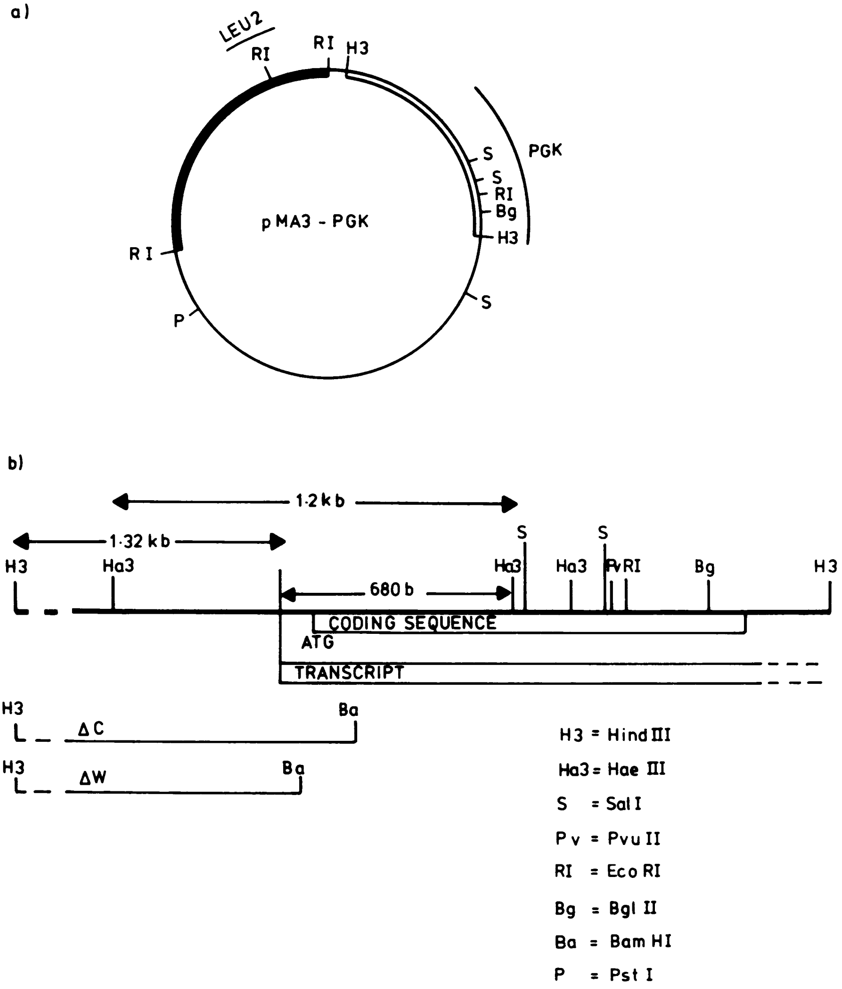

Figure 1. a) Partial restriction map of pMA3-PGK. b) Detailed map of the 2.95kb Hind III fragment from pMA3-PGK. See text for explanation.

3' ends of the coding sequence. The extent of the coding sequence, assuming colinearity, is shown in figure 1b, the initiation codon is about 920 nucleotides to the left of the Eco RI site and the termination codon about 330 nucleotides to the right.

Mapping the 5' transcript terminus:

The position of the 5' end of the PGK transcript was located

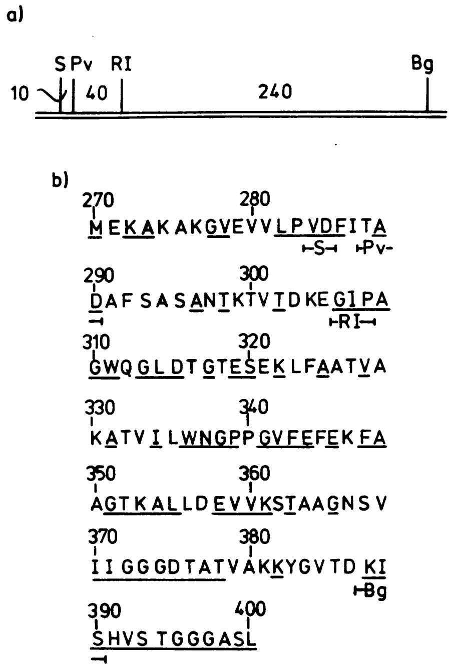

Figure 2. a) Restriction enzyme sites around the Eco RI site in the 2.95kb Hind III PGK fragment. b) Amino acid sequence of PGK from position 270 to 400.The positions of the restriction enzyme sites in a) are marked on the amino acid sequence.The residues which are identical in human,horse and yeast PGK are underlined.

