Volume 12 Number 5 1984 Nucleic Acids Research

Analysis of a eukaryotic β-galactosidase gene: the N-terminal end of the yeast *Kluyveromyces lactis* protein shows homology to the *Escherichia coli lacZ* gene product

Karin D.Breunig, Ulrike Dahlems, Sunil Das and Cornelis P.Hollenberg

Institut für Mikrobiologie der Universität Düsseldorf, Universitätsstrasse 1, 4000 Düsseldorf 1, FRG

Received 23 December 1983; Revised and Accepted 14 February 1984

ABSTRACT

The LAC4 gene of Kluyveromyces lactis, encoding the enzyme ß-galactosidase was mapped on a cloned DNA fragment and the sequence of the 5'end was determined. This sequence includes the 5' regulatory region involved in the induction by lactose and the N-terminal end of the protein coding region. Comparison of the deduced amino acid sequence of this eukaryotic enzyme with the N-terminal end of the Escherichia coli ß-galactosidase revealed substantial homology. Two major RNA initiation sites were mapped at -115 and -105. A number of structural peculiarities of the 5'non-coding region are discussed as in comparison to Saccharomyces cerevisiae genes.

INTRODUCTION

The Escherichia coli lac operon was instrumental in the elucidation of the principles of gene regulation at the molecular level in prokaryotes. In eukaryotes no model system for gene regulation is known in such great detail as yet, although a number of systems are well characterized genetically and are being investigated at the molecular level. Substantial progress has been achieved in yeast e.g. for the galactose/melibiose regulon (1,2). Among the yeasts, Kluyveromyces lactis is one of the few species that can grow on lactose. Moreover, this species is amenable to genetic analysis and recently a transformation system has been developed (3). We have therefore chosen K. lactis to study the regulation of the genes involved in lactose metabolism in order to be able to compare this eukaryotic Lac system with the prokaryotic lac operon. In addition K. lactis genes can be expressed in S. cerevisiae (4) which opens the possibility to analyse the functions of the Lac genes by transfer to S. cerevisiae.

ß-Galactosidase, the key enzyme in lactose metabolism, is an

© IRL Press Limited, Oxford, England.

abundant protein in K. lactis cells growing on lactose. It is encoded by the LAC4 gene which has been shown to be regulated at the transcriptional level by lactose and galactose (5,6). The gene has been cloned on the plasmid pBR322 by complementation of an E. coli ß-galactosidase mutant (7).

In this paper we present a detailed restriction map of the cloned K. lactis DNA fragment and the localization of the LAC4 gene by determination of its RNA initiation and termination sites. Furthermore, we have sequenced the 5' regulatory region and the first 350 bp of the coding region to study sequences involved in the interaction with both RNA polymerase and regulatory proteins. The analysis of the primary protein structure shows substantial homology between the eukaryotic and prokaryotic ß-galactosidase enzymes.

MATERIALS AND METHODS

Bacterial strains and plasmids: E. coli strain K514 hsdR$^-$ hsdM$^+$, a derivative of strain C600 was used for plasmid propagation. Strain RRI ΔM15 leu pro strA hsdR hsdM lacZ ΔM15 thi F'lacI^{q}ZΔM15 pro$^+$ (8) was used to construct pUK11.

Plasmid pK16 (7) was kindly provided by R. Dickson. PTY75-LAC4 consists of pCR1, 2-µm DNA and a SalI fragment from pK16, carrying the LAC4 gene (3), pUK11 has the small ClaI fragment of pK16 inserted into the AccI site of pUR222 (9).

Yeast strains: K. lactis wild type strains were CBS2360a and Y1140 (10). K. lactis SD69, a Lac$^-$ derivative of CBS2360 (3) as well as S. cerevisiae AH22 (11) were used for transformation with PTY75-LAC4.Selection of Saccharomyces transformants was performed on G418 plates as described (3), Kluyveromyces transformants were selected on YNB/lactose.

Yeast RNA was isolated from cells grown in YEP containing 0.5% lactose as described (12).

Mapping of restriction sites: restriction enzyme recognition sites of 6 base pairs have been mapped by various combinations of single and double digests of pK16. For enzymes cutting more often, restriction sites were mapped by partial digests of terminally labeled fragments as described by Birnstiel and Smith (13) using the 2.2 and 2.5 kb EcoRI-BamHI and the 2.6 and 1.1 kb XbaI-

BamHI fragments from pK16 (Fig.1). Restriction enzymes were purchased from Boehringer (Mannheim) or Biolabs.

