JOURNAL OF BACTERIOLOGY, Oct. 1987, p. 4621–4629
0021-9193/87/104621-09$02.00/0

Vol. 169, No. 1

Cloning and Sequencing of the Alkaline Extracellular Protease Gene of *Yarrowia lipolytica*

LANCE S. DAVIDOW,* MICHELE M. O'DONNELL, FRANK S. KACZMAREK, DENNIS A. PEREIRA, JOHN R. DEZEEUW, AND ARTHUR E. FRANKE

Pfizer Central Research, Groton, Connecticut 06340

Received 9 March 1987/Accepted 20 July 1987

The *XPR2* gene encoding an alkaline extracellular protease (AEP) from *Yarrowia lipolytica* was cloned, and its complete nucleotide sequence was determined. The amino acid sequence deduced from the nucleotide sequence reveals that the mature AEP consists of 297 amino acids with a relative molecular weight of 30,559. The gene codes for a putative 22-amino-acid prepeptide (signal sequence) followed by an additional 135-amino-acid propeptide containing a possible N-linked glycosylation site and two Lys-Arg peptidase-processing sites. The final Lys-Arg site occurs at the junction with the mature, extracellular form. The mature protease contains two potential glycosylation sites. AEP is a member of the subtilisin family of serine proteases, with 42.6% homology to the fungal proteinase K. The functional promoter is more than 700 base pairs long, allowing for the observed complex regulation of this gene. The 5′ and 3′ flanking regions of the *XPR2* gene have structural features in common with other yeast genes.

The dimorphic yeast *Yarrowia lipolytica* has been studied for its alkane utilization (2), its lysine metabolism (8), and its secretion of several relatively large proteins, including an alkaline protease (20, 27), an RNase (4), and several acid proteases (29). Recently, DNA-mediated transformation systems have been developed for *Y. lipolytica* (5, 7) on the basis of homologous integration of vectors containing selectable markers. The complete nucleotide sequence of only one *Y. lipolytica* gene, the isopropylmalate dehydrogenase (*LEU2*) gene, has been reported (6). The *LEU2* gene has been used as the selective marker in shuttle vectors and contains many features common to *Saccharomyces cerevisiae* genes and other eucaryotic genes.

The alkaline extracellular protease (AEP) is the major protein secreted by most *Y. lipolytica* strains examined. Between 1 and 2 g of protease per liter has been obtained from cultures grown to a high density, and the purified enzyme has been biochemically characterized for pH optimum and inhibition by typical serine protease inhibitors (27). The enzyme is classified within the subtilisin family (EC 3.4.21.14) of serine proteases. The molecular weight of the AEP has been estimated to be 28,000 to 31,000 by various physical methods (20, 27). The amino acid sequence of the N-terminal 25 residues of the mature extracellular enzyme has been determined (20). Mutants unable to secrete protease fell into 11 different complementation groups (19). One of these complementation groups was the AEP structural gene, *XPR2* (26).

In this paper we describe the cloning and DNA sequencing of the alkaline protease gene and the deduced AEP amino acid sequence. We have also established the extent of upstream DNA required for independent expression of the gene. The secretory signals of the precursor protein have been identified. The analysis of the *XPR2* gene is an important step toward engineering vectors that will allow *Y. lipolytica* to secrete foreign proteins.

* Corresponding author.

MATERIALS AND METHODS

Strains and plasmids. The strains and plasmids used in this work are shown in Table 1 and Fig. 1, respectively.

Media. YPD-rich medium for routine culturing and defined medium for selection have been described previously (5). Skim-milk plates (19) were used to detect protease-secreting colonies by the formation of zones of clearing. Glycerol proteose-peptone liquid medium was used to grow cells induced for AEP production (20). To select for biotin prototrophy and against *bio-6* mutants, desthiobiotin (25 μg/ml) was added to minimal medium lacking biotin.

***Y. lipolytica* molecular biology.** Transformation of *Y. lipolytica* and chromosomal DNA isolation have been described previously (5). The construction of a gene library of *Sau*3AI partially digested *Y. lipolytica* DNA in the *LEU2*-containing vector pLD40 was similar to the construction of a library in pBR322 (5). The *Sau*3AI partially digested NRRL Y-1094 wild-type DNA was size fractionated on agarose gels, and the 3- to 15-kilobase (kb) range was used in a ligation reaction with *Bam*HI-digested pLD40. Approximately 25,000 independent *Escherichia coli* colonies, containing a total of 45,000 to 60,000 kb of insert *Y. lipolytica* DNA, resulted from the bacterial transformation with the ligation mix. Mixed plasmid DNA, prepared from bacteria harvested from the ampicillin selection plates, served as the source of library DNA for use in *Y. lipolytica* transformations.

DNA sequencing. Nucleotide sequence analysis was performed on overlapping restriction fragments prepared from pLD58, pLD84, pLD86, and pLD108 and their subclones. The restriction fragments were either 5′ end labeled with [γ^{32}P]ATP and polynucleotide kinase or 3′ end labeled with [α^{32}P]ddATP and terminal transferase, isolated from 5% polyacrylamide gels by electroelution, and sequenced by the chemical degradation method (14). DNA sequence analysis was aided by computer programs (IntelliGenetics, Inc., Mountain View, Calif.).

TABLE 1. *Y. lipolytica* strains used

Strain	Relevant genotype or description	Source or reference
ATCC 20688[a]	*MATA leu2-35 ura3-11*	5
ATCC 20774[a]	*MATB leu2-40 xpr2-1002 bio-6*	This work
ATCC 20781[a]	Leu$^+$ Xpr$^+$ transformant of ATCC 20774	This work
ATCC 20794[a]	ATCC 20774 transformed with pLD56 (*BIO*)	This work
NRRL Y-1094	Wild type	NRRL[b]

[a] Deposited with the American Type Culture Collection, Rockville, Md., under the terms of the Budapest Treaty.
[b] Northern Regional Research Center, Peoria, Ill.

RNA studies. Poly (A)$^+$ RNA isolation, glyoxal RNA blots, and avian myeloblastosis virus reverse transcriptase primer extension techniques have been described previously (6). The 15-mer oligonucleotide used for primer extension, 5'-ATAGTAAAGGCGGTA-3', is complementary to a region beginning 12 base pairs (bp) into the structural gene. The 21-mer used as a probe for the RNA blot, 5'-ACAACGATGAACTGATCCTTC-3', is complementary to a region beginning 92 bp into the structural gene.

RESULTS

Gene library construction and screening. To clone *XPR2*, the structural gene for the secreted alkaline protease, a library was constructed consisting of *Sau*3A DNA fragments of wild-type *Y. lipolytica* (strain NRRL Y-1094) inserted into the *Bam*HI site of the *LEU2*-containing vector pLD40 (5). Knowledge of the sequence of the *LEU2* selectable marker was used to choose five restriction enzymes (*Apa*I, *Bgl*II, *Bst*XI, *Nco*I, and *Xho*I) that cut only in the *LEU2* region of pLD40, but not in the pBR322 region. Cut DNA is necessary to obtain high transformation frequencies (1,000 to 100,000 transformants per μg of DNA per 10^8 cells), and also to direct integration (target) to chromosomal regions homologous to the cut region, as was first discovered for *S. cerevisiae* (21). This approach was previously used to clone the *URA3* gene with a *leu2 ura3* double mutant recipient and *Apa*I-cut library DNA (L. S. Davidow, D. Apostolakos, I. Stasko, and J. R. DeZeeuw, *in Molecular Biology of Yeast 1985*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., abstract, p. 63, 1985).

