The nucleotide sequence of the yeast \textit{MEL1} gene

Pirkko L. Liljeström

Research Laboratories of the Finnish State Alcohol Company (Alko Ltd), PO Box 350, SF-00101 Helsinki, Finland

Received 13 August 1985; Revised and Accepted 26 September 1985

ABSTRACT

The complete nucleotide sequence of the \textit{MEL1} gene of the yeast, \textit{Saccharomyces cerevisiae}, encoding \(\alpha\)-galactosidase was determined. The nucleotide sequence contains an open reading frame of 1413 bp encoding a protein of 471 amino acids. Comparison with the known N-terminal amino acid sequence of the mature secreted protein indicated that \(\alpha\)-galactosidase is synthesized as a precursor with an N-terminal signal sequence of 18 amino acids. The general features of this signal peptide resemble those of other yeast signal peptides. Molecular weight of the mature \(\alpha\)-galactosidase polypeptide deduced from the nucleotide sequence is 50,049 kDa. The 5' regulatory region has sequences in common with other yeast genes regulated by the GAL4-protein.

INTRODUCTION

The ability of \textit{Saccharomyces cerevisiae} (1) to utilize both galactose and melibiose is dependent on a regulon consisting of the five structural genes \textit{MEL1}, \textit{GAL1}, \textit{GAL2}, \textit{GAL7}, \textit{GAL10} and at least four regulatory genes \textit{GAL3}, \textit{GAL4}, \textit{GAL11} and \textit{GAL80} (2-4). The \textit{GAL} structural genes are responsible for uptake and utilization of galactose. The \textit{MEL1} gene product, \(\alpha\)-galactosidase, cleaves melibiose into galactose and glucose. The expression of the structural genes is induced by galactose and repressed by glucose (2,3,5-8). The \textit{GAL4} product is a positive regulator needed to activate transcription of the target genes (3,5,9-11) by direct binding to their regulatory regions (12). This function of the \textit{GAL4} protein appears to be impaired by the \textit{GAL80} protein (3,13) in the absence of inducer (galactose) (11,14-16).

Only some strains of \textit{S. cerevisiae} carry a \textit{MEL} gene and therefore produce \(\alpha\)-galactosidase. Such strains were previously classified as \textit{S. uvarum} or \textit{S. carlsbergensis} (1). \(\alpha\)-galactosidase is glycosylated and is secreted to the outside of the cell membrane (17,18). Purification studies have indicated an oligomeric structure and a molecular weight of 300 kDa for the native enzyme, 57\% (w/w) of which is carbohydrate (18).

Recently the \textit{MEL1} gene has been cloned and found to code for a single tran-
script of ca. 1.6 kb (9). This corresponds to an approximate size of 54-55 kd for the polypeptide chain.

In order to gain better insight into the structure and function of the MEL1 gene, its complete nucleotide sequence was determined.

**MATERIALS AND METHODS**

**Plasmids and phages**

The MEL1 gene was cloned from the *Saccharomyces* strain ATCC 9080 by complementation of a mell-18 mutation (Ruohola et al., in preparation). Plasmid pALK7, used as source of DNA in the sequence determination, carries the MEL1 gene on a 2.8 kb fragment in between BamH1 and SalI sites of pBR322 (Ruohola et al., in preparation). The phage vectors used for sequencing were M13mp18 and M13mp19 (19). The plasmid DNA was prepared essentially as described previously (20).

**Strains and media**

*Escherichia coli* (E. coli) strain HB101 (proA2, leuB6, thi, rpsL20, lacY1, araL4, galk2, xyl5, mtl1, recA13, supE44, endA, hsdR, hsdM) harboring the plasmid pALK7 was grown in L-broth (21) containing 150 μg/ml ampicillin. The E. coli strain JM103 (Δ(lac-pro), thi, rpsL, endA, sbcB15, supE/F' traD36, proAB+, lacIq, lacZAM15) used for transfection of M13 phages was grown on M9glycerol + thi plates (21). *Saccharomyces cerevisiae* (var. uvarum) strain ATCC9080 was grown in SGal (22) medium (22) and used as source of RNA.

