Curr Genet (1987) 12:175–184

Current Genetics

Sequence and transcription of the β-glucosidase gene of *Kluyveromyces fragilis* cloned in *Saccharomyces cerevisiae*

Alain Raynal, Claude Gerbaud, Marie Claude Francingues, and Michel Guerineau

Laboratoire "Biologie et Génétique Moléculaire", Bâtiment 400, Université Paris-Sud, F-91405 Orsay Cedex, France

Summary. The complete nucleotide sequence of the β-glucosidase gene of *Kluyveromyces fragilis* has been determined. This sequence contains an open reading frame of 2535 base pairs encoding a protein of 845 amino acids. Analysis of the transcription products revealed only one transcript of about 3 kb identical in both *Kluyveromyces fragilis* and in the expression host *Saccharomyces cerevisiae*. The protein molecular weight of 93,811 Kd deduced from the sequence is consistent with the 90,000 Kd determined by SDS polyacrylamide gel electrophoresis with the purified protein. Mapping of the starts of transcription shows that two starting points are used in the natural host *Kluyveromyces fragilis*. A comparison of the amino acid sequence with that of other β-glucosidases revealed three regions of homology. One of these regions contains an amino acid sequence very similar to a peptide isolated from the active site of β-glucosidase A_3 from *Aspergillus wentii* and could be implicated in the catalytic mechanism of these glucolytic enzymes.

Key words: β-glucosidase – *Kluyveromyces fragilis* – DNA sequence – *Saccharomyces cerevisiae*

Introduction

The main pathway for utilization of cellobiose as a carbon source is the hydrolysis of cellobiose into two molecules of glucose by a β-glucosidase often named cellobiase. This enzyme is involved in the last step of the hydrolysis of cellulosic materials. In view of potential applications, several β-glucosidases have been isolated from various organisms and their characteristics studied (Blondin et al. 1983; Ait et al. 1982; Thomas and McCrae 1982). Recently, in vivo cloning of the genes involved in cellobiose utilization in *Erwinia* (Barras et al. 1984), and in vitro cloning of those of the bacteria *Escherichia adecarboxylata* (Armentrout and Brown 1981) has been described. Using recombinant DNA techniques, the structural genes for the β-glucosidases of several eucaryotic organisms have been isolated. We have cloned those of *Kluyveromyces fragilis* (Raynal and Guerineau 1984) and *Aspergillus niger* (Pentilla et al. 1984). The cloning of the *Candida pelliculosa* enzyme has also been reported (Kohchi and Toh-e 1986). Several species of yeast are able to utilize cellobiose as their sole carbon source. The structural gene for β-glucosidase is present but very poorly expressed in *Saccharomyces cerevisiae* (Duerksen and Halvorson 1959). The active structural gene for the β-glucosidase of *Kluyveromyces fragilis* has been cloned and expressed both in *Escherichia coli* and *Saccharomyces cerevisiae* (Raynal and Guerineau 1984). In this paper we report the determination of the nucleotide sequence of the *Kluyveromyces fragilis* β-glucosidase gene and the analysis of its transcription in *Kluyveromyces fragilis* and in *Saccharomyces cerevisiae* introduced by transformation.

Materials and methods

Strains and media. The β-glucosidase gene has been cloned from *Kluyveromyces fragilis* strain ATCC 12424 and expressed both in *Escherichia coli* HB101 (Boyer et al. 1969) and *Saccharomyces cerevisiae* strain OL1 (Boy-Marcotte and Jacquet 1982). Yeast strains were grown in YNB (Sherman et al. 1974) with glucose (2%) and casamino acids (0.2%) supplemented where appropriate with L-leucine and L-histidine (50 μg/ml) *Escherichia coli* strains used in cloning procedures were HB101 (Boyer et al.

Offprint requests to: A. Raynal

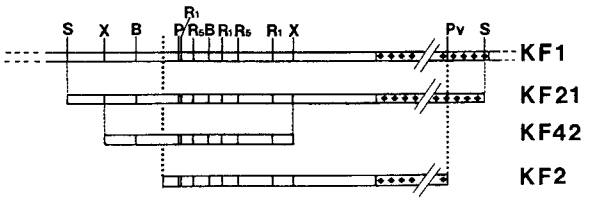

Fig. 1. Subcloning of the *Kluyveromyces fragilis* β-glucosidase gene. Linear representation of the *Kluyveromyces fragilis* DNA carried by the plasmids KF1 KF2 KF21 and KF42. In the original plasmid *Kluyveromyces fragilis* DNA was inserted in the unique *Bam*HI site of the cosmid pHCG3 (Gerbaud et al. 1981). KF2 was previously described (Raynal and Guérineau 1984). KF21 was derived from KF1 by *Sal*I digestion and religating. KF42 was obtained by insertion of the 6 kb *Xba*I fragment of KF21 in the unique *Xba*I site of pHCG3. B = *Bam*HI, P = *Pst*I, R1 = *Eco*RI, R5 = *Eco*RV, Pv = *Pvu*II, S = *Sal*I, X = *Xba*I. (♦ ♦ ♦) = pHCG3

1969) and JM101. JM103 for M13 cloning and single strand preparation for DNA sequencing (Messing 1983).

Vectors. The *Saccharomyces cerevisiae/Escherichia coli* shuttle cosmid PHCG3 (Gerbaud et al. 1981) was used for cloning of *Kluyveromyces fragilis* fragments. The phage vectors used for sequencing were M13mp10 and M13mp11 (Messing 1983).

RNA preparation and analysis. Total RNA was prepared as described by Maccechini et al. (1979), cells were disrupted with glass beads (0.45 mm diameter). Contaminating double-strand DNA was eliminated by precipitation in 3M LiCl (Richter et al. 1980). Poly(A)$^+$ RNA was purified on oligo(dT)-cellulose following the method described by Maniatis et al. (1982). For Northern blots poly(A)$^+$ RNA was denatured with glyoxal and dimethy sulfoxide (McMaster et al. 1977), separated by electrophoresis on a neutral horizontal agarose gel (in 10 mM phosphate buffer) and transferred to Pall biodyne nylon membrane (Thomas 1980). A 1 kb DNA ladder (BRL), denaturated, was used as molecular weight marker. The part of the gel carrying molecular weight markers was stained in ethidium bromide and 50 mM NaOH. S1 mapping was performed as described by Maniatis et al. (1982).

