Volume 14 Number 2 1986 Nucleic Acids Research

Identification, cloning and sequence determination of the genes specifying hexokinase A and B from yeast

Conrad Stachelek, Janet Stachelek, Judith Swan+, David Botstein+ and William Konigsberg

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06510 and +Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 14 November 1985; Accepted 13 December 1985

ABSTRACT

The hexokinase A (HKA) and hexokinase B (HKB) genes of Saccharomyces cerevisiae have been cloned from a library of yeast genomic DNA. Using an *in vitro* glucose phosphorylation assay, the HKB gene was located on a plasmid carrying a 13.6 kb fragment of yeast DNA. After subcloning the relevant restriction fragments, the nucleotide sequence of the HKB gene was determined. Using this information, we were able to locate the HKA gene on a plasmid carrying this gene, which we then sequenced. Approximately 43% of the amino acid sequence of HKB was determined directly from 24 tryptic peptides. The results are in complete agreement with those derived from the DNA sequence and are consistent with the results of x-ray crystallography. Comparison of the amino acid sequences of HKA and HKB show that 378 out of 485 residues are identical. The 5′ flanking region of the A gene contains nucleotide sequences expected for genes that are expressed at relatively high levels in yeast. The 24 base pair hyphenated palindrome at the 3′ end of the HKB gene may be a site for termination of transcription of this gene.

INTRODUCTION

Yeast hexokinase is an allosteric enzyme which catalyzes the first step of several metabolic pathways, forming a hexose-1-phosphate from a hexose and ATP-Mg. It exists as two isozymes, arbitrarily designated as hexokinase A and B. The structure of several crystal forms of hexokinase has been determined by x-ray crystallography (1-6). These structures represent binary or ternary complexes of hexokinase and substrates or substrate analogues at various steps in the reaction pathway of this enzyme. These complexes provide the opportunity of analyzing its catalytic mechanism in exquisite detail.

The HKB structure has been refined by Anderson (5) to 2.5 angstrom resolution. The structure of the HKA isozyme, crystallized as a complex with glucose, has been determined at 3.0 angstrom resolution (6).

In order to build models of these isozymes at atomic resolution, the complete amino acid sequences of HKA and HKB are required. We have determined the nucleotide sequences of the HKA and HKB structural genes which has allowed us to deduce the amino acid sequences of both forms of these

enzymes. In addition, we have obtained amino acid sequence information on a number of tryptic peptides from HKB which are in complete agreement with those derived from the DNA sequence. We have compared the primary structures of HKA and HKB and find that 78% of the residues are identical. The DNA sequence determination of regions flanking the HKA and HKB structural genes provides information concerning transcription initiation and termination signals, as well as data on the sequence context surrounding the translation initiation codon that may help in understanding the high efficiency of translation of the mRNA for these genes. The results of the cloning, DNA sequencing, and the protein chemistry are reported here.

MATERIALS AND METHODS

Strains of S.cerevisiae and E. coli

S. cerevisiae strain DBY1175 (*MAT adel trpl ura3-52 hxk1-1 hxk2-2 can^r*) was employed for the isolation of clones carrying the HKA and HKB genes from a library of yeast genomic DNA. This strain was constructed by standard methods by crossing a strain carrying both *HXK* mutations (F452, originally described by Maitra and Lobo (7)) with a strain carrying the *ura3* mutation (DBY747). *E. coli* strain DB6507 (*hsdS20 [r_B^- m_B^-] recA13 ara-14 proA2 lacY1 galK2 rpsL20 str^r*) was used for transformation and amplification of selected recombinant clones. *E. coli* strain ZSC113 (*lacZ82 or lacZ827 ptsM12 ptsG22 glk-7 rha-4 rpsl223 relA1*) (8), supplied by B. Bachman from the *E. coli* Genetic Stock Center at Yale University, was used for the *in vitro* assay of glucose phosphorylation.

Construction and Screening of Yeast Genomic Library

The clones carrying the HKA and HKB genes were isolated from a YEp24 random-insert genomic library by complementation of the fructose-utilization defect of yeast strain DBY1175. The construction and screening of the library followed the procedure described in detail by Carlson and Botstein (9). The strain DBY1175 was transformed with DNA from the library, selecting first the Ura$^+$ phenotype, pooling batches of transformants, and then replating to select those cells that were able to grow anaerobically using fructose as carbon source on a minimal medium. The frequency of fructose utilizers amoung total transformants was 0.2 to 0.3 percent. The co-segregation of the fructose and Ura$^+$ phenotypes was tested for each independent clone. The DNA from yeast transformants was introduced into *E. coli* strain DB6507 by transformation and selected for ampicillin resistance. Plasmid DNA was isolated and appropriate clones were selected for re-transformation into DBY1175 after

restriction mapping. All transformants selected for the ability to grow on fructose became Ura$^+$ and all those selected for a Ura$^+$ phenotype were able to compensate for the fructose utilization defect of DBY1175.

Identification of HKA and HKB Genes

Identification of the cloned genes with hexokinases A and B was done by the method described by Gancedo et al. (10), which uses hydroxyapatite chromatography to separate the two isozymes. The two hexokinase isozymes can be further distinguished by the ratio of activity with fructose and glucose as substrates. Additional evidence for the assignment of the plasmid clones to the loci specifying hexokinases A and B was obtained by genetic mapping. The method of 2-micron mapping (11) was applied to one plasmid specifying the HKA gene. The location of HKB was determined by linkage studies relative to the ade5 gene on chromosome VII (12).

Location of Structural Genes Within the Plasmid Clones

The structural gene for HKB was located by an in vitro assay for glucose phosphorylating activity in extracts of E. coli strain ZSC113 transformed with plasmid vectors carrying subclones of the original plasmid (pRB62). After transformation of ZSC113 with plasmid DNA, antibiotic resistant colonies were grown in 50 ml of luria medium for preparation of the cell free extracts. Cells were resuspended in 50 mM tris-HCl, pH 7.5, 1 mM EDTA, 1 mM DTT, 1 mM phenyl methyl sulfonyl chloride (PMSF) and lysed by the addition of solid lysozyme to a final concentration of 1 mg/ml and Brij-58 to 0.5%. After a 30 min. incubation at 37°C, DNA and RNA were digested by the addition of DNAse I to a concentration of 10 microgram/ml and RNAse A to a concentration of 10 microgram/ml and incubated for 30 min at 37°C. The lysate was centrifuged for two hours at 45,000 rpm to remove particulate debris. The supernatant was stored at -70°C. Assay mixtures (50 microliters) contained 6 mM $MgSO_4$, 5 mM ATP, 5 mM glucose and 10^6 cpm of D-[U-^{14}C] glucose (Amersham) and 25 microliters of the ZSC113 cell free extract. The reaction mixture was incubated at 37°C for 30 min. The reaction was terminated by the addition of 100% trichloroacetic acid to a final concentration of 5%. The mixture was centrifuged to remove protein, and then extracted with ether. The supernatant was chromatographed on polyethyleneimine (PEI) plates to separate glucose from glucose-6-phosphate using a solvent of 0.5M formic acid and 0.5M lithium chloride.