**Enzymes and reagents**

Restriction endonucleases, T4-DNA ligase, DNA-polymerase I Klenow fragment and M13-hybridization primers were purchased from Boehringer-Mannheim. S1-nuclease, deoxynucleoside triphosphates and dideoxynucleosidetriphosphates were from P.L. Biochemicals. [α-32P]dATP (3000 Ci/mmol) and [α-32P]ATP (3000 Ci/mmol) were from Amersham.

**S1-mapping**

The 5' end of MEL1-encoded transcript was mapped by S1 nuclease protection (23). Total yeast RNA (100 - 300 μg) isolated from galactose induced cells, was ethanol precipitated with 40 000 to 100 000 cpm of probe and suspended in 30 μl of 40 mM PIPES [piperazine-N,N'-bis(2-ethanolsulfonic acid), pH 6.4], 1 mM EDTA, 0.4 M NaCl, 80 % formamide; the suspension was heated at 95°C for 5 min and annealing of nucleic acid was carried out at 48 °C for 3 h. To this mixture 300 μl of ice-cold S1 buffer (180 mM NaCl, 30 mM NaAc(pH 4.5), 4.5 mM ZnSO4) containing 100U/ml of S1 nuclease (P.L. Biochemicals) was added and the reaction was incubated at 37 °C for 30 min. The reaction was stopped by
adding EDTA to a final concentration of 10 mM and phenol extraction. RNA was then hydrolyzed with NaOH for 1 h at 65 °C. The mixture was neutralized and ethanol precipitated. The protected fragments were analyzed on sequencing gels.

Preparation of probe

The probe used for SI nuclease protection was end labeled at XbaI site 149 nucleotides downstream from translation start point. Plasmid containing the 609 bp SalI-XbaI fragment of yeast DNA was cut with XbaI and end labeled using polynucleotide kinase as described (24). The probe for SI nuclease protection was generated by cleaving the end labeled DNA with SalI and purifying the 609 bp SalI-XbaI fragment.

DNA sequencing

Determination of DNA sequence was performed by the dideoxy-chain termination method by using bacteriophage M13 system (25,26).

Dot blot hybridization

Dot blot hybridization was performed essentially as described by Thomas (27).

RESULTS AND DISCUSSION

We have previously cloned the MEL1 gene and localized it on a 2.8 kb SalI-BamHI fragment (Ruohola et al. in preparation). The complete nucleotide

![Open Reading Frame Diagram](attachment:image)

Figure 1. MEL1 DNA sequencing strategy. The restriction sites used for the sequencing are shown along with arrows indicating the length and direction of sequences determined by the chain termination method. The position of the 471-codon open reading frame is shown at the bottom. Abbreviations: A, AluI; Ac, AccI; Be, Bell; B, BamHI; H, HincII; R, RsaI; S, SalI; X, XbaI. Direction of transcription was determined by using the clones marked with *.
sequence of both strands of this fragment was determined and the junction points were sequenced by the overlapping fragments. The strategy for sequence analysis is shown in figure 1 and figure 2 presents the complete nucleotide sequence of the MEL1 region.

**Coding region**

The entire region sequenced contained only one long open reading frame in either direction, beginning at position +1 (in fig. 2) and ending with the stop codon TGA at position +1414. The direction of this open reading frame is consistent with the direction of the MEL1 transcription determined by using DNA-RNA hybridisation and two M13 clones carrying opposite strands of the MEL1 gene as probes. Only one of these clones (marked n:o 1 in fig. 1) hybridizes to total yeast RNA in dot blot hybridization experiment (data not shown, method in ref. 28). The amino acid sequence of α-galactosidase as deduced from the MEL1 DNA sequence is shown in fig. 2. The protein is assumed to begin at the first AUG codon (nucleotides +1 - +3) as no other methionines are present further upstream. There is an in-frame AUG-codon further downstream, at nucleotides +94 - +96 (fig. 2). However it has been found in eukaryotes that the most 5′AUG-codon of the mRNA is generally the one used for translation initiation (29). The translation start at position +1 is also supported by the fact that it has the invariant A at position -3 and a purine, U, at position +4 as has been found around initiation codons of other eukaryotic mRNA’s (29).

The molecular weight of α-galactosidase calculated from the predicted amino acid sequence is 52.044 kd, which is in good agreement with the apparent molecular weight of purified secreted endo-H treated α-galactosidase of 53 kd estimated by SDS-gel electrophoresis (Kopu et al. in preparation). The protein size of 52.044 kd is also in agreement with the size of mRNA of ca. 1.6 kb (9).