Radioactive probes. DNA probes for Northern blot hybridizations were single-strand probes derived from M13 templates (Hu and Messing 1982) labelled with [α-^{32}P]dCTP (600–800 Ci/mM) (Amersham). DNA probes for S1 experiments were obtained as described by Burke (1984) except that the sample was not denatured and the probes were detected and purified as double-stranded DNA on 1.8% agarose gels.

DNA sequencing. DNA sequences were determined with the chain termination procedure (Sanger et al. 1977), using the bacteriophage M13 system, sequencing kits and [α-^{35}S]dATP (600 Ci/mM) from Amersham.

Enzymes. Restriction endonucleases, T4 DNA ligase and DNA polI Klenow fragment were purchased from Boehringer Mannheim, S1 nuclease was from BRL.

Results

Subcloning of the β-glucosidase gene

We have previously cloned the β-glucosidase gene on a *Kluyveromyces fragilis* DNA fragment of 35 kb (Raynal and Guerineau 1984). From this original plasmid a series of plasmids containing smaller inserts and expressing the β-glucosidase at the same level in *Saccharomyces cerevisiae* were constructed. The first was KF21 with a 12 kb insert, from which a 6 kb *Xba*I fragment was isolated and cloned into the plasmid KF42 (Fig. 1). Plasmid KF42 was used as the source of DNA for sequence determination.

Nucleotide sequence of the β-glucosidase gene

The sequenced region of about 4 kb is depicted in Fig. 2. It contains only one long open reading frame beginning at position +1 and ending with the TGA stop codon at position +2538 (Fig. 3), no other open reading frame coding for a protein of the expected molecular weight (about 90 Kd) could be detected. The β-glucosidase amino acid sequence, deduced from the DNA sequence, is shown in Fig. 3. The protein is assumed