After the completion of chromatography, each lane was cut into 1 cm portions, inserted into scintillation vials and covered with scintillation fluid for counting. The amount of ^{14}C glucose converted to ^{14}C glucose-1-phosphate was

compared to control reaction mixtures made from recombinant strains carrying pRB62 and pRB5 (vector without yeast insert).

The location of the HKA gene was determined by DNA sequence analysis of selected restriction fragments of plasmid pRB141; the amino acid sequences specified by these DNA sequences were then compared to the amino acid sequence specified by the HKB structural gene to locate the homologous HKA gene.

DNA Sequencing

DNA sequences were determined by the dideoxynucleotide chain termination method of Sanger (13). Template DNA was prepared from clones propagated in M13 phage derivatives (14).

Peptide Isolation and Amino Acid Sequencing

Hexokinase protein was obtained from Worthington Diagnostics (code HKP II). This preparation was shown to consist of nearly pure hexokinase B by non-denaturing polyacryamide gel electrophoresis (15). This protein was performic acid oxidized by standard methods (16). A tryptic digest of 500 nmol of this derivative was initially separated by cation exchange chromatography on Aminex AG-50W-X4 (BioRad)(17). Peptides were detected by a modified version of the alkaline ninhydrin reaction of Hirs (18). Those peptides which were sufficiently pure after acid hydrolysis and amino acid analysis were chemically sequenced using a modified version of the classic Edman degradation (19).

Mixtures of peptides were further purified by anion exchange chromatography on Dowex AG-1-X2 resin (20). Those peptides which could be purified to homogeneity by this additional step were then chemically sequenced.

RESULTS

Isolation of Clones Carrying the HKA and HKB Genes

Clones derived from a random-insert library of yeast genomic DNA were selected for their Ura^+ phenotype and their ability to grow anaerobically using fructose as carbon source after being transformed into yeast strain DBY1175. Plasmid DNA was isolated and restriction maps were made which allowed the division of the clones into two separate classes: of 24 independent clones analyzed, 16 were in one class and eight in the other. The identity of the isozyme carried by each class was determined by hydroxyapatite chromatography and by the ratio of activity using glucose and fructose as substrates. When the enzyme activity produced by a class I plasmid (pRB141) in DBY1175 was fractionated, a single major peak eluted at a salt concentration of 100 mM sodium phosphate which had a fructose/glucose activity ratio of 1.9,

characteristic of hexokinase A. When the activity produced by a class II plasmid (pRB62) in DBY1175 was fractionated, a single peak eluted at a salt concentration of 70 mM sodium phosphate which had a fructose/glucose activity ratio of 1.1, characteristic of hexokinase B. These experiments sufficed to show that each of the two genes represented by these classes of plasmid clones specifies one of the hexokinase isozymes.

Additional evidence for the assignment of the plasmid clones to the loci specifying hexokinase A and B was obtained by genetic mapping studies. The method of 2-micron mapping was applied to pRB141, which was found to integrate specifically into chromosome VI, where the HXK1 gene (specifying hexokinase A) is located (12). The location of integration of the other gene was found by a more classical method, since close linkage of the HXK2 gene (specifying hexokinase B) to ade5 on chromosome VII has been reported (12). A subclone of pRB62 in the integrating vector YIp5 was used to transform DBY1034; integration by homology was checked by Southern blot experiments. When such integrants were crossed to strains carrying ade5 and ura3, close linkage was observed between the integrated plasmid (carrying URA3$^+$) and ADE5$^+$, indicating an intrachromosomal distance of about 3cM.

Genomic Southern blot experiments confirmed that the homology between the two hexokinase genes is limited to the two regions cloned, suggesting that there are only two hexokinase genes in yeast, and that the gene for glucokinase is less homologous at the DNA level to the hexokinases than the latter are to each other.

Localization and Subcloning of the HKB Gene.

The plasmid pRB62, containing the structural gene for HKB, was mapped using nine restriction enzymes to give the results shown in Fig. 1. The length of the plasmid was calculated to be 22.5kb based on summing the sizes of all fragments obtained with any given restriction endonuclease. The size of the yeast chromosomal DNA insert was found to be approximately 13.5kb kb. To localize the HKB gene within the insert, pRB62 was cut with Sal I, which divided the 13.5kb yeast DNA insert into two regions.

Religation of the diluted Sal I digest gave a deletion derivative of pRB62 (ΔSal I). When this plasmid was transformed into E. coli strain ZSC113, the resulting transformant showed the same glucose phosphorylating activity as strain ZSC113 harboring the pRB5 vector which lacked the yeast DNA insert. In contrast, when the BamHI-SalI half of the 13.5kb fragment was cloned into pBR322 to give pBR322(SalII) and used to transform ZSC113, the resulting strain had almost the same level of glucose phophorylating activity as ZSC113

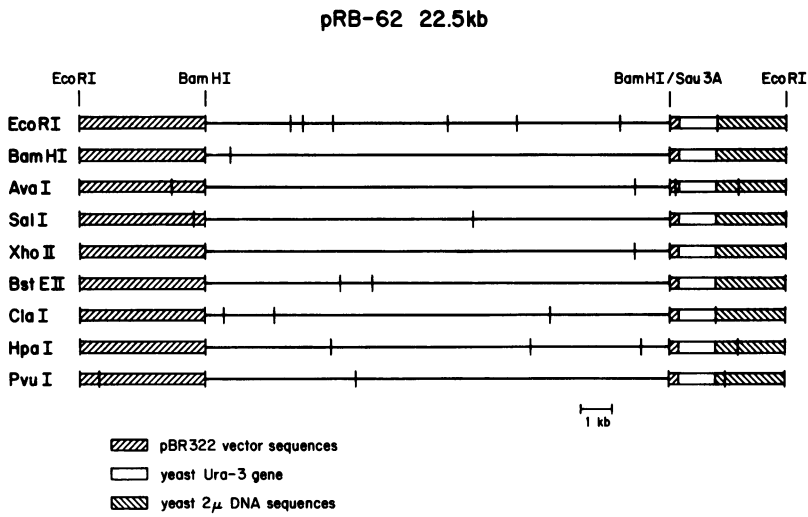

Figure 1.Restriction map of a class II plasmid which complements the fructose utilization defect of yeast strain DBY1175. Hexokinase purified from DBY1175/pRB62 has a fructose to glucose activity ratio characteristic of hexokinase B. Restriction of the plasmid pRB62 with nine enzymes yielded the map shown above. The total length of the plasmid is 22.5 kb. The yeast DNA insert is a partial Sau 3A fragment 13.6 kb in length, which has been inserted into the BamHI site of the vector pRB5. Vector sequences derived from pBR322, the yeast ura3 gene and the yeast 2 micron DNA circle are indicated.

which contained the parent plasmid pRB62, thus localizing the HKB gene to the left half of the yeast DNA insert. Various subclones of pBR322(SalII) were constructed and assayed for glucose phosphorylating activity to more precisely localize the HKB gene prior to DNA sequencing. The resulting 8.5kb plasmid which contained the HKB gene is shown in Fig. 2. This was used for amplification and provided sufficient DNA for cloning into suitable M13 mp vectors prior to Sanger dideoxy sequencing (13,14). The sequencing strategy used for the HKB gene is shown in Fig. 3. The location of the HKB gene was confirmed by translation of the nucleotide sequence that we obtained from subclones which started at the PstI site (nucleotide position 1244) (Fig. 3) and extended in the coding and non-coding directions. The sequence from one of the subclones in one reading frame translated to give the amino acid sequence -Ala-Asp-Gly-Ser-Val-Tyr-Asn-Arg-. This matched the sequence of a tryptic peptide that we had previously determined by Edman degradation. The sequence from the subclone extending in the opposite direction from the PstI