**Figure 2. Nucleotide sequence and deduced amino acid sequence of the MEL1 gene.** The complete sequence of a 2812 nucleotide region containing the MEL1 structural gene is shown. The sequence of the mRNA-identical strand of the MEL1 gene is presented. The numbering of nucleotides begins at the A in the translation initiating ATG. TATA-like sequences preceding the coding region are underlined and the position of the 5′ end of the MEL1 mRNA is marked with an asterisk. Sequence thought to play a role in GAL4-mediated regulation is in brackets and inverted repeats are indicated by arrows. The putative recognition sequences for poly(A) addition and transcription termination are underlined twice. The open reading frame is translated into amino acids. The amino acids are numbered from the N-terminus of the mature protein. Open box indicates the potential glycosylation sites.
Protein structure

Since α-galactosidase is a secreted protein it would be expected to contain a signal sequence. The N-terminus of purified secreted α-galactosidase has been shown to start with the valine at position +1 in figure 2 (Kopu et al. in preparation). This valine is preceded by 18 amino acid residues from the assumed initiating methionine. This peptide shares common features with known signal peptides in both prokaryotes and higher eukaryotes (30,31) strongly suggesting that α-galactosidase is synthesized as a precursor with 18 amino acid long extension at its N-terminus. The N-terminus of this precursor is very hydrophobic, characteristic of known signal sequences. Its length is similar to the length of the few known yeast signal peptides varying from 17 to 30 amino acid residues (table 2). A positively charged residue is usually found within the first five residues in both eukaryotic and prokaryotic signal sequences (31). Specifically in yeast signal peptides at least one positively charged residue is invariantly present, although it is not always located at the N-terminal part of signal peptide. In invertase, killer toxin and the putative α-galactosidase signal sequence a positively charged residue is close to the C-terminus of the signal peptide (table 2), although in the case of killer toxin positively charged residue is also present close to the N-terminus. Whether the position of positive charge has any significance in export efficiency in yeast remains to be shown.

The Kyte and Doolittle plot (32) of hydrophilic and hyrophobic regions in α-galactosidase indicates that the N-terminus, predicted as signal sequence, is the most hydrophobic region of the protein (fig. 4). Mature α-galactosidase is composed of alternating hydrophobic and hydrophilic regions. This hydroplot is consistent with α-galactosidase being a soluble protein, since there are no hydrophobic sequences in the mature protein that would be long and hydrophobic enough to span a membrane (32,33). The deduced amino acid sequence of mature α-galactosidase was analyzed for the expected secondary structure by using the method of Chou and Fasman (34). As shown in figure 4 α-galactosidase is a typical α-β-protein ca. 24 % being predicted as α-helix and ca. 24 % as β-sheet.

Codon usage

The codon usage in the HEL1 gene is summarized in table 1. In the MEL1 gene 60 out of the 61 possible codons are used. This is in contrast to the highly expressed yeast genes such as alcohol dehydrogenase I (35) and glyceraldehyde-3-phosphate dehydrogenase (36,37), in which 96 % of the amino acid residues are encoded by certain 25 of the 61 possible coding triplets. These
preferred codons correspond to the most abundant isoacceptor tRNA species of yeast (38). There is also a good correlation between the degree of the bias towards these selected codons and the extent of expression of a gene, with highly expressed genes showing the most extreme bias. Bennetzen and Hall (38) have defined a codon bias index which essentially quantifies, on a scale of 1.0-0, the extent of the bias towards a set of 22 preferred codons identified in these two very highly expressed genes. On this basis the codon bias index for MEL1 (0.23) indicates a moderately expressed gene.

5` and 3` non-coding regions

The 5` end of the MEL1 transcript was mapped by a SI nuclease protection experiment (23). The probe employed was 606 bp SalI-XbaI fragment 5` end-labeled at the XbaI site (Fig. 3). This probe was hybridized to total RNA

---

Table 1. Codon usage of the MEL1 471-codon open reading frame. Percentages are shown in parentheses.