Mapping of RNA initiation sites: S1 mapping experiments were performed exactly as described by Favorolo (14) with the end-labeled fragments described in the text using either 5 µg of polyA RNA or 50 µg of total RNA. With the larger DNA fragments hybridization was performed at 45°C for 6 h in 80% formamide. The EcoRI-Sau3A fragment was hybridized in 50% formamide, since more stringent conditions did not give any detectable hybridization in the small region of homology. S1 treatment after hybridization was performed at 37°C with 1 unit of S1 (Boehringer) per ml. When total RNA was used, probes were treated with RNase (10 µg/ml) for 15 min prior to loading on the gel. Electrophoresis was carried out under denaturing conditions as indicated in the legends to Figs. 2 and 5.

For primer extension an end-labeled 78 nucleotide DNA fragment (3000 cpm, $8x10^6$ cpm/pmole) was used as a primer and hybridized to 30ug polyA RNA at 45°C for 6 h in a volume of 15 µl in 50 mM Tris-HCl(pH 8.3), 60 mM KCl, 10 mM $MgCl_2$, 30 mM ß-mercaptoethanol. After hybridization cold nucleotide triphosphates were added to a final concentration of 0.5 mM and samples were incubated with 5 units of reverse transcriptase from avian myeloblastosis virus (purchased from NEN) for 45 min at 37°C. The reaction was stopped by the addition of EDTA and RNA was hydrolysed in 0.125 M NaOH for 1h in 65°C. Samples were neutralized and ethanol precipitated before loading on sequencing gels.

Chemical DNA sequencing was performed according to Maxam and Gilbert (15) and partly repeated by enzymatical sequencing using the Sanger method in M13mp8 (16,17).

RESULTS

Size and Direction of Transcription of the LAC4 Gene

On the basis of the molecular weight of the *K. lactis* ß-galactosidase which is 135.000 (10) the *LAC4* gene of *K. lactis* was expected to be about 3.3 kb in size. The gene has been cloned in pBR322 on a 7.3 kb insert of *K. lactis* DNA to yield pK16 (7), and the coding sequence was shown to reside in the 4.9 kb segment of DNA to the left of the single XhoI site of that region (4)

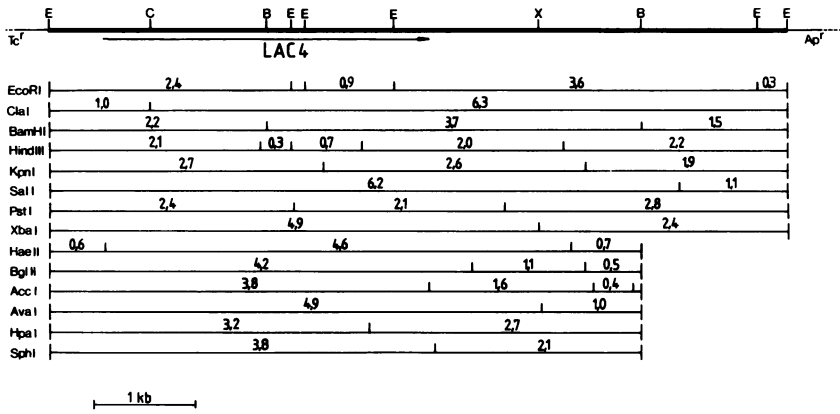

Fig.1: Restriction map of a 7.3 kb K. lactis fragment carrying the LAC4 gene
The upper line shows the plasmid pK16 (7) consisting of a DNA fragment from K. lactis (thick line) with the LAC4 gene cloned into the EcoRI site of pBR322 (thin line). Restriction sites relevant in the text are indicated above this line (E=EcoRI, C=ClaI, B=BamHI, X=XbaI and XhoI (both enzymes cleave very close together)). The arrow gives the direction of transcription and the location of the transcribed region as determined by S1 mapping (Fig.2). For the upper eight enzymes listed, cleavage maps have been determined by double digests of the whole plasmid pK16. The lower six maps have been compiled by limited cleavage of end-labeled fragments as described in material and methods. Sequences right of the BamHI site have not been analysed with this method.

(Fig.1).

We have established a detailed restriction map of that fragment (Fig.1) by limited cleavage of subfragments radioactively labeled at one end (18). The endpoints of the LAC4 gene and the direction of transcription were determined by S1 mapping using RNA isolated from K. lactis CBS2360. Two BamHI fragments of 2.5 and 3.7 kb isolated from pK16 cover most of the cloned K. lactis DNA fragment and therefore should contain at least part of the LAC4 gene. After labeling their 5'ends, only the 2.5 kb fragment was found to be protected from S1 digestion by hybridization to polyA RNA, yielding a fragment of about 1.6 kb (data not shown). Since only one of the two labeled ends lies within K. lactis sequences this result indicates that the direction of transcription of the LAC4 gene is from left to right in Fig.1.