by the S1 protection method (14). The 1.2kb Hae III fragment spanning the 5' end of the coding sequence (figure 1b) was purified from an agarose gel and hybridised to total yeast RNA. The hybrids were treated with various concentrations of S1 nuclease and the products were analysed on a 1.5% agarose gel by Southern hybridisation using the 2.95kb Hind III fragment as probe. Figure 3 shows that the size of the single protected fragment was 680b. On the basis of our previous mapping data this would place the 5' end of the PGK transcript about 960bp to the left of the Eco RI site on the 2.95kb Hind III fragment. This agrees well with our estimate of the position of the initiation codon and suggests that if there are any introns between the 5' end of the transcript and the Bgl II site then they are very small.

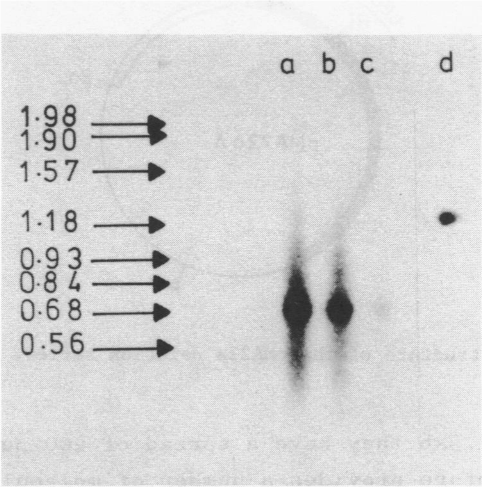

Figure 3. S1 protection of the 1.2kb Hae III fragment from pMA3-PGK. Lane a) 25 units S1, b) 50 units S1, c) 100 units S1, d)1.2kb Hae III fragment untreated.

Deletion construction and sequencing:

The 5' 'control' region of the PGK gene is in a region that contains very few convenient restriction sites, making the design of a sequencing strategy relatively difficult. We adopted a procedure to solve this problem that may be of general use. Plasmid pMA3-PGK was digested with Sal I (figure 1) and then with exonuclease BAL 31 to remove about 500bp from each end. This resulted in the loss of the two small Sal I fragments and the creation of a series of deletions starting at the leftmost Sal I site in the PGK sequence and the Sal I site in pBR322 and ending around the initiation codon in PGK and nucleotide 1150 in pBR322 respectively. These deleted molecules were then ligated in the presence of a 50-fold molar excess of Bam H1 linkers and then used to transform AKEC28 to Leu^+, Amp^r. The general structure of these molecules, designated the pMA22a deletion series is shown in figure 4. Seventy of these deleted molecules have been analysed by measuring the length of the Eco RI - Bam HI fragments containing the 5' region of the PGK gene. While they show a

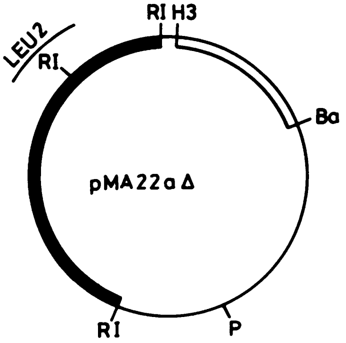

Figure 4. General structure of the pMA22a deletion series.

mean length of 1.5kb they have a spread of 200 nucleotides. This collection therefore provides a number of molecules that are useful for the sequence analysis of the 5' region of the PGK gene. Two such deletions, C and W are shown in figure 1b. The small Eco RI - Bam HI fragments from these molecules were purified and cloned in M13mp701 and sequenced by the dideoxy-chain termination method (20) starting in each case at the Bam HI site and elongating towards the Eco RI site. The nucleotide sequence of 226 nucleotides upstream from the initiation codon and the first seven codons are shown in figure 5. The sequence was confirmed by sequencing four other deletions with overlapping end-points (data not shown).

The amino acid sequence of PGK is highly conserved and shows considerable homology between the human, horse and yeast proteins (23, L. Fothergill unpublished data). The amino acids underlined in figure 5 match exactly those of the human and horse sequence. The sequence of the entire coding region will be the subject of a separate paper.