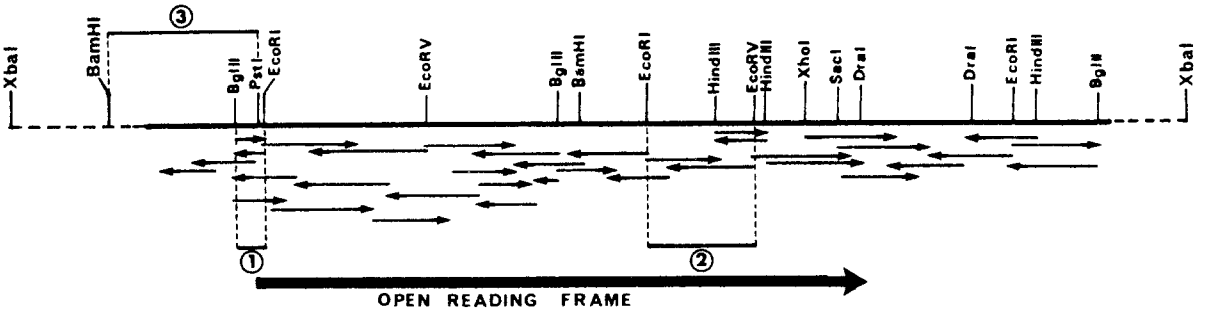

Fig. 2. Sequencing strategy. *Arrows* indicate the length and the direction of the sequences determined. The 845 codon open reading frame is shown at the *bottom* of the figure. Probes 1 and 2 were used in Northern blot analysis, probe 2 in the determination of the direction of transcription and probe 3 in S1 mapping

```
 ATTGAATAATGGGGCTGCAAAAATTAAGGGTGCTAAAATATATACCAGTACTACCTAAAGCAGTATG -545

CATTCCCATTATTACTTGGAGGAAGGGCCTTCGAAAGCGTTTGAAACAGGAGTAGGAGTAGGAGTATT -477

ACGGAGAAATGGGCATTGAATAGAGGTAGCAAAGGAAAAGCCGTGACAATGCGGCAATGTACCCAACC -409

TATGAGTCTGTGTGTGTAAAGTGTGTGTGTCTTTTTTTTTTTAATTTTAATTTTTTTTTTTAATTATT -341
                                                                    (1)
GAAAAACCAGGACTACGTCAGGTTAGGAGAATGGATATTGAAAATATTAATGTACCAGTAGTTTTGCA -273
                                    (1)
TGTATAATTATTATTACCAATAGACGAAGTGGAGGTATTAGAGCGATTTGGGGGAATTGTGATTGTAT -205
                                  (2)       (2)
ATATAGTGTGTATGTTTTACGGTACTGAAAGATCGTGTTGTAAGGAAAAATAAGGCACAGGAAAAATT -137

CAGATTTTTTAAGGGTTTATTACTAGGAGCAGTAGTAGTAGTAGAGTAGTGGTAGCAAAAGCCGAGAT  -69

CTAGGTAGAGACAGATCTGTAATACATAGCTAGTTGAAGTTGAAGAAGTGCAAACGTTAAGAAAAAAA   -1

ATG TCT AAA TTT GAT GTT GAA CAG TTA TTG AGT GAA TTG AAC CAG GAT GAA    51
Met Ser Lys Phe Asp Val Glu Gln Leu Leu Ser Glu Leu Asn Gln Asp Glu

AAG ATT TCC TTA CTT TCT GCA GTT GAT TTC TGG CAT ACT AAG AAG ATT GAA   102
Lys Ile Ser Leu Leu Ser Ala Val Asp Phe Trp His Thr Lys Lys Ile Glu

CGG TTG GGA ATT CCA GCG GTG AGG GTT TCT GAT GGT CCA AAT GGT ATT AGA   153
Arg Leu Gly Ile Pro Ala Val Arg Val Ser Asp Gly Pro Asn Gly Ile Arg

GGG ACG AAG TTC TTT GAT GGG GTT CCT TCA GGA TGT TTC CCT AAT GGT ACC   204
Gly Thr Lys Phe Phe Asp Gly Val Pro Ser Gly Cys Phe Pro Asn Gly Thr

GGG TTG GCA TCT ACT TTT GAT CGC GAT CTG CTT GAG ACA GCA GGT AAG TTG   255
Gly Leu Ala Ser Thr Phe Asp Arg Asp Leu Leu Glu Thr Ala Gly Lys Leu

ATG GCT AAG GAA TCG ATT GCG AAG AAT GCT GCT GTG ATT TTG GGT CCA ACC   306
Met Ala Lys Glu Ser Ile Ala Lys Asn Ala Ala Val Ile Leu Gly Pro Thr

ACA AAC ATG CAA CGT GGT CCT TTG GGT GGT CGT GGT TTT GAA TCA TTC TCT   357
Thr Asn Met Gln Arg Gly Pro Leu Gly Gly Arg Gly Phe Glu Ser Phe Ser

GAA GAT CCA TAT CTT GCT GGT ATG GCT ACT TCT TCT GTT GTT AAA GGT ATG   408
Glu Asp Pro Tyr Leu Ala Gly Met Ala Thr Ser Ser Val Val Lys Gly Met

CAA GGC GAA GGT ATT GCT GCT ACC GTT AAG CAT TTT GTT TGT AAC GAC TTG   459
Gln Gly Glu Gly Ile Ala Ala Thr Val Lys His Phe Val Cys Asn Asp Leu

GAA GAC CAA CGT TTT TCT TCG AAC TCA ATT GTT TCT GAA AGG GCT CTG AGA   510
Glu Asp Gln Arg Phe Ser Ser Asn Ser Ile Val Ser Glu Arg Ala Leu Arg

GAA ATT TAC TTG GAG CCC TTC AGA TTG GCA GTT AAA CAT GCC AAT CCT GTT   561
Glu Ile Tyr Leu Glu Pro Phe Arg Leu Ala Val Lys His Ala Asn Pro Val

TGT ATA ATG ACT GCT TAT AAC AAG GTC AAT GGC GAT CAT TGC TCC CAA TCC   612
Cys Ile Met Thr Ala Tyr Asn Lys Val Asn Gly Asp His Cys Ser Gln Ser

AAG AAG CTA TTG ATC GAC ATT TTG AGA GAC GAG TGG AAA TGG GAC GGT ATG   663
Lys Lys Leu Leu Ile Asp Ile Leu Arg Asp Glu Trp Lys Trp Asp Gly Met

TTA ATG TCC GAC TGG TTC GGT ACA TAT ACG ACT GCC GCA GCT ATC AAG AAT   714
Leu Met Ser Asp Trp Phe Gly Thr Tyr Thr Thr Ala Ala Ala Ile Lys Asn

GGG TTG GAT ATC GAG TTC CCT GGA CCA ACA AGA TGG AGA ACA CGT GCT TTA   765
Gly Leu Asp Ile Glu Phe Pro Gly Pro Thr Arg Trp Arg Thr Arg Ala Leu

GTG TCT CAC TCT CTC AAC TCC AGA GAA CAA ATC ACT ACT GAA GAT GTT GAT   816
Val Ser His Ser Leu Asn Ser Arg Glu Gln Ile Thr Thr Glu Asp Val Asp

GAT CGT GTT AGA CAA GTG CTA AAA ATG ATT AAG TTC GTT GTT GAC AAT TTA   867
Asp Arg Val Arg Gln Val Leu Lys Met Ile Lys Phe Val Val Asp Asn Leu

GAG AAA ACA GGT ATT GTG GAG AAT GGC CCA GAA TCT ACT TCA AAC AAC ACC   918
Glu Lys Thr Gly Ile Val Glu Asn Gly Pro Glu Ser Thr Ser Asn Asn Thr

AAG GAA ACC TCG GAC CTG TTG AGA GAG ATT GCT GCT GAC TCT ATT GTT TTA   969
Lys Glu Thr Ser Asp Leu Leu Arg Glu Ile Ala Ala Asp Ser Ile Val Leu

TTG AAG AAC AAA AAC AAT TAT CTT ACC TCT AAA GAA AGA AGA CAA TAT CAT  1020
Leu Lys Asn Lys Asn Asn Tyr Leu Thr Ser Lys Glu Arg Arg Gln Tyr His

GTC ATT GGC CCA AAT GCT AAA GCA AAG ACT AGT TCC GGT GGT GGT TCA GCA  1071
Val Ile Gly Pro Asn Ala Lys Ala Lys Thr Ser Ser Gly Gly Gly Ser Ala

TCT ATG AAC TCC TAC TAT GTT GTT TCT CCG TAT GAA GGT ATC GTC AAT AAG  1122
Ser Met Asn Ser Tyr Tyr Val Val Ser Pro Tyr Glu Gly Ile Val Asn Lys
```

Fig. 3. Nucleotide sequence and deduced amino acid sequence of the β-glucosidase gene. The complete sequence of 4193 nucleotides containing the β-glucosidase gene is shown. Nucleotide +1 corresponds to the A of the translation initiation codon.