A more precise positioning of the start of transcription was

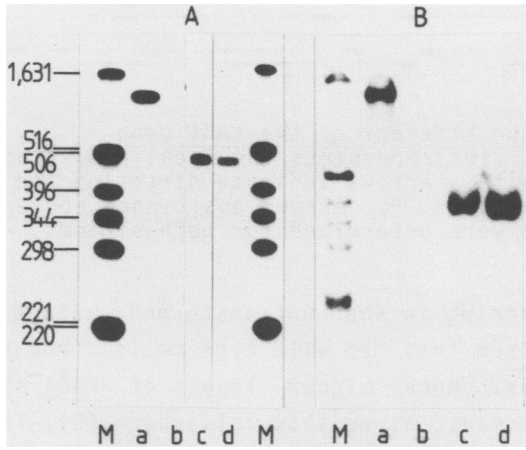

Fig.2: Localisation of the LAC4 gene by S1 mapping of mRNA start- and endpoints
RNA was isolated from wild type K. lactis and transformants grown in YEP medium with lactose and hybridized to end-labeled DNA fragments in 80% formamide, 0.4 NaCl, 40 mM PIPES (pH 6.4), 1 mM EDTA for 6 h at 45°C. After hybridization the samples were treated with S1 (1 U/ml) as described in material and methods.
A:S1 mapping of the start of transcription using a 1.02 kb EcoRI-ClaI fragment from pK16 (Fig.1) 5'end-labeled at the ClaI site. The fragment was denatured and hybridized with the following RNAs: a, 10 ug tRNA, no S1 treatment; b, 10 ug tRNA; c, 5 ug polyA RNA from K. lactis SD69 transformed with PTY75-LAC4; d, d, 5 ug of polyA RNA from K. lactis CBS2360.
Electrophoresis was carried out in 6% polyacrylamid gels containing 7 M urea.
B: S1 mapping of the transcription termination site using a 2.4kb EcoRI-BamHI fragment 3'end-labeled at the EcoRI site. Hybridization was performed with: a, 10 ug tRNA without S1 treatment; b, 10 ug tRNA; c+d, 2 and 10 ug of polyA RNA from K. lactis CBS2360. M, pBR322 cut by HinfI was used as a molecular weight standard. Electrophoresis was performed in 7.5% polyacrylamide after denaturation of the samples in glyoxal/DMSO (29).

achieved by using the 1.02 kb EcoRI-ClaI fragment with a 5'end-label at the ClaI site. In Fig.2A, lanes c and d show that a fragment of approximately 470 bp was protected from S1 digestion by K. lactis RNA. This puts the 5'end of the transcribed region about 550 bp from the EcoRI site on the left hand border of the cloned DNA fragment.

The size of the protected fragment is the same whether the RNA is isolated from wild type K. lactis cells (lane d) or from a Lac⁻ mutant transformed with the plasmid PTY75-LAC4 (3) (lane c).

Fig.3: Sequencing strategy of the LAC4 gene
The continuous line represents the EcoRI-ClaI fragment divided into 100 bp units. Arrows indicate direction and length of sequence determinations. For arrows positioned above and below the line, sequences were determined for both strands.

The LAC4 transcript in the untransformed mutant is also indistinguishable from that in wild type cells (see below, Fig. 5A, lane e), however, much higher levels of LAC4 mRNA are present in the transformant, as we show elsewhere (5). Thus we conclude that the DNA fragment is mainly protected from S1 digestion by plasmid-derived transcripts and that the signals controlling initiation of transcription function on the plasmid in the same way as in the chromosome.

The 3'end has been mapped in the same way using a 3'end-labeled restriction fragment. The 2.4 kb EcoRI-BamHI fragment labeled at the EcoRI site was protected by K. lactis RNA from S1 digestion over 350 bp (Fig.2B) placing the 3'end of the transcribed region 350 bp downstream from this EcoRI site and thereby 3.2 kb from its 5'end. The size of 3.2 kb for the LAC4 coding sequence is close to the value of 3.3 kb suggested by Dickson (10).

Nucleotide Sequence of the 5'End of the LAC4 Gene

In order to be able to analyze structures involved in the control of transcription and in the regulation of the LAC4 gene, we have determined the nucleotide sequence of 1 kb at the 5'end of the gene from the left EcoRI site to the ClaI site (Fig.1). This DNA segment includes about 470 bp of the transcribed region plus 550 bp of sequences 5' to the RNA initiation site. The latter is expected to contain the information for transcriptional control and induction by lactose. For this reason the small ClaI fragment of pK16 (Fig.1) was subcloned into the AccI site of plasmid pUR222 (9) to yield pUK11 (18). The strategy for the sequencing is depicted in Fig.3.