To screen the library for the *XPR2* gene, the recipient strain, ATCC 20774, (*leu2 xpr2 bio*) was transformed to leucine independence and then replica plated to skim-milk

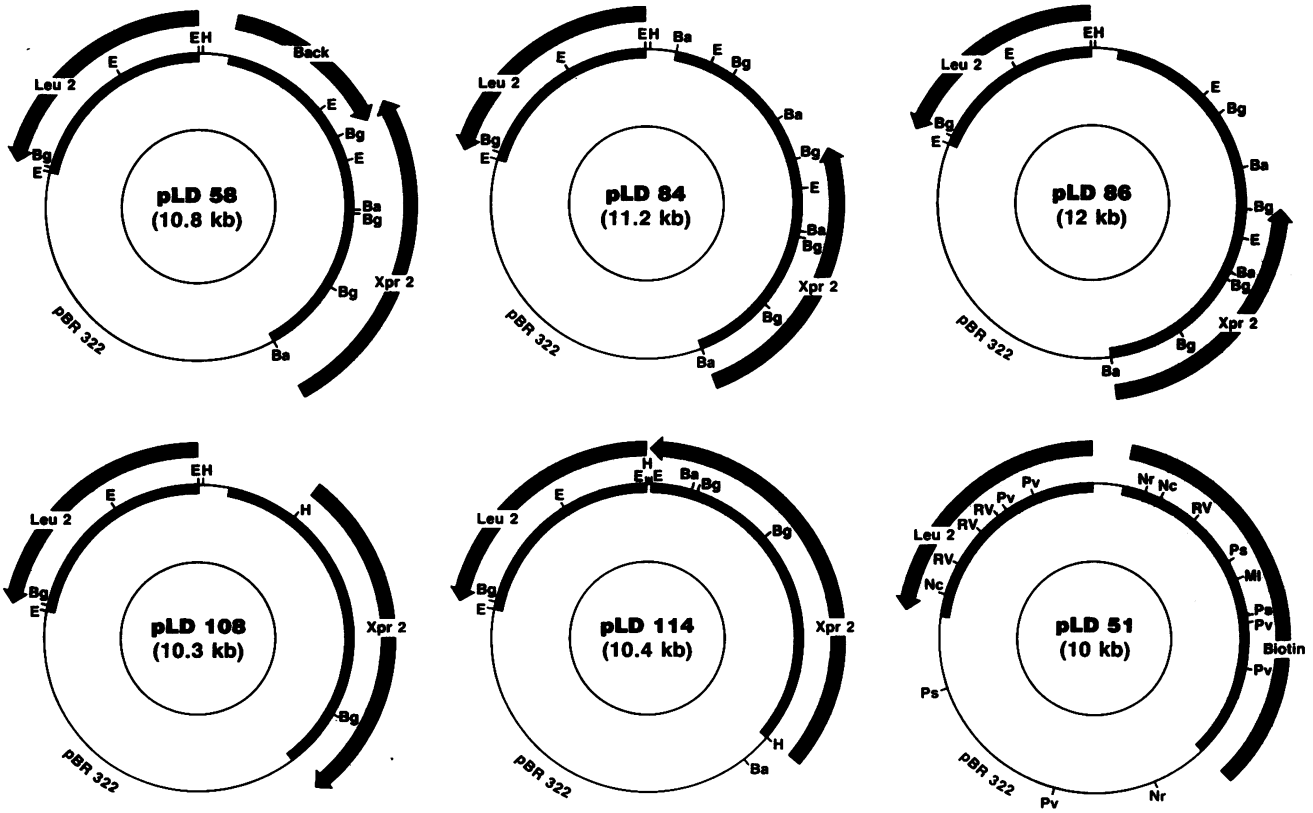


FIG. 1. Plasmids used. These plasmids were recovered from the *Y. lipolytica* gene library or *Y. lipolytica* transformants or were subclones of such plasmids. The *Y. lipolytica* inserts and genes contained on the plasmids are as indicated by the thicker lines. pLD108, pLD84, and pLD86 are library plasmids found by colony hybridization to an *XPR2* probe. pLD114 contains the fully functional *XPR2* promoter plus structural gene and was constructed from segments of pLD108 and pLD84. pLD58 was recovered from a *Bgl*II partial digest of DNA from a protease-positive *Y. lipolytica* transformant and contains the *xpr2-1002* allele and an artificial *Bgl*II junction representing the joining of DNA from the left and right sides of the integrated vector. The *Y. lipolytica* DNA following the artificial junction is marked "Back" to indicate that excision and circularization of the DNA reversed its orientation with respect to the *XPR2* gene. pLD51 was recovered (precisely) from an *Apa*I digest of a biotin-independent and leucine-independent transformant that had integrated a library plasmid at *LEU2*. Restriction sites for *Bam*HI (Ba), *Bgl*II (Bg), *Eco*RI (E), and *Hin*dIII (H) are shown for the *XPR2*-containing plasmids. *Mlu*I (Ml), *Nco*I (Nc), *Nru*I (Nr), *Pst*I (Ps), *Pvu*II, (Pv), and *Eco*RV (RV) are indicated for pLD51.

```
probe series I--protein val thr gln trp gly
                DNA  5' GTX ACX CAU TGG GG 3'
               probe 3' CAX TGX GTY ACC CC 5'
                170     CAA TGX GTY ACC CC
                172     CAT TGX GTY ACC CC
                174     CAG TGX GTY ACC CC
                176     CAC TGX GTY ACC CC

probe series II--protein lys lys ala gln thr
                 DNA  5' AAU AAU GCX CAU AC 3'
                probe 3' TTY TTY CGX GTY TG 5'
                 180     TTC TTY CGX GTC TG
                 182     TTT TTY CGX GTC TG
                 184     TTT TTY CGX GTT TG
                 186     TTC TTY CGX GTT TG
```

FIG. 2. Oligonucleotide probes for the alkaline extracellular protease structural gene. The eight mixed 14-mer probes were based on two different regions (beginning at amino acid residues 7 and 18) of the published 25 N-terminal residues of the mature, secreted alkaline protease (20). Abbreviations: X, all four bases; U, purines; Y, pyrimidines. Probes 174 and 180 gave positive signals with the recovered protease clone pLD58.

indicator plates. No protease-positive transformants were found in experiments with library DNA cut to completion with any of the five enzymes listed above, because the *XPR2* gene contains recognition sites for all five enzymes. However, a protease-positive transformant, ATCC 20781, was obtained following transformation of the recipient with *Bgl*II partially digested (22) library DNA.

Recovering the *XPR2*-containing plasmid. Rescue of an integrated transforming plasmid from *Y. lipolytica* chromosomal DNA requires a restriction digestion, ligation, and an *E. coli* transformation for ampicillin resistance, as in the previous cloning of *URA3*. DNA blot analysis of the *XPR2* transformant ATCC 20781 revealed that the strain was not an integrant in the *LEU2* region. The transforming plasmid had integrated at *XPR2*. Complete *Bgl*II digests of ATCC 20781 total DNA allowed the rescue of a plasmid containing a slight deletion of pBR322 and *LEU2* DNA distal to the *Bgl*II site in *LEU2*, as well as some new DNA from the *XPR2* region up to the first *Bgl*II site. Partial *Bgl*II digests of ATCC 20781 bulk DNA allowed the recovery of larger, overlapping plasmids, the largest of which was designated pLD58 (Fig. 1).