In the 5′ non coding region: (. . .) indicates TATA-like sequences. *Large arrows* correspond to the transcriptional starts. (1) and (2) represent direct repeated sequences. In the 3′ non-coding region, stop codons in frame are underlined by ▾, yeast termination consensus sequences are underlined by ▬ and – – –. Putative polyadenylation site is marked (. . .). (3) and (4) indicate hairpin sequences

```
CTG GGC AAA GAG GTC GAT TAC ACC GTA GGG GCC TAT TCA CAC AAA TCG ATT  1173
Leu Gly Lys Glu Val Asp Tyr Thr Val Gly Ala Tyr Ser His Lys Ser Ile

GGA GGT TTG GCA GAG AGT AGT TTG ATC GAT GCT GCA AAA CCA GCA GAT GCT  1224
Gly Gly Leu Ala Glu Ser Ser Leu Ile Asp Ala Ala Lys Pro Ala Asp Ala

GAA AAT GCT GGA TTA ATT GCC AAG TTT TAC TCC AAT CCA GTA GAA GAG AGA  1275
Glu Asn Ala Gly Leu Ile Ala Lys Phe Tyr Ser Asn Pro Val Glu Glu Arg

TCT GAA GAT GAA GAA CCA TTC CAC GTT ACC AAA GTC AAT AGA TCC AAT GTT  1326
Ser Glu Asp Glu Glu Pro Phe His Val Thr Lys Val Asn Arg Ser Asn Val

CAC TTA TTT GAT TTC AAA CAT GAG AAA GTG GAT CCA AAG AAC CCT TAC TTT  1377
His Leu Phe Asp Phe Lys His Glu Lys Val Asp Pro Lys Asn Pro Tyr Phe

TTT GTA ACC TTA ACC GGA CAG TAC GTG CCT CAA GAA GAT GGT GAT TAT ATC  1428
Phe Val Thr Leu Thr Gly Gln Tyr Val Pro Gln Glu Asp Gly Asp Tyr Ile

TTC AGT CTT CAA GTT TAT GGT TCT GGT TTG TTC TAC TTA AAC GAT GAG TTG  1479
Phe Ser Leu Gln Val Tyr Gly Ser Gly Leu Phe Tyr Leu Asn Asp Glu Leu

ATT ATT GAC CAA AAG CAC AAC CAA GAA AGG GGT AGT TTC TGC TTT GGA GCT  1530
Ile Ile Asp Gln Lys His Asn Gln Glu Arg Gly Ser Phe Cys Phe Gly Ala

GGT ACC AAA GAA AGA ACC AAA AAG TTG ACT TTG AAG AAG GGC CAA GTT TAT  1581
Gly Thr Lys Glu Arg Thr Lys Lys Leu Thr Leu Lys Lys Gly Gln Val Tyr


AAT GTA AGA GTT GAG TAC GGT TCT GGC CCA ACT TCA GGT TTG GTT GGG GAA  1632
Asn Val Arg Val Glu Tyr Gly Ser Gly Pro Thr Ser Gly Leu Val Gly Glu

TTC GGT GCA GGT GGA TTC CAA GCT GGT GTC ATT AAG GCG ATC GAT GAT GAC  1683
Phe Gly Ala Gly Gly Phe Gln Ala Gly Val Ile Lys Ala Ile Asp Asp Asp

GAG GAG ATT AGA AAC GCA GCA GAA TTA GCA GCT AAG CAT GAT AAG GCT GTG  1734
Glu Glu Ile Arg Asn Ala Ala Glu Leu Ala Ala Lys His Asp Lys Ala Val

TTG ATA ATT GGA TTA AAT GGT GAA TGG GAA ACC GAA GGT TAT GAC AGA GAA  1785
Leu Ile Ile Gly Leu Asn Gly Glu Trp Glu Thr Glu Gly Tyr Asp Arg Glu

AAC ATG GAT TTG CCA AAA AGA ACA AAT GAA TTA GTT CGT GCT GTT TTG AAA  1836
Asn Met Asp Leu Pro Lys Arg Thr Asn Glu Leu Val Arg Ala Val Leu Lys

GCA AAT CCA AAT ACT GTT ATC GTT AAC CAA TCA GGT ACC CCA GTT GAG TTC  1887
Ala Asn Pro Asn Thr Val Ile Val Asn Gln Ser Gly Thr Pro Val Glu Phe

CCT TGG TTA GAA GAG GCA AAT GCG CTA GTT CAA GCT TGG TAC GGT GGT AAT  1938
Pro Trp Leu Glu Glu Ala Asn Ala Leu Val Gln Ala Trp Tyr Gly Gly Asn

GAA TTG GGT AAT GCT ATC GCA GAT GTC TTG TAC GGT GAC GTG GTT CCA AAT  1989
Glu Leu Gly Asn Ala Ile Ala Asp Val Leu Tyr Gly Asp Val Val Pro Asn

GGT AAG TTA TCG CTC TCT TGG CCA TTT AAG TTG CAA GAT AAT CCA GCC TTT  2040
Gly Lys Leu Ser Leu Ser Trp Pro Phe Lys Leu Gln Asp Asn Pro Ala Phe

TTA AAC TTC AAG ACC GAG TTC GGA AGA GTT GTT TAC GGT GAG GAT ATC TTT  2091
Leu Asn Phe Lys Thr Glu Phe Gly Arg Val Val Tyr Gly Glu Asp Ile Phe

GTT GGT TAT AGG TAC TAC GAA AAG CTT CAA AGA AAG GTA GCC TTC CCC TTC  2142
Val Gly Tyr Arg Tyr Tyr Glu Lys Leu Gln Arg Lys Val Ala Phe Pro Phe

GGA TAT GGT CTA TCG TAT ACA ACA TTC GAA CTA GAT ATT TCT GAC TTC AAG  2193
Gly Tyr Gly Leu Ser Tyr Thr Thr Phe Glu Leu Asp Ile Ser Asp Phe Lys

GTA ACT GAT GAT AAG ATA GAT ATT TCA GTT GAT GTG AAG AAT ACT GGT GAT  2244
Val Thr Asp Asp Lys Ile Asp Ile Ser Val Asp Val Lys Asn Thr Gly Asp

AAA TTT GCT GGC TCC GAG GTG GTG CAA GTC TAC TTC AGC GCT CTA AAC TCT  2295
Lys Phe Ala Gly Ser Glu Val Val Gln Val Tyr Phe Ser Ala Leu Asn Ser

AAG GTC TCG AGA CCG GTT AAG GAG TTG AAG GGA TTC GAA AAA GTT CAT TTG  2346
Lys Val Ser ArG Pro Val Lys Glu Leu Lys Gly Phe Glu Lys Val His Leu

GAA CCA GGT GAG AAG AAG ACA GTT AAT ATT GAA CTA GAA TTG AAA GAT GCA  2397
Glu Pro Gly Glu Lys Lys Thr Val Asn Ile Glu Leu Glu Leu Lys Asp Ala

ATT TCC TAC TTT AAC GAA GAG CTC GGT AAA TGG CAC GTT GAA GCA GGT GAA  2448
Ile Ser Tyr Phe Asn Glu Glu Leu Gly Lys Trp His Val Glu Ala Gly Glu

TAC TTG GTT TCA GTT GGT ACT TCT TCT GAT GAT ATA CTT TCC GTC AAA GAG  2499
Tyr Leu Val Ser Val Gly Thr Ser Ser Asp Asp Ile Leu Ser Val Lys Glu

TTT AAA GTA GAA AAA GAC TTG TAC TGG AAA GGT TTG TGA   AGAATGCTAAAATG 2552
Phe Lys Val Glu Lys Asp Leu Tyr Trp Lys Gly Leu ▼
```

Fig. 3 (continued)

Fig. 3 (continued)

to begin at the first AUG as no other AUG is present further upstream and only one other in-frame AUG codon is present further downstream (at position +256). A deletion of a *Bgl*II-*Pst*I fragment (−56 to +72) leads to the loss of β-glucosidase activity both in *Escherichia coli* and *Saccharomyces cerevisiae*.