The sequence as shown in Fig.4 revealed only one long open reading frame starting 669 bp from the EcoRI site and continuing to the end of the sequenced region. A comparison of the deduced

```
GAATTCTGTTCACCGCAAAGTTCAGGGTGCTCTGGTGGGTTTCGGTTGGTCTTTGCTTTGCTCTCCCTTGTCTTGCATGTTAATAATAGCCTAGCCTGTG
                -650                                                -600

AGCCGAAACTTAGGGTAGGCTTAGTGTTGGAACGTACATATGTATCACGTTGACTTGGTTTAACCAAGCGACCTGTAGCCAGCCATACCCACACACGTTT
                -550                                                -500

TTTGTATTCTTCAGTATAGTTGTGAAAAGTGTAGCGGAAATATGTGGTCCGAGCAACAGCGTCTTTTTCTAGTAGTGCGGTCGGTTACTTGGTTGACATT
                -450                                                -400

GGTATTTGGACTTTGTTGCTACACCATTCACTACTTGAAGTCGAGTGTGAAGGGTATGATTTCTAGTGGTGAACACCTTTAGTTACGTAATGTTTTCATT
                -350                                                -300

GCTGTTTTACTTGAGATTTCGATTGAGAAAAAGGTATTTAATAGCTCGAATCAATGTGTTATCATTGTGAAGATGTTCTTCCCTAACTCGAAAGGTATAT
                -250                                                -200

GAGGCTTGTGTTTCTTAGGAGAATTATTATTCTTTTGTTATGTTGCGCTTGTAGTTGGAAAAGGTGAAGAGACAAAAGCGCTTAACACTTGAAATTTAGG
                -150                                                -100
                                                                                    met ser cys leu ile pro glu asn
AAAGAGCAGAATTTGGCAAAAAAAATAAAAAAAAAATAAACACACATACTCATCGAGAACTGAAAGAT ATG TCT TGC CTT ATT CCT GAG AAT
                -50                                                 +1
  leu arg asn pro lys lys val his glu asn arg leu pro thr arg ala tyr tyr tyr asp gln asp ile phe glu
  TTA AGG AAC CCC AAA AAG GTT CAC GAA AAT AGA TTG CCT ACT AGG GCT TAC TAC TAT GAT CAG GAT ATT TTC GAA
                                +50
  ser leu asn gly pro trp ala phe ala leu phe asp ala pro leu asp ala pro asp ala lys asn leu asp trp
  TCT CTC AAT GGG CCT TGG GCT TTT GCG TTG TTT GAT GCA CCT CTT GAC GCT CCG GAT GCT AAG AAT TTA GAC TGG
+100                                                 +150
  glu thr ala lys lys trp ser thr ile ser val pro ser his trp glu leu gln glu asp trp lys tyr gly lys
  GAA ACG GCA AAG AAA TGG AGC ACC ATT TCT GTG CCA TCC CAT TGG GAA CTT CAG GAA GAC TGG AAG TAC GGT AAA
                               +200
  pro ile tyr thr asn val gln tyr pro ile pro ile asp ile pro asn pro pro thr val asn pro thr gly val
  CCA ATT TAC ACG AAC GTA CAG TAC CCT ATC CCA ATC GAC ATC CCA AAT CCT CCC ACT GTA AAT CCT ACT GGT GTT
+250                                                 +300
  tyr ala arg thr phe glu leu asp ser lys ser
  TAT GCT AGA ACT TTT GAA TTA GAT TCG AAA TCG AT
                               +350
```

Fig.4: Sequence of the 5'end of the LAC4 gene
The sequence spans an EcoRI-ClaI fragment of 1027 bp including 668 nucleotides upstream and 359 nucleotides downstream of the translation initiation site. Amino acids in the coding region are deduced from the nucleotide sequence.

amino acid sequence encoded by this region with the N-terminal end of the *E. coli* ß-galactosidase showed significant homology. Forty out of 119 amino acid residues were identical in the two polypeptides when they were aligned for optimal correlation. The homology is particularly striking (up to 50%) in the second half

of this segment (see further Discussion, Fig.6).

The nucleotide sequence upstream of the ATG codon shows several features similar to other eukaryotic transcriptional control regions. An AT-rich region, TATAT, which might represent a so-called TATA-box (19), is located around position -170 and is flanked by GC basepairs. Another candidate for a TATA-box is found around position -140: AATTATTATT. The functional significance of either of the two awaits further elucidation. A second element of homology, the CAAT-box (20) is found in many genes of higher eukaryotes but is absent or partially disguised in most S. cerevisiae genes that have been sequenced thus far (21). In LAC4 the sequence AATCAATGT, around position -215 from the ATG has some homology with the consensus sequence GG$^{C}_{T}$CAATCT.