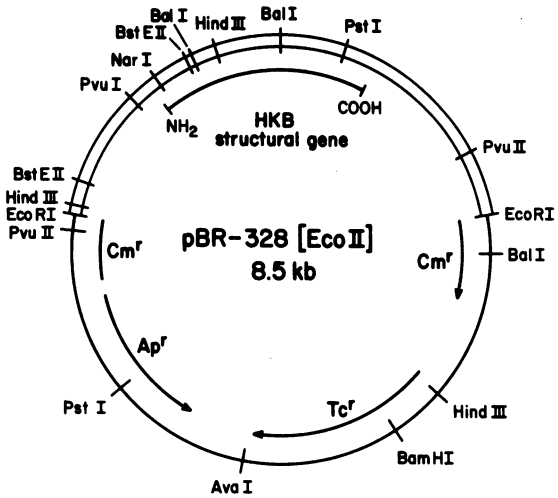

Figure 2.Restriction map of a subclone of pRB62 containing the HKB gene. Cell free extracts made from E. coli strain ZSC113/pEcoII demonstrated the ability to phosphorylate glucose in the in vitro glucose phosphorylating system described in the text. The 3.6 kb yeast DNA insert is shown inserted into the EcoRI site of pBR328, inactivating the chloramphenicol acetyltransferase gene. Deletion of the DNA between the two BstEII sites produces a plasmid which demonstrates no activity in the in vitro glucose phosphorylation assay. The location and orientation of the HKB gene as deduced from the DNA sequence is indicated.

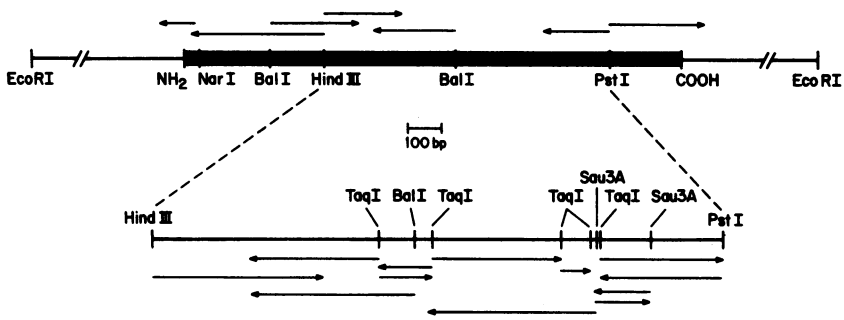

Figure 3.Sequencing strategy of the HKB gene. The HKB gene was located by translating into all 3 reading frames the DNA sequences originating at the PstI restriction site at nucleotide position 1244. Comparison of these derived amino acid sequences to the sequences of purified tryptic peptides of HKB revealed a match with one such peptide located at amino acid residues 385-396 in the 2.5 angstrom x-ray map of Anderson (5). The extent of the structural gene was calculated relative to this Pst site and DNA sequencing was initiated at the other previously identified restriction sites within this region. The 843 bp HindIII/PstI fragment was sequenced by shotgun cloning of a TaqI digest into AccI cleaved mp7. These fragments were ordered by a second digest of the HindIII/PstI fragment. A Sau3A digest was cloned into the BamHI site of mp7 for this purpose.

site was translated to give -Thr-Gly-His-Ile-Ala-Ala-, where the carboxy terminal Ala overlapped the amino terminal Ala from the first subclone. Since this tryptic peptide had been tentatively assigned as spanning residues 385-396 in the approximated structure derived from x-ray crystallography (5), the 5′ and 3′ ends of the HKB structural gene were calculated to be roughly 1170 nucleotides upstream and 210 downstream from the PstI site. Since a number of restriction sites had already been identified in these regions, they were used to obtain suitable DNA fragments from pBR328(EcoII) for dideoxy sequencing. These were cloned into M13mp7 or M13mp8 and provided the sequence information for the regions in Fig.3 shown above the HKB gene. As can be seen from the diagram, the results obtained from this approach did not cover the entire gene nor did they provide information for the sequence of both strands. To obtain the remaining sequences, an 843bp fragment extending from the HindIII to the PstI site (Fig.3) was isolated to allow "shotgun" cloning in this region. The enzyme Taq I was selected, and after digestion, the mixture was ligated into M13mp7 that had been linearized with AccI. Single-stranded phage DNA obtained from transformants of JM103 was used for sequencing. To correctly place the TaqI fragments, overlapping regions from Sau3a and BalI digests were cloned into M13mp7 and M13mp8 and sequenced as indicated in Fig. 3. The order of the the TaqI fragments was also verified by matching the amino acid sequences derived from the translation of the TaqI fragments with the sequences of the HKB tryptic peptides (Fig. 4). This figure includes the complete sequence of the HKB gene, its flanking regions and the translated amino acid sequence for the enzyme.

Isolation and Sequence Determination of Tryptic Peptides from HKB.