The molecular weight of the β-glucosidase calculated from the predicted 845 amino acids sequence is 93,811 Kd. This result is in good agreement with the molecular weight of 325,000 Kd reported by Fleming and Duerksen (1967) for the β-glucosidase of *Kluyveromyces fragilis* since they suggested that this enzyme was made of four identical subunits of about 80,000 Kd. More recently the molecular weight of the purified enzyme from the *Kluyveromyces fragilis* strain used for the cloning experiments and from that produced in *Saccharomyces cerevisiae* have been estimated at 320,000 Kd. In both cases the molecular weight of the subunit is about 90,000 Kd (Arnaud et al. in preparation). In addition Marchin and Duerksen (1968) proposed a molecular weight of about 100,000 Kd for the subunit of the *Kluyveromyces lactis* β-glucosidase which is also a tetramer. A sequence which could be related to a transcription termination, polyadenylation signal can be found, TAAATAAT, at position +2963 to +2970. In addition two consensus sequences similar to those described by Zaret and Sherman (1982) are present in the 3′ non-coding region. These observations place the probable end of the β-glucosidase mRNA around position +2950.

Transcription of the β-glucosidase gene

Total RNA was extracted both from *Kluyveromyces fragilis* and *Saccharomyces cerevisiae* transformed by KF21. Poly(A)$^+$ RNAs were purified and used for transcription analysis. The direction of transcription was previously determined by using DNA-RNA hybridization with two M13 clones carrying opposite strands of *Bgl*II-*Eco*RI fragment (see Fig. 2). It is the same as that determined from the sequence data. The size of the transcripts was determined by using M13 probes 1 and 2 (Fig. 2). Northern analysis showed only one transcript of the same size both in the original strain *Kluyveromyces fragilis* and in the expression host *Saccharomyces cerevisiae* transformed with KF21 plasmid. The size of this transcript appears to be about 3 kb (Fig. 4). These results suggest that the transcription signals could be the same for the β-glucosidase gene both in the *Kluyveromyces fragilis* genome and expressed in *Saccharomyces cerevisiae* on multicopy plasmids.

The 5′ end of the β-glucosidase transcript was mapped using S1 nuclease protection experiments. The probe used was a 900 bp *Bam*HI-*Pst*I fragment (see Fig. 2) labeled on the coding strand. The *Pst*I site is

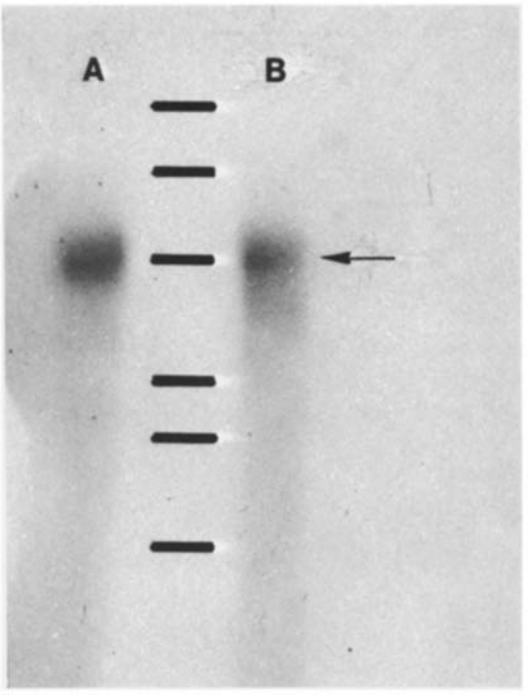

Fig. 4. Northern blot analysis of the β-glucosidase transcripts. *Lane A:* 0.5 μg poly(A)$^+$ RNA from *Saccharomyces cerevisiae* transformed by KF21. *Lane B:* 7 μg poly(A)$^+$ RNA from *Kluyveromyces fragilis*. *Arrow* indicates the position of a molecular weight marker of 3.05 kb. Poly(A)$^+$ RNAs were electrophoresed on a 1% agarose gel, transferred to a nylon membrane and hybridized with probe 2

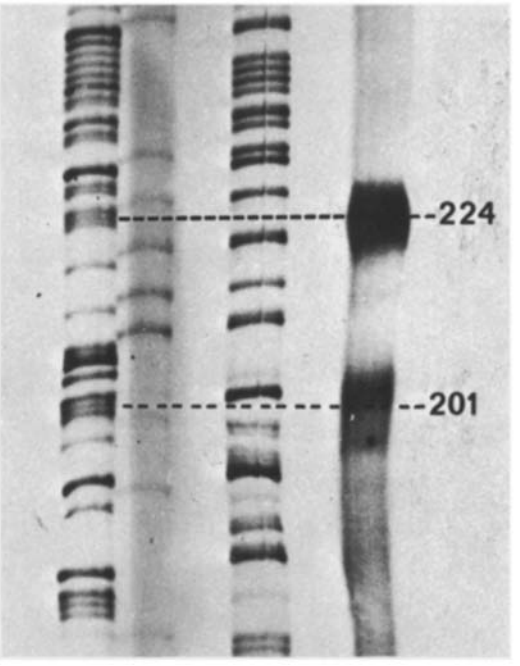

Fig. 5. S1 mapping experiments. Poly(A)$^+$ RNA (20 μg) extracted from *Kluyveromyces fragilis*. After digestion with S1 nuclease (500 U/ml), the nuclease-resistant hybrids were analysed on a sequencing gel. Molecular weight DNA markers of a known sequence were run on the same gel. Coordinates (201, 224) of the 2 major bands are given starting from the *Pst*I site at position +72. Relative to the ATG these coordinates correspond respectively to positions −129 and −152

located 72 nucleotides downstream from translation start point. This probe was hybridized with Poly(A)$^+$ RNA purified from *Kluyveromyces fragilis*. After digestion of the DNA/RNA hybrid with S1 nuclease, the resulting S1-resistant fragments were run on a sequencing gel in parallel with the sequence of a known fragment. From this analysis we can observe two transcription start points located at positions −153/−151 and −130/−128 relative to the beginning of the coding region (Fig. 5). Taking in to account the start point and the size of the mRNA (3.1 kb), the termination of the transcription could be located around +2950.