The most striking structural feature in the leader sequence is located at -51 and consists of a run of 18 A residues interrupted by one T. It precedes the hexanucleotide CACACA which has also been found close to the translational start of several yeast genes (21).

Purine residues at position -3 and +4 have been found to occur preferentially in the initiation region of eukaryotic genes (22). The LAC4 gene has a G at position -3 but a T at +4. Position +6 is a pyrimidine (here a T residue), a common feature for genes encoding the abundant glycolytic enzymes in S. cerevisiae (21). A CT-rich sequence followed after 10 nucleotides by the sequence CAAG, which has been proposed by Dobson et al. (21) to be a common feature of all S. cerevisiae genes that encode an abundant RNA, is not present in the K. lactis LAC4 gene.

Precise Mapping of Transcription Initiation Sites

As has been shown above in Fig.2A, transcription starts about 470bp from the ClaI site (Fig.1) around position -120 in the nucleotide sequence. To allow a more precise positioning of the RNA initiation site(s), we have repeated the S1 mapping experiment using a fragment which had been labeled closer to the expected starting site. The S1 protected subfragment was electrophoresed side by side with a sequencing ladder of the original fragment. The results for the EcoRI-Sau3A fragment (position -668 to +82) labeled at the Sau3A site are shown in Fig.5B, next to the C-track of the sequencing gel. In the autoradiograph a series of

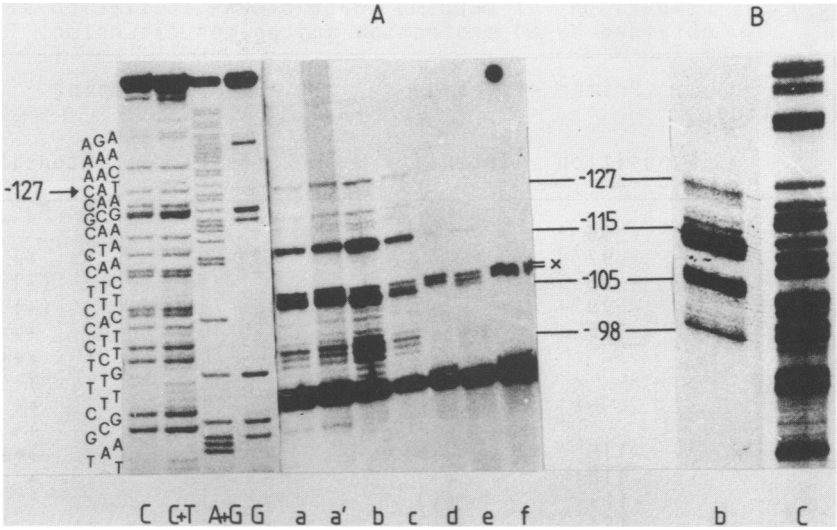

Fig.5: Mapping of transcription initiation sites by primer extension (A) and S1 protection (B)
A: A labeled 78 nucleotide TaqI-HaeII fragment was used as a primer for reverse transcriptase after hybridization to RNA from the following strains:
a, Y1140 polyA RNA; a, Y1140 (using hot nucleotides for primer extension); b, CBS2360; c, SD69 (PTY75-LAC4); d, AH22 (PTY75-LAC4); e, SD69; f, CBS2360 without the addition of reverse transcriptase.
Bands were separated on an 8% polyacrylamide gel containing 7 M urea side by side with a Maxam-Gilbert sequencing reaction of the 140 bp TaqI-DdeI fragment covering the shorter primer fragment. The sequence of the coding strand is shown on the left.
B: A 5'end-labeled Sau3A-EcoRI fragment was hybridized to total RNA from CBS2360 followed by S1 digestion (compare material and methods). The S1 protected fragments were run on a sequencing gel as in Fig.5A parallel to a C-specific sequencing reaction of the same fragment. Numbers between A and B correlate the major bands with the DNA sequence using the numbering of Fig.4. x indicates two bands originating from a contamination of the DNA primer.

bands of different intensities show up with length differences between the major ones of about 30 nucleotides. The endpoints of the various fragments and their relative intensities are listed in Table 1. They are all near position -110 of the LAC4 sequence (Fig.4) which is in agreement with the rough S1 mapping data presented in Fig.2A.