Tryptic peptides derived from 500 nmol of performic acid oxidized HKB were separated by cation exchange chromatography giving an elution profile shown in Fig. 5. Of the expected 52 peptides, 25 were obtained in pure enough form for sequencing. The remainder were present as mixtures from which some of the remaining peptides were isolated by rechromatography on an anion exchange matrix. Sequences of homogeneous tryptic peptides were determined by manual Edman degradation and were matched with the HKB DNA sequence as shown in Fig. 4. In addition to the amino and carboxy terminal peptides, which define the end points of the protein, the remaining peptides are distributed throughout the polypeptide chain and taken together, confirm 43% of the DNA sequence. Of special interest are the tryptic peptides containing cysteic acid residues which account for all of the cysteine residues of HKB. In all cases there was complete agreement between the amino acid sequences of

```
     -240 *                                        -210 *
HKA       ATG TCT CAA CTG CTT CTG TTT CCT CCT TTT CTT TAA AGA GGA ATA TTT CGT ATA TAA GCA
     -180 *                                        -150 *
HKA       ATC GGT TTC ACT TCC TTG GGA ATA TTC TAC CGT TCC TTC ATC TTG TAT TCT TCT CTT TCT
     -120 *                                        -90 *
HKB                                  TAT AAC TTA ACT TCA AGG TTT CTT AAT ATT TTT TCG CTT TTT
HKA       CTT AGC GCA GAA GAG CAA --A G-A AC- -T- GTG GCT -GC AA- -C- CAA --A GAA T-C CAA
      -60 *                                        -30 *                                  -1 *
HKB       CTT TGA AAA GGT TGT AGG AAT ATA ATT CTC CAC ACA TAA TAA GTA CGC TAA TTA AAT AAA
HKA       TAT ATA GTT TC- -TA -TC --A C-C -CC -AA ACA --T C-- -T- -A- TA- -G- AA- --- --G
      +1 *                                        30 *                                   60 *
HKB       ATG GTT CAT TTA GGT CCA AAA AAA CCA CAA GCC AGA AAG GGT TCC ATG GCC GAT GTG CCA
HKA       --- --- --- --- --- --- --G --- --- --G --T --- --- --- --- --- --T --- --- --C
HKB       Met Val His Leu Gly Pro Lys Lys Pro Gln Ala Arg Lys Gly Ser Met Ala Asp Val Pro
HKA
                                                90 *                                   120 *
HKB       AAG GAA TTG ATG CAA CAA ATT GAG ATT TTT GAA AAA ATT TTC ACT GTT CCA ACT GAA ACT
HKA       --- --- --- --- G-T G-- --- C-T CAG --G --- G-T --G --T --A --- GAC -GC --G --C
HKB    20 Lys Glu Leu Met Gln Gln Ile Glu Ile Phe Glu Lys Ile Phe Thr Val Pro Thr Glu Thr
HKA                       Asp Glu     His Gln Leu     Asp Met             Asp Ser
                                                150 *                                  180 *
HKB       TTA CAA GCC GTT ACC AAG CAC TTC ATT TCC GAA TTG GAA AAG GGT TTG TCC AAG AAA GGT
HKA       --G AG- AAG --- GTT --- --- --T --C GA- --- --- A-T --A --- --- A-A --- --G --A
HKB    40 Leu Gln Ala Val Thr Lys His Phe Ile Ser Glu Leu Glu Lys Gly Leu Ser Lys Lys Gly
HKA          Arg Lys     Val                 Asp         Asn             Thr
                                                210 *                                  240 *
HKB       GTT AAC ATT CCA ATG ATT CCA GGT TGG GTT ATG GAT TTC CCA ACT GGT AAG GAA TCC GGT
HKA       --- --- --- --- --- --- --C --- --- --C --- --A --- --- --A --- --A --- --A ---
HKB    60 Gly Asn Ile Pro Met Ile Pro Gly Trp Val Met Asp Phe Pro Thr Gly Lys Glu Ser Gly
HKA                                                 Glu
                                                270 *                                  300 *
HKB       GAT TTC TTG GCC ATT GAT TTG GGT GGT ACC AAC TTG AGA GTT GTC TTA GTC AAG TTG GGC
HKA       A-C -AT --- --- --- --- --- --- --- --T --- --A --- --C --G --G --- --- --- A--
HKB    80 Asp Phe Leu Ala Ile Asp Leu Gly Gly Thr Asn Leu Arg Val Val Leu Val Lys Leu Gly
HKA       Asn Tyr                                                                    Ser
                                                330 *                                  360 *
HKB       GGT GAC CGT ACC TTT GAC ACC ACT CAA TCT AAG TAC AGA TTA CCA GAT GCT ATG AGA ACT
HKA       --- A-- --- --- --- --- --- --- --- --C --- --T -A- C-- --- C-- -AC --- --- --C
HKB   100 Gly Asp Arg Thr Phe Asp Thr Thr Gln Ser Lys Tyr Arg Leu Pro Asp Ala Met Arg Thr
HKA          Asn His                                     Lys         His Asp
                                                390 *                                  420 *
HKB       ACT CAA AAT CCA GAC GAA TTG TGG GAA TTT ATT GCC GAC TCT TTG AAA GCT TTT ATT GAT
HKA       --  A-G C-C -A- --G --G --A --- TCC --- --- --- --- --- --- --G -AC --- --G -TC
HKB   120 Thr Gln Asn Pro Asp Glu Leu Trp Glu Phe Ile Ala Asp Ser Leu Lys Ala Phe Ile Asp
HKA          Lys His Gln Glu             Ser                             Asp     Met Val
                                                450 *                                  480 *
HKB       GAG CAA TTC CCA CAA GGT ATC TCT GAG CCA ATT CCA TTG GGT TTC ACC TTT TCT TTC CCA
HKA       ---, --- GAA TTG -T- AAC -C- AAG --C A-C T-A --- --A --- --- --- --C --G -A- ---
HKB   140 Glu Gln Phe Pro Gln Gly Ile Ser Glu Pro Ile Pro Leu Gly Phe Thr Phe Ser Phe Pro
HKA             Glu Leu Leu Asn Thr Lys Asp Thr Leu                                 Tyr
                                                510 *                                  540 *
HKB       GCT TCT CAA AAC AAA ATC AAT GAA GGT ATC TTG CAA AGA TGG ACT AAA GGT TTT GAT ATT
HKA       --- --C --- --- --G --T --C --- --- --T --- --- --- --- --- --G --- --C --- ---
HKB   160 Ala Ser Gln Asn Lys Ile Asn Glu Gly Ile Leu Gln Arg Trp Thr Lys Gly Phe Asp Ile
HKA
                                                570 *                                  600 *
HKB       CCA AAC ATT GAA AAC CAC GAT GTT GTT CCA ATG TTG CAA AAG CAA ATC TCT AAG AGG AAT
HKA       --- --T G-C --- GG- --- --- --C --C --- T-- C-A --- --A G-- --T --C --- --A G-G
HKB   180 Pro Asn Ile Glu Asn His Asp Val Val Pro Met Leu Gln Lys Gln Ile Ser Lys Arg Asn
HKA             Val     Gly                         Leu             Glu                 Glu
                                                630 *                                  660 *
HKB       ATC CCA ATT GAA GTT GTT GCT TTG ATA AAC GAC ACT ACC GGT ACT TTG GTT GCT TCT TAC
HKA       T-G --T --- --- A-- --A --A --- --T --T --T --- GTT --- --- --A A-- --C --A ---
HKB   200 Ile Pro Ile Glu Val Val Ala Leu Ile Asn Asp Thr Thr Gly Thr Leu Val Ala Ser Tyr
HKA       Leu             Ile                             Val             Ile
```

```