<table>
<thead>
<tr>
<th>Codon</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT-Phe</td>
<td>12 (2.5)</td>
</tr>
<tr>
<td>TAC-Tyr</td>
<td>9 (1.9)</td>
</tr>
<tr>
<td>TGG-Trp</td>
<td>6 (1.3)</td>
</tr>
<tr>
<td>CTT-Leu</td>
<td>9 (1.9)</td>
</tr>
<tr>
<td>CTC-Leu</td>
<td>9 (1.0)</td>
</tr>
<tr>
<td>CTA-Leu</td>
<td>6 (1.3)</td>
</tr>
<tr>
<td>CTG-Leu</td>
<td>7 (1.5)</td>
</tr>
<tr>
<td>ATT-Ile</td>
<td>12 (2.5)</td>
</tr>
<tr>
<td>ATC-Ile</td>
<td>9 (1.0)</td>
</tr>
<tr>
<td>ATA-Ile</td>
<td>2 (0.4)</td>
</tr>
<tr>
<td>ATG-Met</td>
<td>13 (2.8)</td>
</tr>
<tr>
<td>GTT-Val</td>
<td>11 (2.3)</td>
</tr>
<tr>
<td>GTC-Val</td>
<td>4 (0.8)</td>
</tr>
<tr>
<td>GTA-Val</td>
<td>2 (0.4)</td>
</tr>
<tr>
<td>GTG-Val</td>
<td>3 (0.8)</td>
</tr>
<tr>
<td>GCT-Ala</td>
<td>12 (2.5)</td>
</tr>
<tr>
<td>GCC-Ala</td>
<td>9 (1.9)</td>
</tr>
<tr>
<td>GCA-Ala</td>
<td>12 (2.5)</td>
</tr>
<tr>
<td>GCG-Ala</td>
<td>3 (0.6)</td>
</tr>
</tbody>
</table>

---

Table 2. Comparison of yeast signal sequences

<table>
<thead>
<tr>
<th>Signal sequence</th>
<th>STAl</th>
<th>KILLER toxin</th>
<th>pro α-factor 1</th>
<th>pro α-factor 2</th>
<th>pro 4 factor 1</th>
<th>pro 4 factor 2</th>
<th>MEL1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Met</td>
<td>Val</td>
<td>Ile</td>
<td>Leu</td>
<td>Asp</td>
<td>Tyr</td>
<td>His</td>
</tr>
<tr>
<td>Killer toxin</td>
<td>Met</td>
<td>Val</td>
<td>Ile</td>
<td>Leu</td>
<td>Asp</td>
<td>Tyr</td>
<td>His</td>
</tr>
<tr>
<td>Pro α-factor 1</td>
<td>Met</td>
<td>Val</td>
<td>Ile</td>
<td>Leu</td>
<td>Asp</td>
<td>Tyr</td>
<td>His</td>
</tr>
<tr>
<td>Pro α-factor 2</td>
<td>Met</td>
<td>Val</td>
<td>Ile</td>
<td>Leu</td>
<td>Asp</td>
<td>Tyr</td>
<td>His</td>
</tr>
<tr>
<td>MEL1</td>
<td>Met</td>
<td>Val</td>
<td>Ile</td>
<td>Leu</td>
<td>Asp</td>
<td>Tyr</td>
<td>His</td>
</tr>
</tbody>
</table>

* proposed or known end of signal sequence

In the case of STAl it has not been shown which of the two methionines in this signal sequence is used to start translation.
Figure 3. SI mapping of the 5'end of the MEL1 mRNA

Total yeast RNA was hybridised to a radioactive double-stranded SalI-XbaI probe DNA 5' end-labelled at the XbaI end (see lower part of the figure). After treatment with SI nuclease (100 U/ml) the nuclease-resistant hybrids were analyzed on an 8% polyacrylamide gel containing 8M urea. Lane b: 100 μg of tRNA; lane c: 300 μg of total RNA; lane d: 100 μg of total RNA; lanes a,e,f,g : molecular weight DNA markers of known sequence. Their size (in bases) is indicated in the margin.