To determine whether these multiple bands reflect an artifact produced by the S1 mapping technique or a heterogeneity in the

Table 1. Comparison of mapping data of RNA initiation sites obtained by S1 protection and primer extension.

first nucleotide in RNA	Primer extension		S1 mapping	
	position	intensity	position	intensity
C	- 96	(+)		
A	- 97	+	- 97	++
G	- 98	++	- 98	(+)
A	- 99	(+)	- 99	(+)
G	-105	+++	-105	+++
G			-106	+++
A			-107	+
T			-113	++
T			-114	++
G	-115	+++	-115	+++
G	-118	(+)	-118	+
T	-119	(+)		
C	-121	(+)		
G	-127	+	-127	+

RNA initiation sites are deduced from the length of the bands shown in Fig.5A and B. Position of the first transcribed nucleotide refers to the sequence listed in Fig.4. Intensities of the bands of Fig.5 are indicated by the number of +. The two major transcripts are underlined.

initiation of the LAC4 transcripts, the RNA starts were additionally mapped by the primer extension method (23). The primer was isolated from a gel as a 78 nucleotide TaqI-HaeII fragment covering nucleotides -91 to -14 and labeled at the TaqI site. After hybridization to polyA RNA from different strains, reverse transcriptase and cold dXTPs were added to synthesize a cDNA. As shown in Fig. 5A, again multiple starting points were obtained. Two bands (marked x in Fig.5) which are larger than the primer result from contamination of the primer DNA, as is obvious from the control in lane f, where no reverse transcriptase was used. Control hybridization with this contaminating fragment alone showed that, as expected, this cannot be used as a primer for reverse transcriptase (data not shown). Almost all other bands correlate with those obtained by S1 mapping. The position of the major starts are given in the middle of Fig.5 and in Fig.7a; a more complete listing of the endpoints determined by both methods is shown in Table 1.

We conclude that the transcripts of the LAC4 gene have multiple 5'ends, the major initiation sites being at positions -115 and -105. In most cases, the first nucleotide of the RNA is a purine, for the major transcripts it is a G.

The positions and intensities of the bands are the same whether RNA from two different wild type K. lactis strains, CBS2360 (lane a and a') and Y1140 (from which the LAC4 gene has been cloned)(lane b), or a lac4 mutant transformed with the LAC4-gene-containing plasmid PTY75-LAC4 (lane c) were used. The transcripts in the untransformed lac mutant (lane e) are propably also indistinguishable from the wild type LAC4 mRNA, but since the RNA had been isolated from uninduced cells, only the two strongest initiation sites are visible.

As argued above, due to the highly elevated LAC4 mRNA levels in transformants, compared to the untransformed mutant, we conclude that the bands in lane c result mainly from plasmid derived transcripts and that even at the nucleotide level no difference in RNA initiation can be observed whether the gene is located on a plasmid or in the chromosome.

The LAC4 gene has been shown to be expressed in S. cerevisiae (4), but it can not be induced (5). When RNA from Saccharomyces transformed with the same plasmid, was used for primer extension (Fig.5,lane d), the signal is weaker than in the uninduced K. lactis lac4 mutant (lane e). The only band weakly visible correlates with the K. lactis transcript starting at position -115. Further analysis of the transcription and the regulation of the LAC4 gene in S. cerevisiae will be described elsewhere (5).

DISCUSSION

As a start in the study of a eukaryotic lactose system, we have analysed the LAC4 gene from the yeast K. lactis, encoding ß-galactosidase, the key enzyme in lactose utilization. We have established a detailed restriction map, mapped the transcribed region and sequenced the 5'end. From expression studies we know that K. lactis lac4 mutants after transformation with a vector carrying the LAC4 DNA fragment display a wild-type induction phenotype (5). Thus we expect the 5' noncoding region to contain control signals for the regulation of the gene by lactose.

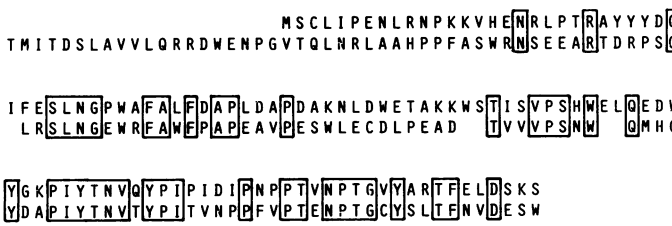

Fig.6: Comparison of the amino acid sequences of the K. lactis and the E. coli ß-galactosidase
The upper line shows the first 119 amino acids of the yeast enzyme; the bottom line the N-terminal 133 amino acids of the E. coli enzyme. For alignment with maximal homology one and two amino acid insertions were allowed in the yeast sequence. Identical amino acids are boxed.

The correlation between the size of the mRNA of 3.3kb (5) and the distance between its transcription initiation and termination sites indicates that the gene has no or at least no large introns. The fact that the LAC4 gene is functionally expressed in E. coli (7) implies the absence of smaller introns as well.