Discussion

Sequence features

From the nucleotide sequence data we have postulated the translation start at position +1 as it is the beginning of the unique long open reading frame. This is also supported by the presence of the invariant A at position −4 and a purine (U) at position +4 found around initiation codons of other eucaryotic mRNAs (Kozak 1984). The start located at position −151/−153 fits with a sequence AATAA. This result is in good agreement with those described by Hahn et al. (1985). They proposed two consensus sequences accounting for about 55% of all yeast transcription initiation sites (TC(G/A)A or PuPuPyPuPu).

Four presumed TATA boxes could be observed in the 5′ region of the β-glucosidase gene at positions −296 to −292, −270 to −257, −236 to −232 and −207 to −200. In yeast promoter regions there are frequently multiple TATA-like sequences but it is generally not clear which, if any, of these sequences are functionally important. In addition the distances between the transcriptional starts and the TATA elements in yeast genes are more variable and usually longer than in higher eucaryotes. Chen and Struhl (1985) suggested that TATA elements could act at distances ranging from 40 to 90 bp upstream of the transcription start site. The transcriptional start points we have determined agree with these observations and thus the two TATA-like sequences starting at positions −236 and −207 would be the most probable.

Transcription termination and polyadenylation signals have not yet been established in yeast. From a

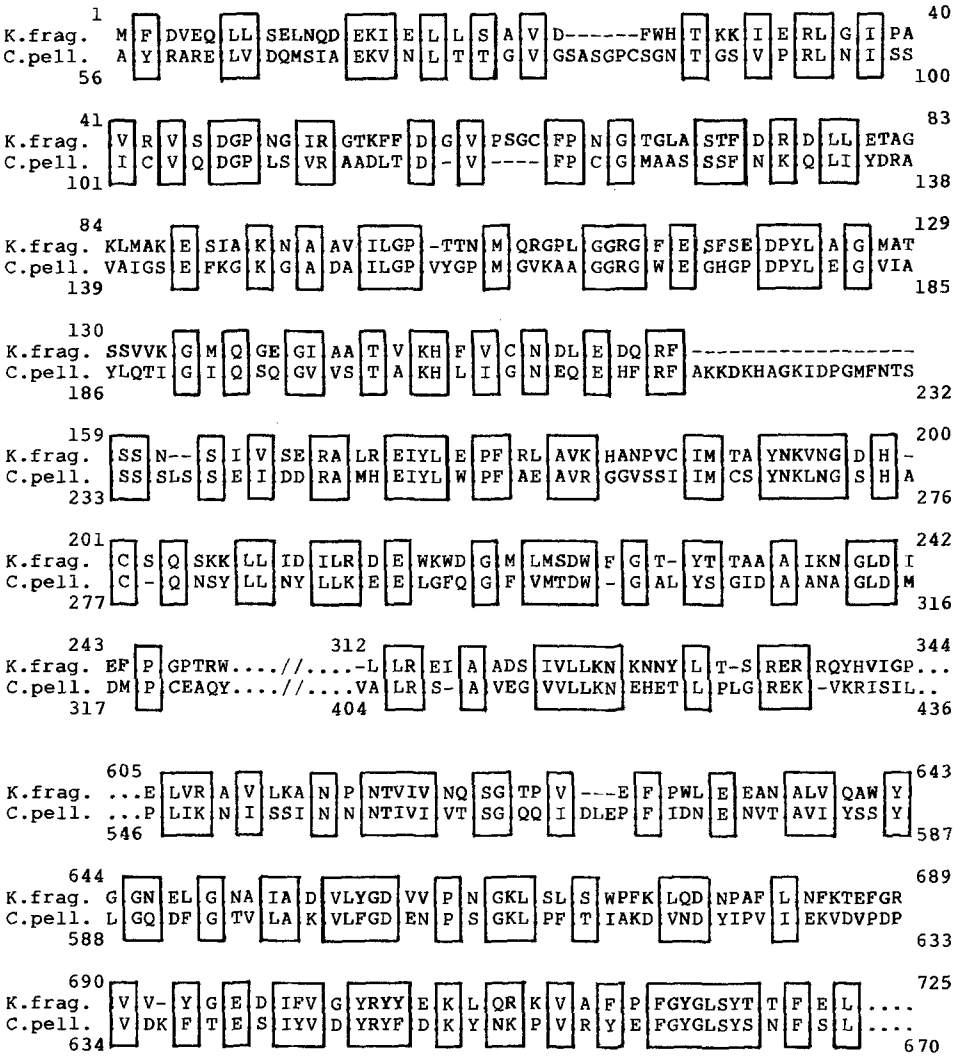

Fig. 6. Homologous regions of the β-glucosidases of *Kluyveromyces fragilis* and *Candida pelliculosa*. Numbering of the residues in the *Candida pelliculosa* sequence is that of the putative primary translation product (Kohchi and Toh-e 1985). *Boxes* indicate identical amino acid positions or those corresponding to an accepted amino acid substitution (Dayhoff 1978). Sequences have been aligned to maximize homology by introducing gaps indicated by a dash (–)

comparison of over 15 yeast genes, a consensus sequence TAG....TAGT..(ATrich)..TTT has been proposed as a transcription termination signal (Zearet and Sherman 1982; Birnsteil et al. 1985). In addition a TAAATAA(T/A) consensus sequence has been suggested as a polyadenylation signal in yeast (Bennetzen and Hall 1982). This consensus sequence related to the AATAAA polyadenylation signal of higher eucaryotic genes remains controversial.