DISCUSSION

We have mapped the position of the yeast PGK gene on the 2.95kb Hind III restriction fragment (figure 1) and shown that, as is the case for most yeast genes, there are either no introns or if there are any they are small and close to the transcription termini. There are several interesting features at the 5' end of

```
-226
 *
 AGCCTGCTCT CACACATCTT TCTTCTAACC AAGGGGTGTT TAGTTTAGTA


-176
 *
 GAACCTCGTG AAACTTACAT TTACATATAT ATAAACTTGC ATAAATTGGT


-126
 *
 CAATGCAAGA AATACATATT TGTCTTTTCT AATTCGTAGT TTTTCAAGTT


-76
 *
 CTTAGATGCT TTCTTTTTCT CTTTTTTACA GATCATCAAG AAGTAATTAT


-26
 *                      -1+1
 CTACTTTTTA CAACAAATAT AAAACA ATG TCT TTA TCT TCA AAG TTG
                              MET SER LEU SER SER LYS LEU


              Total sequence   247  nucleotides
```

Figure 5. Nucleotide sequence of the 5′ region of the PGK gene. Underlined sequences are discussed in the text. The open box marks the approximate position of the 5′ end of the transcript.

the gene, and these are summarised and compared with sequences from 16 other yeast genes in figure 6.

Kozak (33) has suggested that the sequence environment of the translation initiation codon modulates the efficiency with which the 40S ribosomal subunit binds to the 5' end of a message. The essential features of the preferred eukaryotic initiation region are a purine at -3 and a purine, usually G, at +4 (33). In yeast this is only partially true. PGK has an adenine residue at -3, so too have all the other yeast genes examined, although in the in vitro fusion of the ADH 1 5' control region to a human LeIFD gene (24) there was a uracil residue at -3 and efficient translation of the human gene was observed. Eleven of the seventeen genes compared in figure 6 have a purine at +4. PGK has a pyrimidine yet is an abundant protein suggesting that efficient translation does not depend on a purine at this

	ATG ENVIRONMENT	CACACA	TATA	GGPyAATCT	CT Block	CT-CAAG
	-10 -1+1					
PGK	ATATAAAACA.ATG.TCT.TTA.	-	-114 -152	-129	-49.-68	10
ENO8	TACAATAATA.ATG.GCT.GTC.	-12	?	?	-40.-77	8
ENO46	CTAAATCAAA.ATG.GCT.GTC.	-20	-146	?	-47.-110	10
GAP63	ACAAAACAAA.ATG.GTT.AGA.	-18	?	?		
GAP491	AATAAACAAA.ATG.GTT.AGA.	-18	-143	?	-48.-64	8
ADH1	CTCATATACA.ATG.TCT.ATC.	-25	?	?	-50.-83	12
CYC-1	TAAATTAATA.ATG.ACT.GAA.	-16	-124 -177	?	-93.-113	-
CYC-2	TAAACAAAAC.ATG.GCT.AAA.	-	-167		-110.-120	30
ACTIN	TGAATTAACA.ATG.GAT.TCT.	-	-200	-224	-27.-40 -130.-149	9
H2B-1	CACACATACA.ATG.TCT.GCT.	-5	-142	-192	-	
H2B-2	TCTACAAATA.ATG.TCC.TCT.	-	-141 -153		-	
MATa1	GAAGGACAAC.ATG.GAT.GAT.	-	-70	-	?	
MAT 1	CATTCACAAT.ATG.TTT.ACT.	-	-200	-	?	
MAT 2	GCAAGAAAAA.ATG.AAT.AAA.	-	-70	-	-	
TRP1	AGCTTGGAGT.ATG.TCT.GTT.	-21	-227	-	-120.-133	47
HIS3	GAAGGCAAAG.ATG.ACA.GAG.	-	-70	?	-	
HIS4	TTTCTGAATA.ATG.GTT.TTG.	-	-70 -120	?	-	

Figure 6. Comparison of the 5′ regions on 17 yeast genes. Columns 3,4,5 and 6 give the positions of sequences discussed in the text. Column 7 gives the distance from the CT block to the CAAG sequence (see text). - = sequence absent; ? = sequence may be present but with considerable variation from the canonical sequence. A gap in the figure means that the data are not available. Data are from the folowing: ENO8,ENO46,GAP63,GAP491 (8); ADH1 (24); ISO1 (25); ISO2 (26); ACTIN (27); H2B-1,H2B-2 (28); MATa1, 1, 2 (29); TRP1,nucleotides +6 to -103 (30),nucleotides -103 to 250 (M.F.Tuite, S.M.Kingsman and A.J.Kingsman, unpublished data); HIS3 (31); HIS4 (32) .

position. Perhaps more striking is the presence of a pyrimidine, usually U, at position +6. PGK and the abundant glycolytic genes obey this 'rule'.