                                                                690 *                                                                720 *
HKB       TAC ACT GAC CCA GAA ACT AAG ATG GGT GTT ATC TTC GGT ACT GGT GTC AAT GGT GCT TAC
HKA       --- --- --- --- --G --- --- --- --- --G --T --- --- --- --- --- --C --- --A -T-
HKB   220 Tyr Thr Asp Pro Glu Thr Lys Met Gly Val Ile Phe Gly Thr Gly Val Asn Gly Ala Tyr
HKA                                                                                      Phe
                                                                750 *                                                                780 *
HKB       TAC GAT GTT TGT TCC GAT ATC GAA AAG CTA CAA GGA AAA CTA TCT GAT GAC ATT CCA CCA
HKA       --T --- --- --- --- --- --- --- --- T-G G-G --C --- T-- G-A --C --T --- --- AGT
HKB   240 Tyr Asp Val Cys Ser Asp Ile Glu Lys Leu Gln Gly Lys Leu Ser Asp Asp Ile Pro Pro
HKA                                                 Glu                  Ala                  Ser
                                                                810 *                                                                840 *
HKB       TCT GCT CCA ATG GCC ATC AAC TGT GAA TAC GGT TCC TTC GAC AAT GAA CAT GTC GTT TTG
HKA       AAC T-- --- --- --T --- --T --- --- --T --- --- --- --T --- --- --- T-G --C ---
HKB   260 Ser Ala Pro Met Ala Ile Asn Cys Glu Tyr Gly Ser Phe Asp Asn Glu His Val Val Leu
HKA       Asn Ser                                                                    Leu
                                                                870 *                                                                900 *
HKB       CCA AGA ACT AAA TAC GAT ATC ACC ATT GAT GAA GAA TCT CCA AGA CCA GGC CAA CAA ACC
HKA       --- --- --C --G --- --- G-T G-T G-C --C --- C-- --- --- --- --T --T --- --- G-T
HKB   280 Pro Arg Thr Lys Tyr Asp Ile Thr Ile Asp Glu Glu Ser Pro Arg Pro Gly Gln Gln Thr
HKA                           Val Ala Val         Gln                                    Ala
                                                                930 *                                                                960 *
HKB       TTT GAA AAA ATG TCT TCT GGT TAC TAC TTA GGT GAA ATT TTG CGT TTG GCC TTG ATG GAC
HKA       --- --- --G --- A-C --C --- --- --- --G --- --- T-G --- --- C-A -TG --A C-T --A
HKB   300 Phe Glu Lys Met Ser Ser Gly Tyr Tyr Leu Gly Glu Ile Leu Arg Leu Ala Leu Met Asp
HKA                     Thr                           Leu              Val      Leu Glu
                                                                990 *                                                               1020 *
HKB       ATG TAC AAA CAA GGT TTC ATC TTC AAA AAC CAA GAC TTG TCT AAG TTC GAC AAG CCT TTC
HKA       T-A A-- G-G A-G --C --G --G --G --G G-T --- --T C-A AGC --- --G A-A C-A --A -A-
HKB   320 Met Tyr Lys Gln Gly Phe Ile Phe Lys Asn Gln Asp Leu Ser Lys Phe Asp Lys Pro Phe
HKA       Leu Asn Glu Lys      Leu Met Leu     Asp                      Leu Lys Gln     Tyr
                                                               1050 *                                                               1080 *
HKB       GTC ATG GAC ACT TCT TAC CCA GCC AGA ATC GAG GAA GAT CCA TTC GAG AAC CTA GAA GAT
HKA       A-- --- --T --C --C --- --- --A --- --- --- --T --- --- --T --A --- T-G --- ---
HKB   340 Val Met Asp Thr Ser Tyr Pro Ala Arg Ile Glu Glu Asp Pro Phe Glu Asn Leu Glu Asp
HKA       Ile                                       Asp
                                                               1110 *                                                               1140 *
HKB       ACC GAT GAC TTG TTC CAA AAT GAG TTC GGT ATC AAC ACT ACT GTT CAA GAA CGT AAA TTG
HKA       --T --- --- A-- --- --- --G --C --T --- G-- --G --C --- C-G -C- --- --- --G ---
HKB   360 Thr Asp Asp Leu Phe Gln Asn Glu Phe Gly Ile Asn Thr Thr Val Gln Glu Arg Lys Leu
HKA                   Ile         Lys Asp         Val Lys         Leu Pro
                                                               1170 *                                                               1200 *
HKB       ATC AGA CGT TTA TCT GAA TTG ATT GGT GCT AGA GCT GCT AGA TTG TCC GTT TGT GGT ATT
HKA       --T --- A-A C-T -G- --- --- --C --- A-C --- --- --- --- --A G-T --- --- --- ---
HKB   380 Ile Arg Arg Leu Ser Glu Leu Ile Gly Ala Arg Ala Ala Arg Leu Ser Val Cys Gly Ile
HKA                     Cys                     Thr                     Ala
                                                               1230 *                                                               1260 *
HKB       GCT GCT ATC TGT CAA AAG AGA GGT TAC AAG ACC GGT CAC ATC GCT GCA GAC GGT TCC GTT
HKA       --C --- --T --C --- --- --- --- --- --- --T --- --- --T --C --T --- --- --T --C
HKB   400 Ala Ala Ile Cys Gln Lys Arg Gly Tyr Lys Thr Gly His Ile Ala Ala Asp Gly Ser Val
HKA
                                                               1290 *                                                               1320 *
HKB       TAC AAC AGA TAC CCA GGT TTC AAA GAA AAG GCT GCC AAT GCT TTG AAG GAC ATT TAC GGC
HKA       --T --- -A- --- --- --- --- --G --- GCC --C --T --G -G- --- -GA --T --C --T --A
HKB   420 Tyr Asn Arg Tyr Pro Gly Phe Lys Glu Lys Ala Ala Asn Ala Leu Lys Asp Ile Tyr Gly
HKA               Lys                     Ala         Lys Gly     Arg
                                                               1350 *                                                               1380 *
HKB       TGG ACT CAA ACC TCA CTA GAC GAC TAC CCA ATC AAG ATT GTT CCT GCT GAA GAT GGT TCC
HKA       --- --- GGT GAA AAC GC- AG- A-A G-T --- --T -C- --- --- --A --- --G --- --- --A
HKB   440 Trp Thr Gln Thr Ser Leu Asp Asp Tyr Pro Ile Lys Ile Val Pro Ala Glu Asp Gly Ser
HKA               Gly Glu Asn Ala Ser Lys Asp         Arg
                                                               1410 *                                                               1440 *
HKB       GGT GCT GGT GCC GCT GTT ATT GCT GCT TTG GCC CAA AAA AGA ATT GCT GAA GGT AAG TCC
HKA       --- --A --- --T --- --- --- --- --A --- T-- G-- --- --- --- --C --- --- --- GT-
HKB   460 Gly Ala Gly Ala Ala Val Ile Ala Ala Leu Ala Gln Lys Arg Ile Ala Glu Gly Lys Ser
HKA                                           Ser Glu                                    Val
                                                               1470 *
HKB       GTT GGT ATC ATC GGT GCT TAA ACT TAA TTT GTA AAT TAA GTT TGA
HKA       TC- --- --- --T --C --- --- TGA A-- -G- AAT G-A
HKB   480 Val Gly Ile Ile Gly Ala stop
HKA       Ser                   stop
```