Comparison with published β-glucosidase amino acid sequences

We compared the amino acid sequence of the β-glucosidase of *Kluyveromyces fragilis* to that of *Candida pelliculosa* (Kohchi and Toh-e 1985) by using the NBRF programs "Align" and "Relate" implanted in the Centre Interuniversitaire de Traitement de l'Information (C.I.T.I.2 Paris). In both cases the scoring matrix was the mutation data matrix (Schwartz and Dayhoff 1979). The amino acid sequence of the β-glucosidase of *Candida pelliculosa* was deduced from the nucleotide sequence of the gene (Kohchi and Toh-e 1985). It shows a smaller number of amino acids: 805 *vs* 845 for that of *Kluyveromyces fragilis* and a potential signal peptide of 20 amino acids. Three domains showing similarities could be described (Fig. 6). The first similarity is found in the N-terminal part of the two proteins. These regions are about 250 amino acids in length and the homology observed was 42%. The second similarity (25 amino acids, 62% homology) is found in

```
common CS   **  *   * **** * * *  *        *     *  *   **  *  *******  ****
S.comm.     -LVQAVAD NENTIVAVNTVGAII---TEAWIEHPNKAVVWSGLPGQEAGNSVADILYGAYNPSG
Common SK    ** **   * **** **  *  *      ***  *  *      *** ** *******   * *
K.frag.     ELVRAVLKANPNTVI-VNQSGTPV---EFPWLEEANALV-QAWYGGNELGNAIADVLYGDVVPNG
Commun KC    *** *   * **** *  **  *    *   *   ***    * **  *  ** *****  * *
C.pell.     PLIKNISSINNNTIV-IVTSGQQIDLEPPFIDNENVTAVI-YSSYLGQDFGTVLAKVLFGDENPSG
common 3     **  *   * **** *   *  *        *     *      **  *  ** ***    * *

common CS   ********   **   *         *  *** * **** **      ** *
S.comm.     RLPYTIAKSADDYPAQVLYESSAQVPDIDYSEGLLVDYRHFDANGIEPRFGF
common SK   **  *      * ** * * *       * * * * ** *        * *
K.frag.     KLSLSWPFKLQDNPAFLNFKTEFGRVV--YGEDIFVGYRYYEKLQRKVAFPFGYGLSYTTFELDIS
common KC   **  *    ***    *       *   * * *** **** * ** * * ******** * ***
C.pell.     KLPFTIAKDVNDYIPVIEKVDVPDPVDK-FTESIYVDYRYFDKYNKPVRYEFGYGLSYSNFSLDIEL
common 3    ** *      *    *           * * * * ** *        * *
```

Fig. 7. Homologies between the C-terminal part of the β-glucosidases of *Kluyveromyces fragilis*, *Candida pelliculosa* and the published partial sequence of the β-glucosidase of *Schizophyllum commune*. *Stars* indicate identical amino acid positions or those corresponding to an accepted amino acid substitution. "Common KC" indicates identity between β-glucosidase of *Kluyveromyces fragilis* and *Candida pelliculosa*. "Common SK" indicates identity between β-glucosidase of *Schizophyllum commune* and *Kluyveromyces fragilis*. "Common CS" indicates identity between β-glucosidase of *Candida pelliculosa* and *Schizophyllum commune*. "Common 3" indicates identity between the three enzymes

```
                <-------------------->
K.fragilis  D E W K W D G M L M S D W F G T Y T T A A A I K N G L D I E F P
common AK     *         *   * * * * *                       * *
A.wentii    A E L G F Z G F V M S D W A A H H A G V S G A L A G L N M G S M
common AC     * * * * * * * * * * *   *       * *         * * *   *
C.pellicu.  E E L G F Q G F V M T D W G A L Y S G I D A A N A G L D M D M P
common 3      *         *   * * * * *                       * *
```

Fig. 8. Similarity of the β-glucosidases of *Kluyveromyces fragilis* and *Candida pelliculosa* with the peptide from the active site of β-glucosidase A_3 from *Aspergillus wentii*. Stars indicate identical amino acid positions or those corresponding to an accepted amino acid substitution. "Common AK" indicates identity between β-glucosidase of *Aspergillus wentii* and *Kluyveromyces fragilis*. "Common AC" indicates identity between β-glucosidase of *Aspergillus wentii* and *Candida pelliculosa*. "Common 3" indicates identity between the three enzymes. The residues GLU and ASP thought to have a catalytic function are indicated by <– – –>

the middle of the two sequences. The third region is in the C-terminal part of the proteins (120 amino acids, 50% homology). Moranelli et al. (1986) have published a partial sequence of the β-glucosidase of *Schizophyllum commune*. This partial sequence shows a good homology with the C-terminal corresponding part of the β-glucosidase both of *Kluyveromyces fragilis* and *Candida pelliculosa* (Fig. 7). As β-glucosidases are tetrameric proteins the conserved sequences may be involved in the activity and/or monomer association. Nevertheless using the results of Legler et al. (1979) and Bause and Legler (1980) a putative active site sequence for β-glucosidase could be proposed. These authors observed that D-glucal and conduritol B epoxide, inhibitors analogous to the substrate, bound to the Asp in the sequence Val-Met-Ser-Asp-Trp in *Aspergillus wentii* β-glucosidase A_3. They reported also an amino acid sequence encompassing this putative active site within 63 residues. This sequence was compared to those *Kluyveromyces fragilis* and *Candida pelliculosa*. Two conclusions can be drawn from this comparison. First, the amino acids sequence around the aspartic acid binding to the inhibitors is highly conserved. Second, by analogy with lysozyme, the active site of β-glucosidase could be supposed to be composed of the aspartic acid previously described and of an upstream glutamic acid. As shown in Fig. 8 such a glutamic acid is found in the three sequences 9 amino acids upstream from the inhibitor-binding aspartic acid. As β-glucosidase (EC 3.2.1.21) and lysozyme (EC 3.2.1.17) are both glycosyl hydrolases these observations would imply that the active site lies between amino acids 200 and 230 of the β-glucosidase of *Kluyveromyces fragilis* and between the amino acids 275 and 305 of the β-glucosidase of *Candida pelliculosa*.