A number of arguments led us to the conclusion that the translation of the LAC4 mRNA starts with the AUG codon at the position indicated in Fig.4:

(i) This AUG codon is followed by the only open reading frame which continues to the end of the sequenced region.

(ii) It is the AUG closest to the 5' ends of the major mRNAs.

(iii) The amino acid sequence deduced from the open reading frame shows homology with the E. coli ß-galactosidase (24) in 40 out of 119 residues.

This homology is particularly interesting in the C-terminal half of these 119 amino acids (Fig.6). Between residue 84 and 109, 65% of the amino acids are identical with perfect colinear alignment. However, identical amino acids in the two proteins are often encoded by different codons, so that the homology at the nucleotide level is not significant. It seems that amino acids 66 to 116 (83 to 130 in E. coli) are part of a functionally important domain of the protein and homology within this region might reflect a divergent evolution of the two genes from a common ancestor.

The extreme codon bias, found in S. cerevisiae for some of the

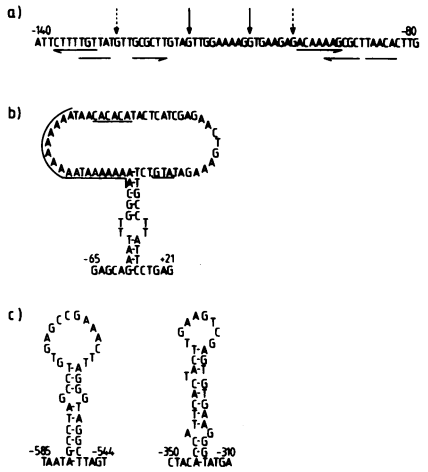

Fig.7: Analysis of the LAC4 5'control region for potential secondary structures
a) Transcription initiation sites (vertical arrows, solid for strong initiation sites, dotted for weaker ones), surrounded by two sets of inverted repeats as underlined by horizontal arrows.
b) Stem and loop structure including the start of translation and peculiar sequences in its upstream region.
c) Inverted repeats upstream of the transcriptional start.
Data were obtained by computer analysis of the DNA sequence in a Tinoco-diagram (30) on an Apple II computer. The program was kindly supplied by G. Steeger and W. Rapport.

most abundant proteins (25) is not found for the LAC4 gene of K. lactis. Thirty-nine out of the 119 codons sequenced are rarely used in S. cerevisiae, which suggests that K. lactis has evolved far enough from S. cerevisiae to allow for a different codon bias. As LAC4 is the first gene for an abundant protein from K. lactis that has been sequenced, we cannot exclude that genes coding for other major proteins such as the glycolytic enzymes follow the S. cerevisiae codon bias. Sixty-three codons of the LAC4 gene correspond to those frequently used in S.cerevisiae.

The sequenced DNA fragment carrying the LAC4 gene extends to 668 nucleotides upstream from the ATG initiation codon and therefore should contain the sequences interacting with RNA polymerase. Two elements resembling a TATA-box might be involved in directing transcription initiation. They are located at -140 and -170 about 30 and 60bp from the center of the multiple transcriptional initiation sites. The two major starts are located ten basepairs apart at positions -115 and -105; a slightly weaker one was found at position -98. Mapping experiments by S1 protection as well as primer extension gave almost the same results. The significance of a strong band at position -106 present only in the S1 protection experiment is unclear. Both techniques revealed a number of weaker bands indicating a rather heterogeneous LAC4 mRNA population with 5'ends at almost every

nucleotide between -98 and -105 and a few start-points further upstream. In this respect it will be interesting to see whether any functional correlation exists between either of the two TATA-boxes and any of the single RNA initiation sites. For each RNA species the deduced ATG initiation codon (Fig.4) is the first from the 5'end of the molecule. We have no information yet about the efficiency of translation of the different transcripts.

It is interesting to note that all transcription initiation sites fall between two sets of inverted repeats which are shown in Fig.7a. The possibility that these sequences have caused an artifact in mapping the 5'end of the mRNA is unlikely, since at least in the primer extension experiment premature termination of reverse transcriptase due to secondary structure in the RNA template would be expected at the bottom of a stem rather than in the loop of a potential hairpin structure. We believe that the upstream half of the inverted repeat is not transcribed and if there is any significance in the presence of the inverted repeats it will be at the DNA rather than at the RNA level.