the tryptic peptides and the DNA sequence. No tryptic peptides were isolated which could not be matched with the DNA sequence.

Nucleotide Sequence of the HKA Gene.

The determination of the nucleotide sequence of the HKA gene was greatly simplified by knowledge of the HKB gene sequence. Two overlapping recombinant plasmids established an approximately 4kb overlap region containing the HKA gene and its control sequences as shown in Fig.6. A centrally located 1.5 kb Bgl II fragment from this region was cloned into the Bam HI site of M13mp9 in both orientations to allow determination of the nucleotide sequences of its 5′ and 3′ termini. The sequences and their complements were translated in all three reading frames and compared to the amino acid sequence of HKB. The translation of the complement of the nucleotide sequence originating at the righthand BglII site (Fig. 6) showed strong homology to the sequence of HKB at amino acid residues 240-331, thus establishing the location and coding direction of the structural gene for HKA within the 4 kb overlap region. Since the HKA structural gene was assumed to be very similar in size to the HKB structural gene, the limits of the structural gene were calculated relative to the Bgl II site identified in the restriction map of pRB141 (Fig. 6).

The strategy used to complete the nucleotide sequence of the HKA gene was very similar to that employed for the HKB gene. Nucleotide sequences obtained around BglII, SalI, and Hind III sites (Fig. 7) allowed the identification of "secondary" sites which were used to extend the previous sequence data or to check it by determination of the nucleotide sequence of the opposite strand (Fig. 7). As with the HKB gene the interior of the HKA gene was found to be devoid of restriction sites suitable for cloning and sequencing in the M13 system. Thus, the entire 1.5 kb Bgl II fragment (which contained about two thirds of the structural gene) was cut with Taq I and the digest ligated into AccI cleaved M13mp7. In the case of the HKA gene, the high degree of amino acid sequence homology between HKB and HKA made ordering of the restriction fragments considerably easier. Overlaps were

Figure 4.Nucleotide and amino acid sequences of the HKA and HKB genes. The nucleotide sequence of the HKB gene and its derived amino acid sequence are given, and only the differences between the HKA and HKB sequences are indicated. Regions of the amino acid sequence in italic type are areas of HKB whose sequence was determined directly from peptides isolated from a tryptic digest of HKB protein. Nucleotide +1 is the first nucleotide of the ATG translation initiation codon of these proteins. Both structural genes are 1455 nucleotides in length and code for proteins of 485 amino acids.

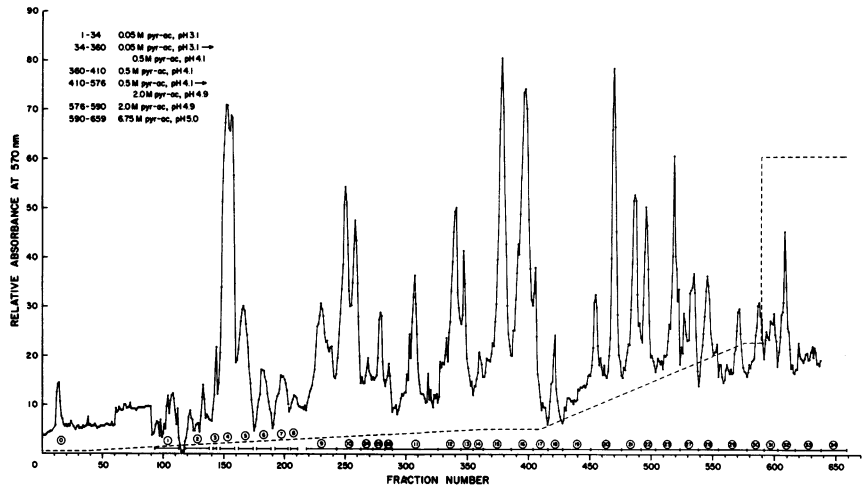

Figure 5. Elution profile of a tryptic digest of HKB protein. 500 nmol of a tryptic digest of performic acid oxidized HKB was loaded on a 0.9 x 60 cm column of Aminex AG-50W-X4 equilibrated in 0.05M pyridinium acetate, pH 3.1. Fractions of 1.5 ml were collected at a flow rate of 50 ml/hr. The gradients employed are indicated in the figure. Peptides were detected by the alkaline ninhydrin method on 50 microliter aliquots of each fraction. Pooled fractions whose amino acid composition was analyzed after acid hydrolysis are indicated.

obtained for most of the fragments, both by sequencing upstream from the single HindIII site at nucleotide position 896, and by performing a secondary "shot gun" cloning of the same fragment using the enzyme MspI. The extent of individual nucleotide sequences and their overlaps are shown in Fig. 7. The nucleotide sequence of the HKA gene is shown in Fig. 4, aligned with the sequence of the HKB gene.

DISCUSSION

One of the major reasons for our attempting to work out the primary structures of HKA and HKB was to allow the construction of models based on x-ray data at the level of atomic resolution. Up to now, this has not been possible because of the lack of amino acid sequence information.

The concerted use of recombinant DNA technology and DNA sequencing, together with protein chemistry, has made it possible to elucidate the primary structure of these proteins with a high degree of confidence. To illustrate the advantages of this combined approach, we will cite just a few examples

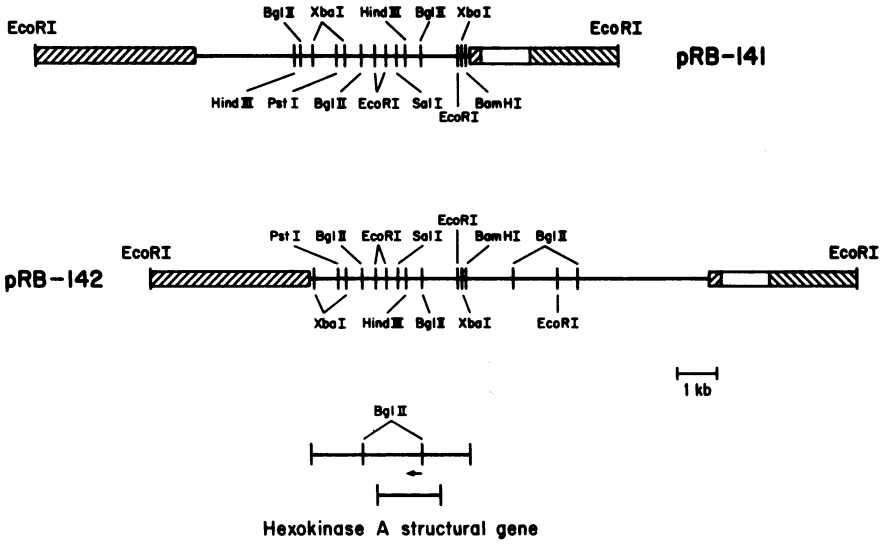

Figure 6.Restriction maps of two overlapping class I plasmids. The maps are aligned to indicate the area of yeast genomic DNA common to both plasmids. The HKA structural gene was located by comparing the amino acid sequences derived from the DNA sequences determined at each end of a 1.5kb BglII fragment (indicated) to the known amino acid sequence of HKB. DNA sequence originating at the BglII site on the right hand side of this fragment yielded an amino acid sequence which was extremely homologous to residues 240-331 of HKB, identifying the location and orientation of the HKA gene within the plasmids.

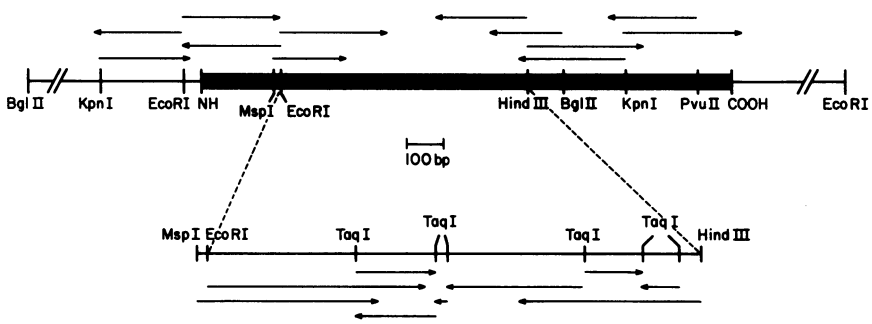

Figure 7.Sequencing strategy of the HKA gene. The extent of nucleotide sequences originating at various restriction sites is indicated. The sequence of the interior of the gene was determined from clones obtained by a "shotgun" cloning of the 1.5 kb BglII fragment which had been cleaved with the enzyme TaqI. The TaqI fragments were ordered by overlaps obtained from a second "shotgun" cloning of the 1.5 kb BglII fragment cleaved with MspI and also by comparing the derived amino acid sequences of the TaqI fragments to the amino acid sequence of HKB.