Mixed, 14-mer oligonucleotide probes (Fig. 2) based on the known N-terminal amino acid sequence of the extracellular protease (20) were used to demonstrate the presence of the alkaline protease structural gene in the recovered plasmids. Since a 900-bp *Bgl*II fragment hybridized to both probes 174 and 180, it was concluded that this region of the plasmids contained the structural gene. DNA sequencing (described below) of the recovered plasmid pLD58 (Fig. 1) verified this conclusion.

Colony hybridization to obtain the NRRL Y-1094 *XPR2* allele. To be certain of obtaining the unrecombined NRRL Y-1094 positive allele of *XPR2*, we used the recovered protease gene as a colony hybridization probe against the original gene library. *E. coli* colonies containing plasmids designated pLD84, pLD86, and pLD108 (Fig. 1) were among those recovered as hybridizing to a 2-kb *Pvu*I-*Eco*RI fragment (which contained the structural gene) from the plasmid pLD58.

Retransformation as a functional test of *XPR2*-containing plasmids. The functionality of the recovered plasmids was tested by transformation of the original *leu2 xpr2* double-mutant recipient with *Bgl*II partial digests of pLD58, pLD84,and pLD86. For pLD84 and pLD86, the majority of the leucine transformants were protease positive. Therefore these two plasmids contain the wild-type (NRRL Y-1094) allele of *XPR2*. A different result was obtained for pLD58. None of the leucine-independent transformants was protease positive. Presumably, pLD58, which was rescued from the original protease-positive transformant, contained the *xpr2-1002* negative allele.

To be certain that a plasmid-specific effect was not responsible for the apparent protease-negative phenotype of pLD58, identically sized 2-kb *Pvu*I-*Eco*RI fragments from both pLD58 and pLD86 were subcloned, with linkers, into the *Hin*dIII site of the *LEU2* vector pLD40. Transformation experiments again showed that the subclones from pLD58 were protease negative, whereas the subclones from pLD86 were protease positive. Hybrid plasmid constructs, taking restriction fragments from the positive and mutant alleles, localized the defective site as farther 3′ than an *Xba*I site in the region coding for amino acid 13 of the mature protein. DNA sequencing showed a single-base-pair difference in the C-terminal region of the structural gene between the two subclones (Fig. 3, legend).

DNA sequencing. Plasmids containing the recovered *XPR2* gene region were sequenced to localize the structural gene for the alkaline protease and its 5′ and 3′ flanking regions. Since different gene library plasmids supplied the extreme 5′ and 3′ ends of the region sequenced, we performed a DNA blot experiment to verify that an *XPR2* probe hybridized to a unique 3.7-kb *Hin*dIII-*Eco*RI fragment of NRRL Y-1094 chromosomal DNA identical in size to the large, reconstructed *XPR2* insert in pLD114 (Fig. 1). A detailed restriction map and the strategy used for nucleotide sequence analysis of the 4.1 kb of the yeast genomic DNA between a *Hin*dIII site and a *Bgl*II site containing the *XPR2* gene and flanking regions are shown in Fig. 3. The nucleotide sequence of the *XPR2* gene region is shown in Fig. 4.

AEP structural gene. To locate the alkaline protease coding region, we took advantage of the previously determined N-terminal amino acid sequence of the mature enzyme (20).

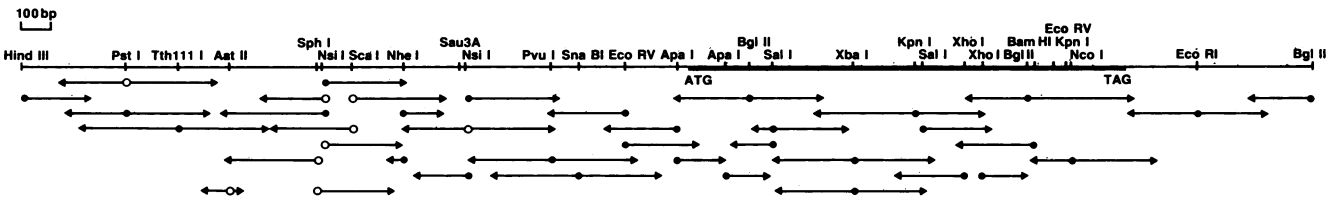
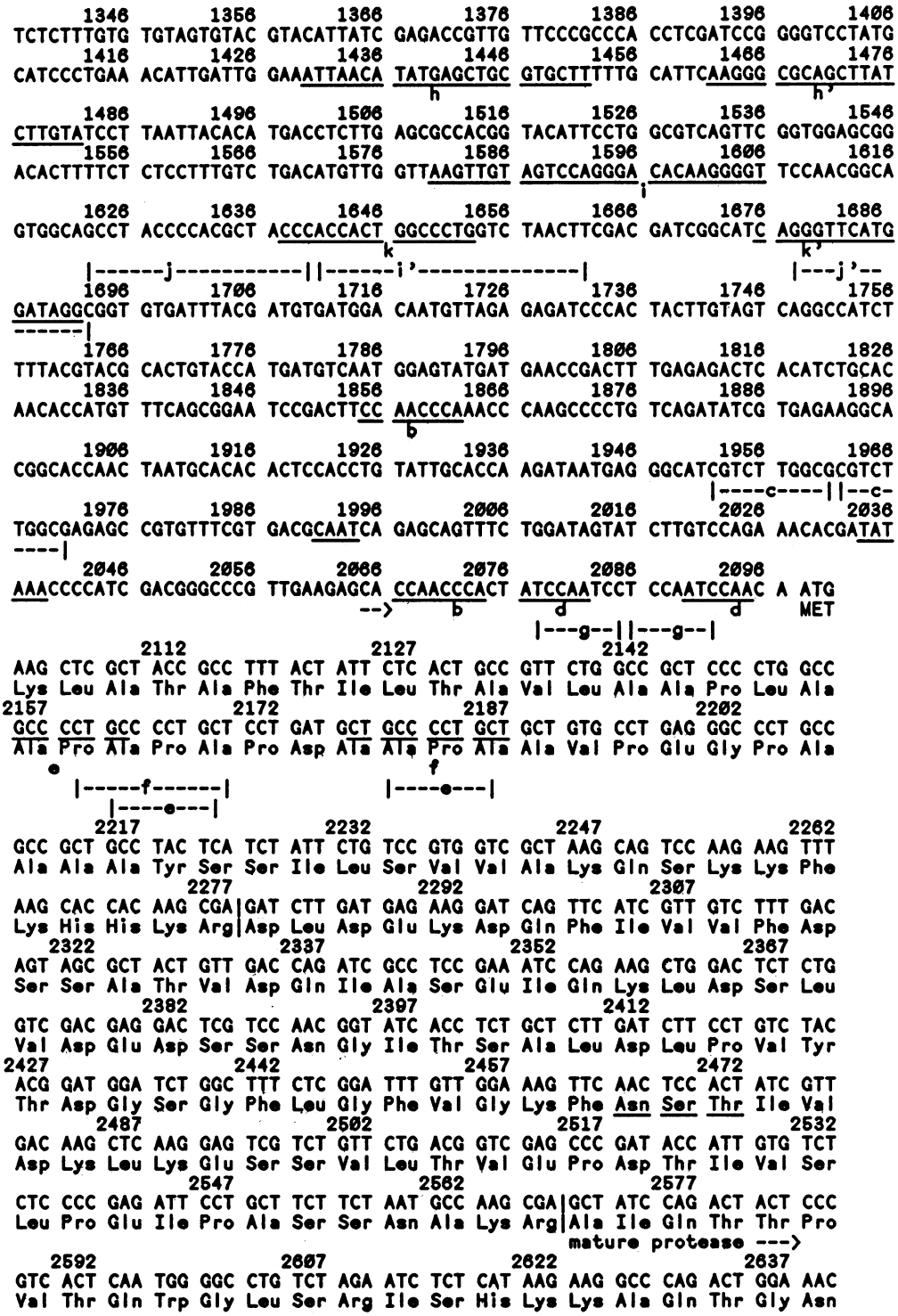