The hexanucleotide CACACA (or CAPyACA) has been found close to the initiation codon in several (8/17) yeast genes (figure 6). There is however no obvious correlation between the abundance of a gene product and the presence of this sequence. PGK lacks such a sequence.

At least two sequences are thought to be important for transcription initiation in eukaryotes. The first of these is an A/T rich region with the canonical sequence TATAT/AAT/A (34) (the TATA Box) which is usually located 25-32bp upstream from the transcription initiation site. The second is the CAAT box with the canonical sequence GCC/TCAATCT (35) which has been found in several eukaryotic genes about 80bp upstream from the site of initiation of RNA synthesis. While the TATA box has been found in almost all yeast genes examined the CAAT box is generally either clearly absent or partially disguised (figure 6). PGK has two possible TATA boxes, the first starting at -152, TATATAAA (figure 5) and the second, an imperfect example, starting at -114, TACATA (figure 5). There is also an almost perfect (7/9) CAAT box at -29 between the two TATA boxes. Given the position of these sequences relative to the transcription initiation site (34) it seems likely that the TATA box at -114 is the relevant site.

Holland and colleagues (8) showed that the ENO and GAP genes shared common features at their 5' ends and speculated that these may be instrumental in the coordinate control of glycolysis (8). However from our comparison of PGK with all the other yeast genes for which there is data we would like to suggest that the sequences common to the 5' regions of the glycolytic enzyme genes are, in fact, not specific for the glycolytic enzymes but rather these structures are common to genes encoding relatively abundant products. In PGK this structure comprises a pyrimidine rich block (CT block) on the coding strand at position -49 to -68 followed, 10 nucleotides later, by the sequence CAAG (figure 5). We have also noted that the relative positions of these sequences where they occur are highly conserved and correlate with the highly expressed genes for the glycolytic enzymes, ADH and actin (figure 6). The exceptions are the iso-2-cytochrome C and TRP1 genes which are not expressed at high levels and which do not conform to the 'normal' spacing of the CT-block and CAAG sequence. The CAAG sequence is part of a 20 nucleotide sequence that is conserved to a slightly lesser extent and which contains the most significant homology between the ENO and GAP genes (8). This 20 nucleotide sequence starts about 7 nucleotides before the CAAG and ends

```
     8-10bp
CT------------AAAAAACCAAG-AACTAAG-CT      CONCENSUS  ENO/GAP
               T   C          T T T
              * * ** **** ** ****  *
      10bp
TT------------AGATCATCAAG-AAGTAATTAT         PGK
```

Figure 7. Comparison of the CAAG region of glycolytic enzyme genes. The top sequence is a concensus from reference 8 with alternatives shown. The lower sequence is from PGK. A * marks homologous positions. A - signifies any nucleotide.

about 10 nucleotides after it. There is 75% homology amongst the glycolytic enzyme genes in this region; about 50% homology between ADH and actin. Figure 7 shows a comparison of the PGK sequence with a concensus sequence around the CAAG site which we have derived from the ENO and GAP sequences.

In summary, the 5' region of the yeast PGK gene has many of the features associated with other eukaryotic genes. Within the 150-250 nucleotides upstream from the translational initiation codon that have been examined in five glycolytic enzyme genes there is no sequence that is an obvious candidate for mediator of coordinate control. However the CT-block----CAAG structure may be a requirement for high efficiency gene expression in yeast.