prepared from yeast strain ATCC9080 grown on galactose. The DNA/RNA hybrids were digested with SI nuclease and the resulting SI-resistant fragments were analyzed on the sequencing gel in parallel with radioactively labeled restriction fragments of known size (fig. 3). From this analysis it can be concluded that the major 5' end of the MEL1 mRNA is located at a position between -20 and -24 in the non-coding sequence. There also seems to be a minor start point between -48 and -53. Multiple start sites are common to a number of yeast genes including GAL1 and GAL10 (39). A number of yeast genes that encode abundant mRNAs seem to have their transcription start sites at or very near the sequence CAAG and a CT rich block is usually found upstream of the CAAG sequence in the vicinity of the mRNA start site (40). The major 5' end of the
Figure 4. Hydropathic profile and predicted secondary structure of α-galacto-
sidase precursor. Hydropathic Index is calculated according to Kyte and
Doolittle (32) and expected secondary structure is predicted by the method of
Chou and Fasman (34). $\alpha$-helix; $\beta$-sheet; $\square$, $\beta$-turn; --, coil.

MEL1 transcript is located at the sequence CAAC (from -23 to -20) and a CT-
rich sequence, 5'-TTTCTTACTTGATTCT-3', is located between positions -31 and -46 ca. 10 bp 5' to the major mRNA start.

Two TATA boxes (41) were found in the 5' region of the MEL1 gene at posi-
tions -119 to -113 and -110 to -104. Whereas in higher eukaryotes TATA boxes are located at ca. -35 bp from the transcriptional start, in yeast genes this distance is more variable and usually longer (42).

Transcription of the MEL1 gene as well as the GAL1, GAL7 and GAL10 genes encoding galactose-metabolizing enzymes, is induced by galactose (3,5,9). The GAL1 and GAL10 genes are divergently transcribed and are separated by 680 bp of DNA that contains a sequence called the GAL upstream activating sequence (UAS$_G$), which is required for this induction (5,39). Induction also depends on the positive regulatory protein encoded by the GAL4 gene (3,10). It has recently been shown that the GAL4-protein binds directly to UAS$_G$ to activate transcription of the GAL1 and GAL10 genes (12). The consensus GAL4-protein recognition sequence proposed is a 17 bp dyad symmetric sequence 5'-CGGAC$^A$$^C$$^C$$^T$$^C$$^C$$^G$$^C$-3'. One related sequence is also found in the upstream region of a third gene controlled by GAL4-protein, the GAL7 gene. Also the expression of the MEL1 gene is controlled by GAL4-protein and thus it is expected that the 5' region of this gene has a similar sequence. Indeed a
related sequence, 5'-CGGCCATATGTCTTCCG-3' (from -253 to -237), is found at about 220 bp upstream from the major transcriptional start point. In addition, two copies of a short sequence, 5'-GCAACA-3' proposed earlier (43) to play a role in GALA-mediated induction, are present in the 5' region of the MEL1 gene; one is part of the 17 bp consensus (bottom strand from -238 to -244) and an inverted repeat of this sequence is found in the sequence from -223 to -216.

Transcription termination and polyadenylation signals have not yet been established in yeast, although several proposals have been made. A consensus sequence TAAATAAATT, related to the postulated polyadenylation signal AATAAA (35,44) of higher eukaryotic genes, has been suggested to be involved in termination and/or polyadenylation in yeast. In the MEL1 gene, the sequence TAAATAAT, at position +1508 to +1515 (fig. 2) is clearly related to both these sequences. In addition Zaret and Sherman (45) have proposed that the tripartite sequence TAG...TAGT...TTT or TAG...TATG...TTT may have a role in transcription termination and polyadenylation in yeast genes. Also this sequence is present in the 3' non-coding region of the MEL1 gene (fig. 2). Poly(A) addition occurs 10 to 40 bp downstream from the TAGT-part of this signal in those yeast mRNAs that have been analyzed so far. This places the likely end of the MEL1 mRNA at nucleotides 1560 to 1590 (fig. 2). Thus the size of the mRNA based on this sequence analysis would be ca. 1580-1610. This is consistent with the MEL1 mRNA size of ca. 1.6 kb deduced from Northern blot analysis (9, Ruohola et al. in preparation).

ACKNOWLEDGEMENTS

I wish to thank Peter Liljestrom for his expert advise on DNA sequencing techniques and for helpful discussions. I thank Dr. Roy Tubb for critical reading of the manuscript and Tuula Roikonen-Kari for skilful technical assistance.

REFERENCES