FIG. 3. Restriction endonuclease sites and DNA sequencing strategy. The 4,049 bp of the *XPR2* region of *Y. lipolytica* chromosomal DNA is diagrammed. The sequence was obtained from several overlapping plasmids containing either the wild-type or the *xpr2-1002* allele. The mutant differed from the wild type (shown) by a single-base-pair change (to AGT for serine) at bp 2701 (Fig. 4). ●, →, ←, 5′-labeled termini and the extent of useful sequence information; ○, 3′-labeled termini; ▬, protein-coding region for the preproalkaline extracellular protease gene.

```
                    15                  30                  45
 AAG CTT AGA GTT GAC TAC AAC CGG CTC AAG AAG GAG TGG TGG GTA CAG GAG GAC
 Lys Leu Arg Val Asp Tyr Asn Arg Leu Lys Lys Glu Trp Trp Val Gln Glu Asp
     60                  75                  90                 105
AAG GAG CGG GAC GAC TTT TGG CGA GAC CAA CTT TCC CGA ATC GAA AAA GAC GTG
Lys Glu Arg Asp Asp Phe Trp Arg Asp Gln Leu Ser Arg Ile Glu Lys Asp Val
             120                 135                 150
CAC CGT ACC GAC CGA AAC ATC ACA TTT TTT GCC GAG TGT GAC GCC AAA AAG GAC
His Arg Thr Asp Arg Asn Ile Thr Phe Phe Ala Glu Cys Asp Ala Lys Lys Asp
165                 180                 195                 210
GGG GAT GAT GAC AAC TAC GAT AAG GAT GAG TTT GGG TTT TCG TCC CAG ATA AAC
Gly Asp Asp Asp Asn Tyr Asp Lys Asp Glu Phe Gly Phe Ser Ser Gln Ile Asn
         225                 240                 255                 270
TCC AAC ATT CAT TTG ATC CAG CTC CGT GAC ATG TTG ATT ACC TAC AAC CAG CAT
Ser Asn Ile His Leu Ile Gln Leu Arg Asp MET Leu Ile Thr Tyr Asn Gln His
                 285                 300                 315
AAC AAG AAT CTG GGC TAT GTC CAG GGC ATG TCA GAC CTC CTA TCA CCG CTG TAT
Asn Lys Asn Leu Gly Tyr Val Gln Gly MET Ser Asp Leu Leu Ser Pro Leu Tyr
    330                 345                 360                 375
GTC GTG CTG CAG GAT GAC ACA CTG GCA TTT TGG GCG TTT TCG GCC TTC ATG GAG
Val Val Leu Gln Asp Asp Thr Leu Ala Phe Trp Ala Phe Ser Ala Phe MET Glu
             390                 405                 420
CGC ATG GAG CGA AAC TAC CTC CGG GAC CAG AGT GGC ATG AGA AAC CAG CTT CTT
Arg MET Glu Arg Asn Tyr Leu Arg Asp Gln Ser Gly MET Arg Asn Gln Leu Leu
435                 450                 465                 480
TGT CTG GAC CAT TTG GTC CAA TTT ATG CTT CCC TCA CTG TAC AAG CAC CTT GAG
Cys Leu Asp His Leu Val Gln Phe MET Leu Pro Ser Leu Tyr Lys His Leu Glu
         495                 510                 525                 540
AAG ACC GAG TCA ACC AAT CTG TTT TTC TTC TTC AGA ATG CTG CTG GTG TGG TTC
Lys Thr Glu Ser Thr Asn Leu Phe Phe Phe Phe Arg MET Leu Leu Val Trp Phe
                 555                 570                 585
AAG CGA GAG TTG CTC TGG GAT GAC GTT TTG CGT CTG TGG GAG GTG TTG TGG ACA
Lys Arg Glu Leu Leu Trp Asp Asp Val Leu Arg Leu Trp Glu Val Leu Trp Thr
    600                 615                 630                 645
GAT TAC CTG TCG TCC CAA TTT GTT CTA TTT GTG TGC CTG GCT ATC CTC GAT AAG
Asp Tyr Leu Ser Ser Gln Phe Val Leu Phe Val Cys Leu Ala Ile Leu Asp Lys
             660                 675                 690
CAC AAG GAC GTC ATG ATT GAC CAT CTG GCT GGG TTT GAT GAG ATT CTG AAG TAC
His Lys Asp Val MET Ile Asp His Leu Ala Gly Phe Asp Glu Ile Leu Lys Tyr
705                 720                 735                 750
ATG AAC GAG CTG TCC ATG ACC ATC GAT TTG GAC GAG CTT CTT GTT CGT GCC GAG
MET Asn Glu Leu Ser MET Thr Ile Asp Leu Asp Glu Leu Leu Val Arg Ala Glu
         765                 780                 795                 810
CTC TTG TTC TAC CGA TTC AGA CGT ACG GTC GAG CTT ATT GAC CGA AAG AAC GAG
Leu Leu Phe Tyr Arg Phe Arg Arg Thr Val Glu Leu Ile Asp Arg Lys Asn Glu
                 825                 840                 855
GAC AGA CGC AAC TCA GCG GAC GGC TCC GAG CCT GTT TCC ATC ACA GAG GAC CTG
Asp Arg Arg Asn Ser Ala Asp Gly Ser Glu Pro Val Ser Ile Thr Glu Asp Leu
    870                 885                 900                 915
CGG GAA TTG TTA TCT CGG AAA GTC ATT GTT GTG CGT GAG GGT GAG CGT CCT GAA
Arg Glu Leu Leu Ser Arg Lys Val Ile Val Val Arg Glu Gly Glu Arg Pro Glu
             930                 946         956         986         976          986
GGC GTA ATG GGT GGG TAG GTAATGCAGT TTGCATGCAT GAAGACACTA AACAAGCCAA CCATACAGCA
Gly Val MET Gly Gly .termI
       996         1006        1016        1026        1036        1046        1056
GAAGTATGTA GCCTTGCACA TGATTTATTG ACAGGCCACC CAAACAGGCG TATGTATAGT ACTGTACCTT
    termII              termIII
      1066        1076        1086        1096        1106        1116        1126
CAGTAGACTA TTGTAGCTAA CATGTCGTTG CGTGGCGTAT GTACCAAGCC ACAGAAATTA TGTCAGAGAT
      1136        1146        1156        1166        1176        1186        1196
AAGGTCGCGA CAGTTAGAGC AGCAACGCGT GGAGAGTTTG GGTTTTGGGT TACGTACGTA GAGCCGTTTG
      1206        1216        1226        1236        1246        1256        1266
ATAGATGGTA CATCCACCGG CTAGCGGAAC ACAGTGTCAA GACAAGCCTG CAACACAGTC ATAATATTTG
      1276        1286        1296        1306        1316        1326        1336
CGATATTCAG GCGTATCAGG TACAATCTGA GGTGTCTCAC AAGTGCCGTG CAGTCCCGCC CCCACTTGCT
```