REFERENCES

1. Holland,M.J. and Holland,J.P.(1978) Biochemistry 17,4900-4907.
2. Hommes,F.A.(1966) Arch.Biochem.Biophys. 114, 231-233.
3. Maitra,P.K. and Lobo,Z.(1971) J.Biol.Chem. 246,475-488.
4. Hitzeman,R.H.,Chinault,A.C.,Kingsman,A.J. and Carbon,J.A. (1979) in ICN-UCLA Symposium on Molecular and Cellular Biology,Maniatis,T.and Fox,C.F.Eds.,Vol 14 ,pp.57-68,Academic Press ,New York.
5. Holland,M.J.and Holland,J.P. (1979) J.Biol.Chem. 254,5466-5474.
6. Holland,J.P.and Holland,M.J. (1979) J.Biol.Chem. 254,9839-9845.
7. Holland,J.P.and Holland,M.J. (1980) J.Biol.Chem. 255,2596-2605.
8. Holland,M.J.,Holland,J.P.,Thill,G.P. and Jackson,K.A. (1981) J.Biol.Chem. 256,1385-1395.
9. Maniatis,T.,Jeffrey,A. and van deSande,H. (1975) Biochemistry 14,3787-3794.
10. Fothergill,L.A.and Harkins,R.N. (1982) Proc.R.Soc.Lond.B. 215,in press.
11. Ozols,J,Gerard,C.and Stachelek,C. (1977) J.Biol.Chem. 252,5986-5989.
12. Kingsman,A.J.,Gimlich,R.L., Clarke,L., Chinault,A.C. and Carbon (1981) J.Mol.Biol. 145,619-632.
13. Tabak,H.F.and Flavell,R.A. (1978) Nucl.Acids.Res. 5,2321-2332.
14. Berk,A.J. and Sharp,P.A. (1978) Proc.Natl.Acad.Sci.USA 75,1274-1278.
15. Southern,E.M. (1975) J.Mol.Biol. 98,503-517.

16. Chinault,A.C. and Carbon,J. (1979) Gene 5,111-126.
17. Rigby,P.W.J., Dieckman,M., Rhodes,C. and Berg,P. (1977) J.Mol.Biol. 113,237-251.
18. Anderson,S., Gait,M.J., Mayol,L. and Young,I.G. (1980) Nucl.Acids.Res. 8,1731-1743.
19. Maxwell,I.H., Maxwell,F. and Hahn,W.E. (1977) Nucl.Acids.Res. 4,241-246.
20. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) Proc.Natl.Acad.Sci.USA. 74,5463-5467.
21. Hitzeman,R.A., Clarke,L. and Carbon,J. (1980) J.Biol.Chem. 255,12073-12080.
22. Murray,N.E., Brammar,W.J. and Murray,K. (1977) Molec.Gen.Genet. 150,53-61.
23. Banks,R.D., Blake,C.C.F., Evans,P.R., Haser,R., Rice,D.W., Hardy,G.W., Merrett,M. and Phillips,A.W. (1979) Nature 279,773-777.
24. Hitzeman,R.A., Hagie,F.E., Levine,H.L., Goeddel,D.V., Ammerer,G. and Hall,B.D. (1981) Nature 293,717-722.
25. Faye,G., Leung,D.W., Tatchell,K., Hall,B.D. and Smith,M. (1981) Proc.Natl.Acad.Sci.USA. 78,2258-2262.
26. Montgomery,D.L., Leung,D.W., Smith,M., Shalit,P., Faye,G. and Hall,B.D. (1980) Proc.Natl.Acad.Sci.USA. 77,541-545.
27. Gallwitz,D., Perrin,F. and Seidel,R. (1981) Nucl.Acids.Res. 9,6339-6350.
28. Wallis,J.W., Hereford,L. and Grundstein,M. (1980) Cell. 22,799-805.
29. Astell,C.R., Ahlstrom-Jonasson,L., Smith,M., Tatchell,K., Nasmyth,K.A. and Hall,B.D. (1981) Cell. 27,15-24.
30. Tschumper,G. and Carbon,J. (1980) Gene 10,157-166.
31. Struhl,K. (1981) Proc.Natl.Acad.Sci.USA. 78,4461-4465.
32. Farabaugh,P.J. and Fink,G.R. (1980) Nature 286,352-356.
33. Kozak,M. (1981) Nucl.Acids.Res. 9,5233-5252.
34. Gannon,F., O'Hare,K., Perrin,F., LePennec,J.P., Benoist,C., Cochet,M., Breathnach,R., Royal,A., Garapin,A. and Chambon,P. (1979) Nature 278,428-434.
35. Benoist,C., O'Hare,K., Breathnach,R. and Chambon,P. (1980) Nucl.Acids.Res. 8,127-142.