where it not only speeded up the work but also helped us to avoid errors. First of all, the availability of portions of the amino acid sequence of HKB was crucial in allowing us to definitively locate the position of the HKB structural gene in the recombinant plasmid and also to define the 5′ and 3′ termini based on our knowledge of the amino and carboxyterminal sequences of HKB. Reading frame shifts in the nucleotide sequence due to deletion or addition of bases became obvious when two sequenced peptides appeared in different reading frames of the nucleotide sequence. This discrepancy alerted us to the necessity of going back and checking the DNA sequence once again. Many of the sequenced peptides provided the information needed to confirm the ordering of restriction fragments, especially when the sequencing was carried out by the "shotgun" cloning of TaqI and Sau3A fragments. Discrepancies were occasionally noticed between the primary structure of the chemically sequenced tryptic peptides and an amino acid sequence derived from the region of the structural gene which we suspected coded for that peptide. In all but two instances, the amino acid sequence data was found to be correct when we rechecked the area of the gene in question by repeating the DNA sequencing.

Much of the nucleotide sequence of both hexokinase structural genes was determined form both strands of DNA. Approximately 68% of the sequence of HKA gene was determined from both coding and non-coding strands. With the HKB gene, approximately 60% was determined from both strands of the DNA. This, together with the protein sequence data, served as a check for sequence errors that might have been caused by sequence-specific secondary structure effects. An additional degree of confidence was gained by comparing the nucleotide sequences of the HKA and HKB structural genes with each other. For example, after the completion of the DNA sequence of the HKA gene, comparison with HKB showed that there was a small gap in the HKB sequence upstream of the Hind III site at nucleotide 407. Although very close to the Hind III site from which sequencing was initiated, the sequence in this section (approximately from nucleotide 375 to nucleotide 400) presented problems which might have been due to secondary structural effects in the template DNA. The lack of amino acid sequence data in this area prevented further efforts to verify the nucleotide sequence in this section. However, an homologous region of HKA gene was found to match the HKB gene on both sides of the presumed gap but contained 25 extra nucleotides not present in the HKB gene. Reinvestigation of this section of the HKB gene showed that these nucleotides of the template DNA had apparently been skipped over by the

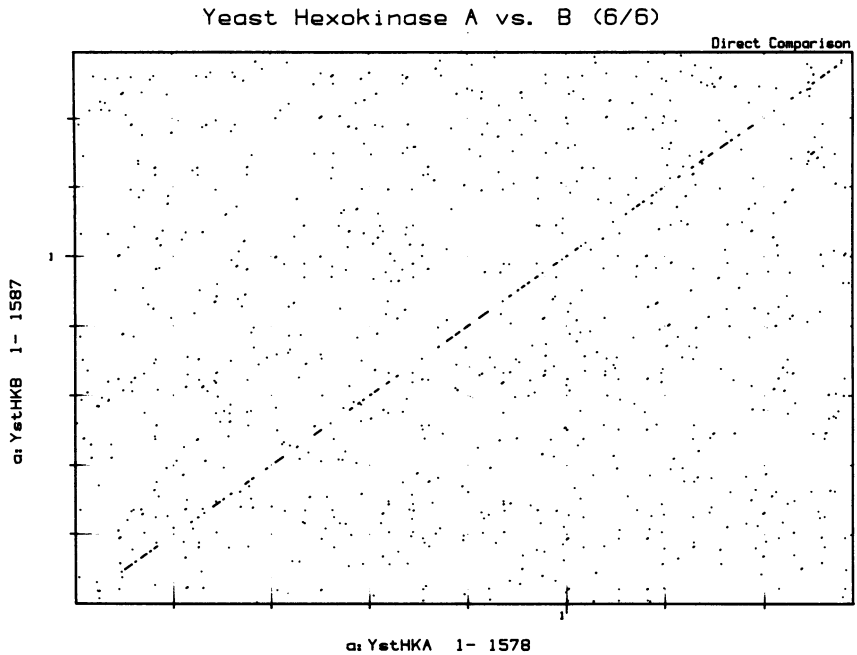

Figure 8. Comparison of the HKA and HKB genes. A two dimensional plot of HKA vs HKB is shown in which regions of identical nucleotide sequence are indicated as a solid line on the diagonal of the plot.

Klenow fragment of DNA polymerase in multiple sequence determinations. When this region of HKB gene was sequenced in the opposite (coding) direction, starting from the BalI site located at nucleotide 148 (see Fig. 4) this effect was not seen. No obvious reason for this behavior can be discerned from an analysis of the nucleotide sequence of this region; no sequences appear to be present which might base pair and cause structural distortion of the template DNA. After this correction, the correspondence between the HKA and HKB structural, in terms of number of residues, matched exactly.

Comparison of HKA and HKB Sequences

The amino acid sequences specified by the HKA and HKB genes are shown in Fig. 4. Comparison of these sequences reveals that 378 out of 485 amino acid residues are identical. Fig. 9 shows a two dimensional matrix comparison of the HKA gene vs. the HKB gene. Using this method, it can readily be observed that several discreet areas of nucleotide sequence disparity exist. Dissimilar regions of amino acid sequence cluster in four areas. In the 2.5 angstrom model of Anderson (5), all except one of these regions of differing sequence

occurs on the exterior of the hexokinase B monomer. None of these clusters appear to be involved in hexose binding, although the third cluster, residues 318-329, is located very close to the site of ATP binding. Since glucose repression in yeast seems to be associated only with the presence of a functional HKB gene (21), it is tempting to speculate that the existence of a functional regulatory domain in HKB is directly attributable to primary structural differences observed in these regions. The identity of the amino acid residues responsible for the observed differences in catalysis and regulation will undoubtably become clearer as the three dimensional structures of the hexokinase A and hexokinase B isozymes are refined with the aid of this new data (Harrison & Steitz, manuscript in preparation).

5′ Untranslated Sequences

Most eukaryotic mRNA coding genes possess an AT-rich promoter with the concensus sequence TATAAA (22). In multicellular eukaryotes this sequence is located 25-35 nucleotides upstream of the transcription initiation site and is bounded by GC-rich sequences; initiation usually occurs about 32 nucleotides downstream of this site (23). The TATAAA sequence is also one element of promoters in S. cerevisiae which are recognized by RNA polymerase II (24). The second elememt, an upstream activator sequence, is located a considerable distance upstream of the TATA box (25). The TATA box has been found in nearly all yeast genes examined, but in contrast to multicellular eukaryotes, is not surrounded by GC-rich sequences (26,27). Its location is also more variable with respect to the site of transcriptional initiation and is usually further upstream than its multicellular eukaryotic counterpart (28).

Two sequences in the 5′ untranslated region of hexokinase A have a high degree of homology with the consensus TATA promoter sequence. These are located at nucleotide positions -188 (TATAA) and -60 (TATATA) with respect to the ATG start codon. In the hexokinase B gene, a TATAA sequence is located at nucleotide position -102. None of these three sequences are surrounded by GC-rich sequences.