The region of the DNA sequence that encodes this portion of the protease was found with the aid of a computer program. The amino acid sequence predicted from the DNA sequence for the *XPR2* gene product is shown in Fig. 4. Reading back in frame from the first codon of the mature protease, a methionine codon was encountered 157 codons upstream (at bp 2098). This ATG codon was followed by an open reading frame of 453 amino acids. Since all other methionine codons in the several hundred nucleotides immediately preceding this frame were followed shortly thereafter by an in-frame stop codon, we concluded that the methionine codon at bp 2098 probably defines the translational initiation codon of the protease precursor. The context supports this methionine codon as the initiation codon, although unequivocal proof is lacking. As discussed below, the nucleotide sequence upstream of this putative translational start contains structural



features known to be important for the transcription of other yeast genes. In most cases, translation is initiated at the AUG closest to the 5′ end of a eucaryotic mRNA. As shown in our mRNA studies (discussed below), this Met codon is the closest to the 5′ end of the *XPR2* transcript. On the basis of these considerations, the *XPR2* coding region contains 1,362 bp (Fig. 4).

The alkaline protease is synthesized as a precursor protein that is proteolytically processed, possibly in several steps, to the secreted or mature enzyme (S. Matoba, J. Fukayama, and D. Ogrydziak, Yeast 2:S231, 1986). From the translation of the nucleotide sequence, the unprocessed precursor is a polypeptide of 454 amino acids, with a calculated relative molecular weight of 46,942 (Fig. 4). Cleavage of the 157

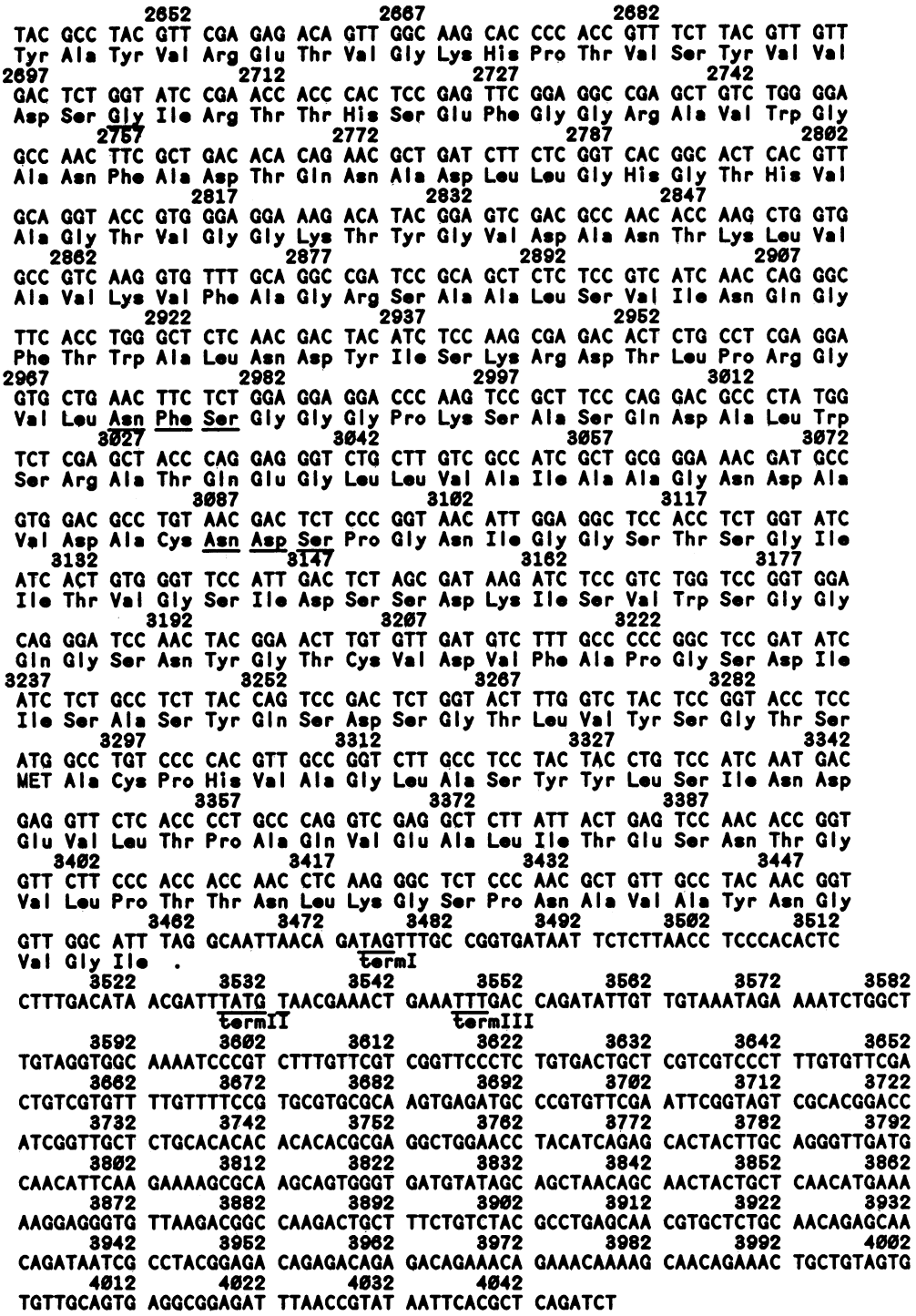


FIG. 4. Nucleotide sequence and amino acid translation of the *Y. lipolytica XPR2* gene including the 3′ end of the hypothesized nearest upstream gene. The preproalkaline extracellular protease-coding region begins at bp 2098. The presumed 5′ TATA box (bp 2034) and CAAT box (bp 1991) are underlined, as are the tripartite components of the presumed transcription terminator (bp 3475 to 3549). The longest transcriptional start is indicated by an arrow (bp 2065). The suspected signal sequence cleavage site is in the Ala-Pro-Ala region (bp 2155 to 2169). Vertical lines show the two Lys-Arg processing sites (bp 2277 and 2568). The three possible N-linked glycosylation sites (bp 2464, 2971, and 3085) are underlined, as is the Gly codon (bp 2701) that is mutated to Ser in *xpr2-1002*. Repeated sequences b to g and possible dyads (imperfect inverted repeats) h to k often overlap, as indicated.