Codon usage

The codon usage in the β-glucosidase gene is summarized in Table 1. Sixty out of the 61 possible codons are used. For yeast genes that are highly expressed, such as alcohol dehydrogenase I the authors observed that more than 90% of the amino acid residues are encoded by 25 of the 61 possible coding triplets. Bennetzen and Hall (1982) observed that a set of 22 preferred codons corresponds to the most abundant isoacceptor yeast tRNA species. They defined a codon bias index on a scale of 1.0 to 0 which shows a good correlation with the level of expression of a gene (0.92 for ADH1, 0.15 for iso-2-

Table 1. Codon usage of the β-glucosidase 845 codon open reading frame

F	TTT	17.0	S	TCT	24.0	Y	TAT	15.0	C	TGT	3.0
F	TTC	24.0	S	TCC	12.0	Y	TAC	17.0	C	TGC	2.0
L	TTA	16.0	S	TCA	10.0	*	TAA	0.0	*	TGA	0.0
L	TTG	34.0	S	TCG	7.0	*	TAG	0.0	W	TGG	11.0
L	CTT	7.0	P	CCT	8.0	H	CAT	8.0	R	CGT	6.0
L	CTC	3.0	P	CCC	2.0	H	CAC	6.0	R	CGC	1.0
L	CTA	7.0	P	CCA	19.0	Q	CAA	18.0	R	CGA	0.0
L	CTG	4.0	P	CCG	2.0	Q	CAG	3.0	R	CGG	1.0
I	ATT	26.0	T	ACT	15.0	N	AAT	27.0	S	AGT	6.0
I	ATC	11.0	T	ACC	15.0	N	AAC	20.0	S	AGC	1.0
I	ATA	4.0	T	ACA	10.0	K	AAA	27.0	R	AGA	21.0
M	ATG	11.0	T	ACG	2.0	K	AAG	39.0	R	AGG	4.0
V	GTT	41.0	A	GCT	26.0	D	GAT	38.0	G	GGT	47.0
V	GTC	10.0	A	GCC	6.0	D	GAC	15.0	G	GGC	8.0
V	GTA	7.0	A	GCA	19.0	E	GAA	44.0	G	GGA	12.0
V	GTG	12.0	A	GCG	4.0	E	GAG	24.0	G	GGG	6.0

cytochrome c). On this basis, the calculated codon bias index for the β-glucosidase gene (0.344) would suggest a moderate expression in *Saccharomyces cerevisiae*. Nevertheless, a deletion in the 5′ non-coding region at position −287 (plasmid KF2, see Fig. 1) leads to a high expression level in *Saccharomyces cerevisiae* transformed with this plasmid (Raynal and Guérineau 1984). The β-glucosidase in these transformants could represent as much as 15% of the total protein. From nucleotide sequence data it is interesting to note that the deletion at position −287 leads to the loss of a palindromic region which is only composed of A and T. Involvement of this sequence at the expression level and study of the promotor organisation are under investigation and will be discussed in a further paper.

Acknowledgements. We wish to thank Gisele Boz for her skillful technical assistance. We thank Doctors M. Cassan, A. Hénaut, and P. Vigier for helpful discussions and suggestions. This work was supported in part by A.T.P. no. 90, 1621 from the C.N.R.S.

References

Ait N, Creuzet N, Cattaneo J (1982) J Gen Microbiol 128:569–577

Armentrout RW, Brown RD (1981) Appl Environ Microbiol 41:1355–1362

Barras F, Chambost JP, Chippaux M (1984) Mol Gen Genet 197:486–490

Bause E, Legler G (1980) Biochim Biophys Acta 626:459–465

Bennetzen JL, Hall BD (1982) J Biol Chem 257:3018–3025

Bennetzen JL, Hall BD (1982) J Biol Chem 257:3026–3031

Birnsteil ML, Busslinger M, Strub K (1985) Cell 41:349–359

Blondin B, Ratomahenina R, Arnaud A, Galzy P (1983) Appl Microbiol Biotech 17:1–6

Boyer HW, Roulland-Dussoix D (1969) J Mol Biol 41:459–472

Boy-Marcotte E, Jacquet M (1982) Gene 20:433–440

Burke JF (1984) Gene 30:63–68

Chen W, Struhl K (1985) EMBO J 12:3273–3280

Duerksen JD, Halverson H (1959) Biochim Biophys Acta 36: 47–55

Fleming W, Duerksen JD (1967) J Bacteriol 96:135–141

Gerbaud C, Elmerich C, Tandeau de Narsac N, Chocat P, Charpin N, Guerineau M, Aubert JP (1981) Curr Genet 3:173–180

Hahn S, Hoar ET, Guarente L (1985) Proc Natl Acad Sci USA 82:8562–8566

Hu N, Messing J (1982) Gene 17:271–277

Kohchi C, Toh-e A (1986) Mol Gen Genet 203:89–94

Kozak M (1984) Nucleic Acids Res 12:857–872

Legler G, Roeser KR, Illig HK (1979) Eur J Biochem 101:85–92

Maccecchini ML, Rubin Y, Blobel G, Schatz G (1979) Proc Natl Acad Sci USA 76:343–347

McMaster GK, Carmichael GG (1977) Proc Natl Acad Sci USA 74:4835–4839

Maniatis T, Fritsch E, Sambrook J (1982) Molecular Cloning. A laboratory manual. Cold Spring Harbor Laboratory Press, New York

Marchin GL, Duerksen JD (1968) J Bacteriol 96:1187–1190

Messing J (1983) Methods Enzymol 101:20–78

Moranelli F, Barbier JR, Dove MJ, MacKay RM, Seligy VL, Yaguchi M, Willing GE (1986) Biochem Int 12:905–912

Pentilla ME, Nevalainen KMH, Raynal A, Knowles JKC (1984) Mol Gen Genet 194:494–499
Raynal A, Guerineau M (1984) Mol Gen Genet 195:108–115
Richter K, Ammerer G, Hartter E, Ruis H (1980) J Biol Chem 255:8019–8022
Sanger F, Nicklen S, Coulson AR (1977) Proc Natl Acad Sci USA 74:5463–5467
Schwartz RM, Dayhoff MO (1979) In: Dayhoff MO (ed) Atlas of protein Sequence and Structure, vol 5. National Biomedical Research Foundation, Washington D.C., pp 353–358
Thomas M, McCrae SI (1982) J Gen Microbiol 41:1355–1362
Thomas PS (1980) Proc Natl Acad Sci USA 77:5201–5206
Zaret KS, Sherman F (1982) Cell 28:563–573

Communicated by P. P. Slonimski

Received January 21, 1987