Other 5′ noncoding sequences have been noted in mRMA coding yeast genes, but their function is more speculative. A pyrimidine rich cluster has been noted between the TATA box and the 5′ coding end of several yeast genes, followed soon after by the sequence CAAG (26,27,28,29). Although initially described in the 5′ untranslated region of yeast glycolytic enzyme genes (26), Hobson (29), and Montgomery (30) have suggested that these sequences are common to any highly expressed yeast gene. Transcription starts at

sequences identical or related to CAAG in several yeast genes, suggesting a role for this sequence in transcription initiation or capping of mRNAs (31).

In the hexokinase A gene, the TATA box at position -188 is followed by a cluster of pyrimidines at position -133. This is followed by the sequence CAGAAG. The TATA box at -60 is also followed by a string of pyrimidines, but no CAAG related sequence is found in the immediate downstream region. The hexokinase B gene has a pyrimidine rich cluster starting at position -74 and is followed by the sequence GAAAG. The sites of transcription initiation for both of these genes is not known, so it is not possible to relate these observed sequences to experimentally identified transcriptional start sites.

3′ Untranslated Sequences

At least three putative transcription termination sequences have been proposed in S. cerevisiae. Nucleotide sequences related to AATAAG have been shown to occur upstream of the site of poly(A) addition in several yeast mRNAs; this seguence itself is homologous to the sequence AATAAA which is presumed to signal the site of poly(A) addition in multicellular eukaryotes (32,33). Bennetzen and Hall have proposed that the concensus sequence, TAAATAA(A/G), may represent a typical yeast terminator structure (28). Zaret and Sherman have identified a tripartate structure with the concensus sequence TAG .... TATGT .... TTT (34). Deletion of this sequence from the 3′ end of the yeast cyc1 gene greatly reduces the amount of mRNA terminating at the wild type site. Furthermore, spontaneous revertants of this deletion with near normal mRNA termination have sequences which tend to resemble the original terminator structure (35). Finally, Henikoff et al. (36) have shown that the sequence AATAAA is not needed for termination of a drosophila gene which complements an Ade8 mutation in yeast. Their deletion analysis suggests that the sequence TTTTTATA is necessary for efficient termination, although Zaret and Sherman have noted that most yeast genes lack an identical or related sequence (35).

Our sequences of the HKA and HKB genes lack sufficient 3′ flanking sequences to allow a comparison which might reveal the existence of a sequence homologous to the putative terminators described earlier.
However, one structure is noteworthy in the hexokinase B gene namely a 24 bp region of dyad symmetry which occur immediately 3′ to the coding sequence and includes the TAA stop codon. This hyphenated palindrome contains a sequence starting at nucleotide 1472 which is clearly related to the concensus sequence of Bennetzen and Hall. It is similar to rho dependent terminators seen in prokaryotes in that it is an AT rich sequence occuring in a region of dyad

symmetry (37), and may therefore represent a terminator structure which is analogous to certain terminators seen in bacteria.

Acknowledgements

The authors would like to acknowledge the expert help of Norma Neff in the cloning of the HKA and HKB genes. Support for this work was provided by USPHS Grant GM12607.

REFERENCES

1. Fletterick, R.J., Bates, D.J., and Steitz, T.A. (1975) Proc. Natl. Acad. Sci. USA 72, 38-42.
2. Steitz, T.A., Fletterick, R.J., Anderson, W.A. and Anderson, C.M. (1976) J. Mol. Biol. 104, 197-222.
3. Shoham, M. and Steitz, T.A. (1982) Biochim. Biophys. Acta 705, 380-384.
4. Anderson, C.M., McDonald, R.C. and Steitz, T.A. (1978) J. Mol. Biol. 123, 1-13.
5. Anderson, C.M., Stenkamp, R.E. and Steitz, T.A. (1978) J. Mol. Biol. 123, 15-33.
6. Bennett, W.S., Jr., and Steitz, T.A. (1978) Proc. Natl. Acad. Sci. USA 75, 4848-4852.
7. Lobo, Z. and Maitra, P.K. (1977) Genetics 86, 727-744.
8. Curtis, S.J. and Epstein, W. (1975) J. Bacteriol. 122, 1189-1199.
9. Carlson, M. and Botstein, D. (1982) Cell 28, 145-154.
10. Gancedo, J.M., Clifton, D., and Fraenkel, D.G. (1977) J. Biol. Chem. 252, 4443-4444.
11. Falco, S.C., and Botstein, D. (1983) Genetics 105, 857-872.
12. Mortimer, R.K., and Schild, D. (1982) in The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression (J.N. Stratesu, E.W.Jones and J.R. Boach, Eds.) Cold Spring Harbor Press, pp. 639-650.
13. Sanger, F., Nicklens, S., and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 3642-3647.
14. Messing, J. and Vieira, J. (1982) Gene 19, 263.
15. Lazarus, N.R., Ramel, A.H., Rustum, Y.M. and Barnard, E.A. (1966) Biochemistry 5, 4003-4016.
16. Hirs, C.H.W. (1967) Methods in Enzymology 11, 197-199.
17. Schroeder, W.A. (1967) Methods in Enzymology 11, 351-360.
18. Hirs, C.H.W. (1967) Methods in Enzymology 11, 325-329.
19. Tomita, M., Furthmayr, H., Marchesi, V.T. (1978) Biochemistry 17, 4756-4770.
20. Schroeder, W.A. (1967) Methods in Enzymology 11, 361-368.
21. Entian, K.D., Kopetzki, E., Frohlich, K., and Mecke, D. (1984) Mol. Gen. Genet. 198, 50-54.
22. Breathnach, R. and Chambon, P. (1981) Annu. Rev. Biochem. 50, 349-383.
23. Faye, G., Leung, D.W., Tatchell, K., Hall, B.D. and Smith, M. (1981) Proc. Natl. Acad. Sci. USA 78, 2258-2267.
24. Guarente, L. (1984) Cell 36, 799-800.
25. Guarente, L., Lalonde, B., Gifford, P. and Alanui, E. (1984) Cell 36, 503-511.
26. Holland, J.P. and Holland, M.J. (1980) J. Biol. Chem. 255, 2596-2605.
27. Holland, J.P. and Holland, M.J. (1979) J. Biol. Chem. 254, 9839-9845.
28. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3018-3025.

29. Dobson, M.J., Tuite, M.F., Roberts, N.A., Kingman, A.J., and Kingman, S.M. (1982) Nucl. Acids Res. 10, 2625-2637.
30. 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.
31. Burke, R.L., Tekamp-Olson, P., and Najarian, R. (1983) J. Biol. Chem. 258, 2193-2201.
32. Bennoist, C., O'Hare, K. Breathnach, R., and Chambon, P. (1980) Nucl. Acids Res. 8, 127-142.
33. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
34. Zaret, K.S. and Sherman, F. (1982) Cell 28, 563-573.
35. Zaret, K.S. and Sherman, F. (1984) J. Mol. Biol. 176, 107-135.
36. Henikoff, S. and Cohen, E.H. (1984) Mol. Cell Biol. 4, 1515-1520.
37. Platt, T. (1981) Cell 24, 10-23.