TTT-Phe	7 (1.5)	TCT-Ser	21 (4.6)	TAT-Tyr	0 (.0)	TGT-Cys	3 (.7)
TTC-Phe	6 (1.3)	TCC-Ser	23 (5.1)	TAC-Tyr	13 (2.9)	TGC-Cys	0 (.0)
TTA-Leu	0 (.0)	TCA-Ser	1 (.2)	TAA- .	0 (.0)	TGA- .	0 (.0)
TTG-Leu	1 (.2)	TCG-Ser	2 (.4)	TAG- .	1 (.2)	TGG-Trp	5 (1.1)
CTT-Leu	8 (1.8)	CCT-Pro	10 (2.2)	CAT-His	1 (.2)	CGT-Arg	0 (.0)
CTC-Leu	10 (2.2)	CCC-Pro	11 (2.4)	CAC-His	7 (1.5)	CGC-Arg	0 (.0)
CTA-Leu	1 (.2)	CCA-Pro	0 (.0)	CAA-Gln	1 (.2)	CGA-Arg	9 (2.0)
CTG-Leu	12 (2.6)	CCG-Pro	0 (.0)	CAG-Gln	13 (2.9)	CGG-Arg	0 (.0)
ATT-Ile	8 (1.8)	ACT-Thr	14 (3.1)	AAT-Asn	2 (.4)	AGT-Ser	1 (.2)
ATC-Ile	17 (3.7)	ACC-Thr	16 (3.5)	AAC-Asn	17 (3.7)	AGC-Ser	2 (.4)
ATA-Ile	0 (.0)	ACA-Thr	3 (.7)	AAA-Lys	0 (.0)	AGA-Arg	1 (.2)
ATG-MET	2 (.4)	ACG-Thr	2 (.4)	AAG-Lys	22 (4.8)	AGG-Arg	0 (.0)
GTT-Val	18 (4.0)	GCT-Ala	22 (4.8)	GAT-Asp	12 (2.6)	GGT-Gly	14 (3.1)
GTC-Val	15 (3.3)	GCC-Ala	28 (6.2)	GAC-Asp	17 (3.7)	GGC-Gly	12 (2.6)
GTA-Val	0 (.0)	GCA-Ala	3 (.7)	GAA-Glu	1 (.2)	GGA-Gly	18 (4.0)
GTG-Val	9 (2.0)	GCG-Ala	1 (.2)	GAG-Glu	12 (2.6)	GGG-Gly	0 (.0)

FIG. 5. Codon usage for the *XPR2* gene. The numbers indicate the frequency of use of each codon in the gene. Numbers in parentheses express the frequencies as percentage of total codons.

amino acid prepropeptide from the precursor is then predicted to yield a mature protein of 297 amino acids and a calculated relative molecular weight of 30,559. This value agrees well with the previously reported relative mass estimations of the AEP of 28,000 to 31,000 daltons (20, 27).

Precursors of hormones and other secreted proteins in mammalian (24) or yeast (11) cells are often processed proteolytically after pairs of basic residues. The *KEX2* gene codes for one such processing protease in *S. cerevisiae*. One pair of Lys-Arg residues immediately precedes the mature AEP N terminus. Another potential processing site (Lys-Arg) exists in the prosequence of the AEP precursor 100 amino acids before the N terminus of the mature protein. The putative precursor molecule generated by cleavage at this *KEX2*-like cleavage site would have a calculated relative molecular weight of 40,924. This might explain one of the many precursor species found for AEP.

Analysis of the N-terminal amino acid sequence deduced from the nucleotide sequence suggested the existence of a signal peptide in the protease precursor polypeptide. This signal peptide contains 22 amino acids and has structural features similar to those of higher eucaryotic and procaryotic signal peptides (23). The assignment of a presumptive signal peptide to the first 22 amino acids of the protease precursor was based on the following interpretation, which was consistent with the empirical rules of typical presecretory sequences. A hydrophobic core of 13 amino acids with a predicted β-sheet structure and a repeated element (Thr-Ala) was preceded by a positively charged amino acid (Lys). Core termination was defined by the occurrence of a Pro interrupting the β-sheet structure. We propose the existence of a signal peptidase cleavage site (Fig. 4, coordinate 2163), 6 amino acid residues after the proline and following the most frequently observed recognition signal Ala-X-Ala (where X is any amino acid residue). Since there is also an Ala-X-Ala following and overlapping the proposed signal peptidase site, it is possible that cleavage occurs at this alternative site.

Codon usage. The codon usage in the *XPR2* structural gene (Fig. 5) shows a bias, as others have found for many genes of the yeast *S. cerevisiae* (3). Of the 61 possible codons, 49 are used in the *XPR2* gene; however, 90% of the amino acids are coded for by only 29 codons. A codon bias is evident in the codon representation of several amino acids. For example, the codon AAG for lysine is used 22 times and the codon AAA is not used. For the amino acids arginine and proline, two or more codons are absent: CGU, CGC, and AGG for arginine and CCA and CCG for proline. The codon usage in the *XPR2* gene is not identical with all the codon biases of *S. cerevisiae* genes. The codon bias in *XPR2* is similar to that of the *Y. lipolytica LEU2* gene.

Base composition. The coding region of the *XPR2* gene had a base composition of 19% A, 32% C, 24% G, and 25% T. This distribution of 44% A+T is slightly less than the overall *Y. lipolytica* base composition of 49.6% A+T (13). The relative scarcity of A and abundance of C are surprisingly similar to the distribution seen for the *LEU2* gene. The presumed 3′ fragment of the upstream open reading frame had a very even distribution of bases (24% A, 23% C, 27% G, and 26% T), as did the sequence between the two coding regions (26% A, 26% C, 24% G, and 25% T).

Glycosylation. The purified AEP has been shown to contain no more than 1.8% carbohydrate as determined by the phenol-sulfuric acid method (20). N-linked glycosylation of eucaryotic proteins occurs at the tripeptide sequences Asn-X-Thr and Asn-X-Ser, where X may be any amino acid except possibly aspartate (9). The amino acid sequence of the precursor protease includes three such tripeptide sequences (indicated in Fig. 4), one of which contains Asp as the middle residue. Two of the potential glycosylation sites are in the mature protease coding sequence, and one is in the prosequence region of the structural gene.

Functional unit size. To determine whether the plasmids contained a whole, functioning *XPR2* gene, the plasmids containing the wild-type allele were integrated at a site(s) other than *XPR2*, and transformants were assayed for protease production. Since the plasmids used were derivatives of pBR322, we decided to integrate a copy of pBR322 into the recipient host to serve as a receptor site, or docking platform, for the integration of other plasmids. The biochemically uncharacterized biotin marker in the recipient allowed selective integration of a pBR322-containing plasmid into the *BIO* region of ATCC 20774. Plasmid pLD51 (Fig. 1), one of the *BIO*-containing plasmids recovered from an *Apa*I-cut library-treated transformant, contained no *Eco*RI sites, except those in *LEU2*. Therefore the *LEU2* gene was precisely excised from pLD51 to create pLD56, a plasmid containing only pBR322 and the *BIO* gene at the *Bam*HI site. The unique *Mlu*I site within the *BIO* region was used to target pLD56 for integration into the *BIO* region of ATCC 20774 to create the strain ATCC 20794. The DNA structure of ATCC 20794 was verified, by a DNA blot experiment, to be a single integrant (data not shown). Integrations into the chromosomal pBR322 of ATCC 20794, as well as integrations into the *XPR2* region, were used to determine the promoter and terminator functions of *XPR2* subclones.