Another obvious structural feature in the non-translated region is a run of 18 A-residues interrupted by one T. In contrast to other yeast genes (CYC1 (26), ADH2 (27), and PDC1 (28)), where a long A-stretch has been found to precede the transcription initiation sites, in LAC4 this sequence is part of the leader RNA. The stretch of 16 of these A-residues, the hexanucleotide CACACA found in many yeast genes, and the ATG initiation codon can hypothetically be looped out in a snap-back structure with a short imperfect stem of 9 base-pairs (Fig.7b).

Since the cloned LAC4 gene still responds to regulation by lactose, we have analysed the region upstream of the transcription initiation site for additional symmetrical structures that might be involved in interaction with regulatory proteins. Two such sequences are centered around positions -565 and -330, respectively (Fig.7c). We are presently determining which sequences are involved in the regulation of the LAC4 gene.

ACKNOWLEDGEMENTS

We like to thank R.C. Dickson for supplying us with the cloned LAC4 gene in the plasmid pK16 and Petra Kuger for excellent technical assistance.

REFERENCES

1. St. John, T.P. and Davis, R.W. (1981) J.Mol.Biol.152,285-316
2. Johnston, S.A. and Hopper, J.E. (1982) Proc.Natl.Acad.Sci USA 79, 6971-6975
3. Das, S. and Hollenberg, C.P. (1982) Curr.Genet. 6,123-128
4. Dickson, R.C. (1980) Gene 10,347-356
5. Das, S., Breunig, K.D. and Hollenberg, C.P., submitted for publication
6. Lacy, L.R. and Dickson, R.C. (1981) Mol.Cell.Biol. 1,629-634
7. Dickson, R.C. and Markin, J.S. (1978) Cell 15,123-130
8. Rüther, U. (1982) Nucl.Acids Res.10,5765-5772
9. Rüther, U., Koenen, M., Otto, K. and Müller-Hill, B. (1981) Nucl.Acids Res.7,1513-1523
10. Dickson, R.C., Dickson, L.R. and Markin, J.S. (1979) J.Bact. 139,135-141
11. Hinnen, A., Hicks, J.B. and Fink, G.R. (1978) Proc.Natl.Ac. Sci USA 75,1929-1933
12. Breunig, K.D., Mackedonski, V. and Hollenberg, C.P. (1982) Gene 20, 1-10
13. Birnstiel, M.L., Smith, H.O. (1976) Nucl.Acids Res.3,2387-2398
14. Favorolo, J., Treisman, R. and Kamen, R. (1980) Methods Enzymol. 65,718-749
15. Maxam, A.M. and Gilbert, W. (1980) in Methods in Enzymology Colowick, S.P. and Kaplan, N.O. Eds., Vol 65, pp499-560, Academic Press, New York
16. Sanger, F., Nicklen, S. and Coulsen, A.R. (1977) Proc.Natl. Sci USA 74,5463-5467
17. Vieira, J. and Messing, J. (1982) Gene 19,259-268
18. Dahlems, U. (1983) Diplomarbeit Universität Düsseldorf
19. Gannon, F., O'Hare, K. Perrin, F., LePennec, J.P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A. and Chambon, P. (1979) Nature 278,428-434
20. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucl.Acids Res.8,127-142
21. Dobson, M.J., Tuite, M.F., Roberts, N.A., Kingsman, A.J., Kingsman, S.M., Perkins, R.E., Conroy, S.C., Dunbar, B. and Fothergrill, L.A. (1982) Nucl.Acids Res.10,2625-2637
22. Kozak, M. (1981) Nucl.Acids Res.9,5233-5252
23. Ghosh, P.K., Reddy, V.B., Swinscoe, J., Lebowitz, P. and Weissmann, S.M. (1978) J.Mol.Biol.126,813-846
24. Kalnins, A., Otto, K., Rüther, U. and Müller-Hill, B. (1983) Embo J.2,593-597
25. Bennetzen, J.L. and Hall, B.D. (1982) J.Biol.Chem. 257,3026-3034
26. Smith, M., Leung, D.W., Gillam, S., Astell, C.R., Montgomery, D.L. and Hall, B.D. (1979) Cell 16,753-761
27. Russel, D.W., Smith, M., Williamson, V.M. and Young, E.T. (1983) J.Biol.Chem.258,2674-2682
28. Hollenberg, C.P., Roggenkamp, R., Erhart, E., Breunig, K. and Reipen, G. in Gene Expression in Yeast. Proceedings of the Alko Yeast Symp. Helsinki 1983, ed. by Korhola, M. and Väisänen, E. (1983) Foundation for Biotechnical and Industrial Fermentation Research 1,73-90
29. McMaster, G.K., Carmichael, C.G. (1977) Proc.Natl.Acad.Sci USA 74,4835-4838
30. Tinoco, I., Uhlenbeck, O.C., Levine, M.D. (1971) Nature 230, 362-367