Three similar plasmids, designated pLD92, pLD100, and pLD114, containing 428 bp (*Pvu*I), 707 bp (*Sau*3AI), and 2,097 bp (*Hin*dIII), respectively, of DNA before the ATG at position 2098 were compared for *XPR2* promoter function. Each contained its respective 5′ untranslated region, the *XPR2* structural gene, and the 3′ untranslated region through the *Eco*RI site (at 3702 in Fig. 4) cloned (with *Hin*dIII linkers) into pLD40 in the same orientation. To direct integration of each plasmid into the chromosomal copy of pBR322 in ATCC 20794, pLD92 and pLD100 were digested with *Nru*I (which cuts once in the pBR322 region) or *Xmn*I (which makes a gap in the pBR322 region), whereas pLD114 was digested only with *Xmn*I, since the additional 5′ DNA contained a *Nru*I site. Fewer than one in 1,000 leucine transformants from pLD92 and pLD100 formed significant zones of clearing within 48 h of replica plating onto skim-milk plates. Most of the leucine transformants from pLD114 formed zones similar in size to those formed by wild-type NRRL Y-1094 colonies. Most transformant colonies with plasmid pLD100 did form small zones by 4 to 5 days of incubation. We conclude, therefore, that more than 707 bp

(and presumably less than 2,097 bp) of 5′ DNA is necessary to make up a functioning *XPR2* promoter. To test this hypothesis further, gene disruption-type experiments were done at the *XPR2* locus of the *leu2 XPR2*$^+$ strain ATCC 20688. We constructed plasmids containing various lengths of 5′ DNA along with the negative allele, *xpr2-1002*, and directed them to integrate at the chromosomal *XPR2* locus by a *Sna*BI digest. Most of the leucine-independent transformants resulting from plasmids with either 428 or 707 bp of 5′ region became protease deficient, as above. However, few of the leucine transformants resulting from the longest construct became protease deficient. These gene disruption results support the conclusion that between 707 and 2,097 bp of 5′ DNA is part of the functional promoter.

Upstream gene. The 5′-most 933 bp sequenced constitutes an open reading frame that probably represents the neighboring chromosomal gene to *XPR2*. Following the TAG translation termination of this open reading frame, a tripartite transcription terminator (30) is found, as underlined and labeled "term" in Fig. 4. If this is the 3′ end of another gene, as the sequence suggests, we suspect that this DNA will not be involved in *XPR2* regulation.

mRNA studies. The 5′ ends of the mRNA molecules were determined by the reverse transcriptase primer extension method. Three bands were found (data not shown), corresponding to the positions of three different CA pairs, one at the C located 10 bp before ATG, a second at the A located 29 bp before ATG, and the largest band corresponding to the C located 33 bp before ATG. It is possible that some of the bands represent strong reverse transcriptase stops, rather than true ends of the mRNA. An RNA blot experiment gave the size estimation for the mRNA as about 1,525 bp.

***XPR2* gene 5′ flanking sequence.** The 5′ flanking sequence contains structural features known to be important for the transcription of eucaryotic genes. The Goldberg-Hogness, or TATAAA consensus, sequence is generally found 25 to 32 nucleotides upstream from the mRNA start in higher eucaryotes or eucaryotic viruses and appears to be important in the positioning of the transcription start. The 5′ upstream region of *XPR2* contains a TATAAA sequence 65 bp in front of the translational start and 30 bp in front of the primary mRNA start. A second sequence thought to be important for transcription initiation in eucaryotes is the CAAT box, which is located about 80 nucleotides upstream from the site of mRNA synthesis. The *XPR2* gene has a CAAT sequence 73 bp in front of the transcription initiation site (Fig. 4).

The efficiency of ribosome binding to the mRNA is likely to be influenced by the sequence immediately preceding the ATG codon (12). The essential features of the preferred eucaryotic initiation region are a purine, usually an A, at −3 and a purine at +4. These features are observed in the *XPR2* mRNA and similarly in many other known yeast mRNAs, including the *Y. lipolytica LEU2* gene. The hexanucleotide CACACA has been found close to the initiation codon in several yeast genes including the *Y. lipolytica LEU2* gene (6); however, the *XPR2* gene lacks such a sequence. However, the 5′ untranslated mRNA is very A+C-rich (28 of the 33 residues).

The paucity of G residues in the 5′ untranslated region has been noted for many highly expressed *S. cerevisiae* genes (1). There are no G residues in the 5′ untranslated *XPR2* mRNA.

***XPR2* gene 3′ flanking region.** A comparison of the 3′ untranslated region of several *S. cerevisiae* genes revealed a sequence 5′-TAG. . .TA(T)GT. . .TTT-3′ as being important for transcription termination (21). This sequence or a variation occurs at a variable distance before the polyadenylation site of most yeast mRNAs. In the *XPR2* gene, at 66 to 87 nucleotides following the stop codon, a homologous sequence can be recognized (Fig. 4).

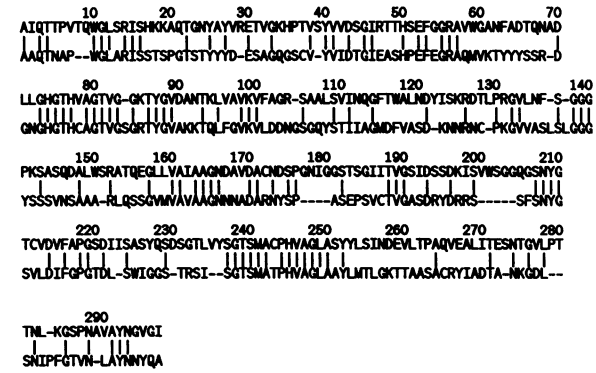


FIG. 6. Homology between the mature AEP (upper lines) and the eucaryotic subtilisin family enzyme proteinase K (lower lines). Vertical lines indicate identical amino acid residues. Regions of maximum homology correspond to the following suspected functional regions (10): a charge relay system composed of Asp-43, His-74, and Ser-240; and regions Ala-163 to Asn-166 and Ser-137 to Gly-139, which are involved in substrate binding. Overall, 118 of the 277 (42.6%) amino acid residues of proteinase K are identical in AEP.

On the basis of the determined transcription start and the proposed location of the polyadenylation site, the calculated size of the *XPR2* mRNA is approximately 1,490 bases plus a poly(A) tail. This length is consistent with the measured mRNA size of 1,525 nucleotides for the poly(A)$^+$ *XPR2* mRNA. The mRNA size, considered together with the sizes of the precursor protease and mature protein, suggests that there are no introns within the *XPR2* coding region.

DISCUSSION

Ogrydziak and Scharf (20) determined the amino acid sequence (25 residues) of the N terminus of the alkaline extracellular protease purified from *Y. lipolytica* CX161-1B. As described above, we have cloned the *Y. lipolytica XPR2* gene and determined its complete nucleotide sequence. A segment of the amino acid sequence deduced from the nucleotide sequence is almost identical to the published partial sequence. The two differences between the sequences occur in residues that were considered only tentatively identified in the sequencing of the protein, and they are therefore more likely to represent uncertainties in the protein sequencing rather than differences in the genes of the two *Y. lipolytica* strains used.

The predicted protein sequence of the mature AEP shows strong homology to other subtilisin family serine proteases, such as 32% homology to *Bacillus subtilis* DY subtilisin (17), 32% homology to *Thermoactinomyces vulgaris* thermitase (15), and 42.6% homology (Fig. 6) with the eucaryotic *Tritirachium album* proteinase K (10). The highest homologies are found in regions previously identified in homologous enzymes (10) as the active-site serine, the active-site histidine, a charge transfer site, and substrate-binding sites. The sequenced mutation that we designate *xpr2-1002* changes an evolutionarily conserved glycine to a serine. The precursor segment of AEP did not bear substantial homology to the precursor segment of *Bacillus amyloliquefaciens* subtilisin (28).

A surprising finding of this study was that the functional *XPR2* promoter region is quite large: greater than 700 bp. The large size of the promoter may be a result of the complex regulation of the gene. The *XPR2* gene is turned on by starvation for nitrogen or sulfur, growth on a poor carbon source, and the presence of extracellular protein (18). An example of an *S. cerevisiae* gene with a large promoter (1,400 bp) showing complex regulation is the *HO* gene for homothallic mating-type switching (16, 25).

Regulatory regions of DNA often involve short repeated sequences or inverted repeats. Computer searches in the 5′ region revealed many examples of such sequences (Fig. 4). A physiological dissection of the *XPR2* promoter is necessary before any functions can be ascribed to these features.

The translated DNA sequence suggests sites as candidates for junctions of a signal peptide, Lys-Arg-terminated precursor peptides, and the mature, secreted alkaline protease. These sites can be used to genetically engineer vectors that direct *Y. lipolytica* to secrete foreign proteins.

ACKNOWLEDGMENTS

We thank Diane Apostolakos for assistance in gene library construction and plasmid subcloning; Irene Stasko for assistance in construction of the *xpr2* mutant strain; Marlene R. Lauth for assistance on RNA work; David Ogrydziak for prepublication data about protease precursors; Glenn Andrews and Lenny Contillo for synthetic oligonucleotides; Alan Proctor, Jean-Marc Nicaud, and Fred Wright for useful discussions; and Gail Welch for graphic arts work.

LITERATURE CITED

1. **Ammerer, G.** 1983. Expression of genes in yeast using the ADC1 promoter. Methods Enzymol. **101**:192–201.
2. **Bassel, J. B., and R. K. Mortimer.** 1982. Genetic and biochemical studies of N-alkane non-utilizing mutants of *Saccharomycopsis lipolytica*. Curr. Genet. **5**:77–88.
3. **Bennetzen, J. L., and B. D. Hall.** 1982. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.
4. **Cheng, S.-C., and D. M. Ogrydziak.** 1986. Extracellular RNase produced by *Yarrowia lipolytica*. J. Bacteriol. **168**:581–589.
5. **Davidow, L. S., D. Apostolakos, M. M. O'Donnell, A. R. Proctor, D. M. Ogrydziak, R. A. Wing, I. Stasko, and J. R. DeZeeuw.** 1985. Integrative transformation of the yeast *Yarrowia lipolytica*. Curr. Genet. **10**:39–48.
6. **Davidow, L. S., F. S. Kaczmarek, J. R. DeZeeuw, S. W. Conlon, M. R. Lauth, D. A. Pereira, and A. E. Franke.** 1987. The *Yarrowia lipolytica LEU2* gene. Curr. Genet. **11**:377–383.
7. **Gaillardin, C., A. M. Ribet, and H. Heslot.** 1985. Integrative transformation of the yeast *Yarrowia lipolytica*. Curr. Genet. **10**:49–58.
8. **Heslot, H., C. M. Gaillardin, J. M. Beckerich, and P. Fournier.** 1979. Control of lysine metabolism in the petroleum yeast *Saccharomycopsis lipolytica*, p. 54–60. *In* O. K. Sebek and A. L. Laskin (ed.), Genetics of industrial microorganisms. American Society for Microbiology, Washington, D.C.
9. **Hubbard, S. C., and R. J. Ivatt.** 1981. Synthesis and processing of asparagine-linked oligosaccharides. Annu. Rev. Biochem. **50**:555–583.
10. **Jany, K.-D., G. Lederer, and B. Mayer.** 1986. Amino acid sequence of proteinase K from the mold *Tritirachium album* Limber. FEBS Lett. **199**:139–144.
11. **Julius, D., A. Brake, L. Blair, R. Kunisawa, and J. Thorner.** 1984. Isolation of the putative structural gene for the lysine-arginine-cleaving endopeptidase required for processing of yeast prepro-alpha-factor. Cell **37**:1075–1089.
12. **Kozak, M.** 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol. Rev. **47**:1–45.
13. **Kurtzman, C. P., H. J. Phaff, and S. A. Meyer.** 1983. Nucleic acid relatedness among yeasts, p. 139–166. *In* J. F. T. Spencer, D. M. Spencer, and A. R. W. Smith (ed.), Yeast genetics. Springer-Verlag, New York.
14. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.
15. **Meloun, B., M. Baudys, V. Kostka, G. Hausdorf, C. Frommel, and W. E. Hohne.** 1985. Complete primary structure of thermitase from *Thermoactinomyces vulgaris* and its structural features related to the subtilisin-type proteinases. FEBS Lett. **183**:195–200.
16. **Nasmyth, K.** 1985. At least 1400 base pairs of 5′-flanking DNA is required for the correct expression of the *HO* gene in yeast. Cell **42**:213–223.
17. **Nedkov, P., W. Oberthur, and G. Braunitzer.** 1983. Die primarstruktur von Subtilisin DY. Hoppe-Seyler's Z. Physiol. Chem. **364**:1537–1540.
18. **Ogrydziak, D. M., A. L. Demain, and S. R. Tannenbaum.** 1977. Regulation of extracellular protease production in *Candida lipolytica*. Biochim. Biophys. Acta **497**:525–538.
19. **Ogrydziak, D. M., and R. K. Mortimer.** 1977. Genetics of extracellular protease production in *Saccharomycopsis lipolytica*. Genetics **87**:621–632.
20. **Ogrydziak, D. M., and S. J. Scharf.** 1982. Alkaline extracellular protease produced by *Saccharomycopsis lipolytica* CX161-1B. J. Gen. Microbiol. **128**:1225–1234.
21. **Orr-Weaver, T. L., J. W. Szostak, and R. J. Rothstein.** 1981. Yeast transformation: a model system for the study of recombination. Proc. Natl. Acad. Sci. USA **78**:6354–6358.
22. **Parker, R. C., R. M. Watson, and J. Vinograd.** 1977. Mapping of closed circular DNAs by cleavage with restriction endonucleases and calibration by agarose gel electrophoresis. Proc. Natl. Acad. Sci. USA **74**:851–855.
23. **Perlman, D., and H. O. Halvorson.** 1983. A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. J. Mol. Biol. **167**:391–409.
24. **Rholam, M., P. Nicolas, and P. Cohen.** 1986. Precursors for peptide hormones share common secondary structures forming features at the proteolytic processing sites. FEBS Lett. **207**:1–6.
25. **Russell, D. W., R. Jensen, M. J. Zoller, J. Burke, B. Errede, M. Smith, and I. Herskowitz.** 1986. Structure of the *Saccharomyces cerevisiae HO* gene and analysis of its upstream regulatory region. Mol. Cell. Biol. **6**:4281–4294.
26. **Simms, P. C., and D. M. Ogrydziak.** 1981. Structural gene for the alkaline extracellular protease of *Saccharomycopsis lipolytica*. J. Bacteriol. **145**:404–409.
27. **Tobe, S., T. Takami, S. Ikeda, and K. Mitsugi.** 1976. Production and some enzymatic properties of alkaline protease of *Candida lipolytica*. Agric. Biol. Chem. **40**:1087–1092.
28. **Wells, J. A., E. Ferrari, D. J. Henner, D. A. Estell, and E. Y. Chen.** 1983. Cloning, sequencing and secretion of *Bacillus amyloliquefaciens* subtilisin in *Bacillis subtilis*. Nucleic Acids Res. **11**:7911–7925.
29. **Yamada, T., and D. M. Ogrydziak.** 1983. Extracellular acid proteases produced by *Saccharomycopsis lipolytica*. J. Bacteriol. **154**:23–31.
30. **Zaret, K. S., and F. Sherman.** 1982. DNA sequence required for efficient termination in yeast. Cell **28**:563–573.