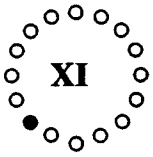
YEAST VOL. 10: S81–S91 (1994)



Yeast Sequencing Reports

Sequencing of a 13·2 kb Segment Next to the Left Telomere of Yeast Chromosome XI Revealed Five Open Reading Frames and Recent Recombination Events with the Right Arms of Chromosomes III and V

DESPINA ALEXANDRAKI†‡* AND MARIA TZERMIA†

†*Foundation for Research and Technology-HELLAS, Institute of Molecular Biology and Biotechnology and* ‡*University of Crete, Department of Biology, P.O. Box 1527, Heraklion 711 10 Crete, Greece*

Received 15 August 1993; accepted 1 December 1993

We report the entire sequence of a 13·2 kb segment next to the left telomere of chromosome XI of *Saccharomyces cerevisiae*. A 1·2 kb fragment near one end is 91% homologous to the right arm of chromosome III and 0·7 kb of that are 77% homologous to the right arm of chromosome V. Five open reading frames are included in the sequenced segment. Two of them are almost identical to the known YCR104W and YCR103C hypothetical proteins of chromosome III. A third one contains a region homologous to the Zn (2)-Cys (6) binuclear cluster pattern of fungal transcriptional activators. The fourth one, part of which is similar to the mammalian putative transporter of mevalonate, has the structure of membrane transporters. The fifth one is similar to yeast ferric reductase. The sequence has been deposited in the EMBL data library under Accession Number X75950.

KEY WORDS — Genome sequencing; *Saccharomyces cerevisiae*; chromosome XI; chromosomal recombination; telomeres; gene redundancy.

INTRODUCTION

In the course of the European Community (BRIDGE) project to sequence *Saccaromyces cerevisiae* chromosome XI, we have determined the complete sequence of 13 213 base pairs on a DNA fragment mapped next to the left telomere (about 200 nucleotides away from a 1 kb telomeric sequence, unpublished results). This fragment contains five open reading frames (ORFs), the potential function of which will be discussed below.

MATERIALS AND METHODS

Strains and vectors

Cosmids pUKG040 and pEKG100 were provided in *Escherichia coli* strain TG1 (Δ(*lac pro*), *thi1, supE44, hsdD5*, F′ (*traD36, proA*$^{+}$ *B*$^{+}$ *lacI*^{Q} *lacZ*ΔM15)) from Agnès Thierry and Bernard Dujon (Thierry and Dujon, in preparation). They are cosmids from libraries of chromosome XI, derivatives of cosmids pOU61 cos and pWE15 respectively, containing overlapping partial *Sau*3AI yeast DNA fragments. *Escherichia coli* strain DH5α (*supE*44 Δ*lac*U169 (φ80*lacZ*ΔM15) *hsdR*17 *recA1 endA1 gyrA96 thi-1 relA1*), and pUC18 and pUC19 vectors were used for all subsequent subcloning and sequencing steps.

Sequencing strategy

We have used directed sequencing of ordered restriction fragments. Cosmid DNAs were digested with *Eco*RI and electrophoresed in low melting point agarose. Four *Eco*RI fragments were purified and subcloned into pUC18 or pUC19 vectors.

*Author to whom correspondence should be addressed.

CCC 0749–503X/94/SQ0081–11

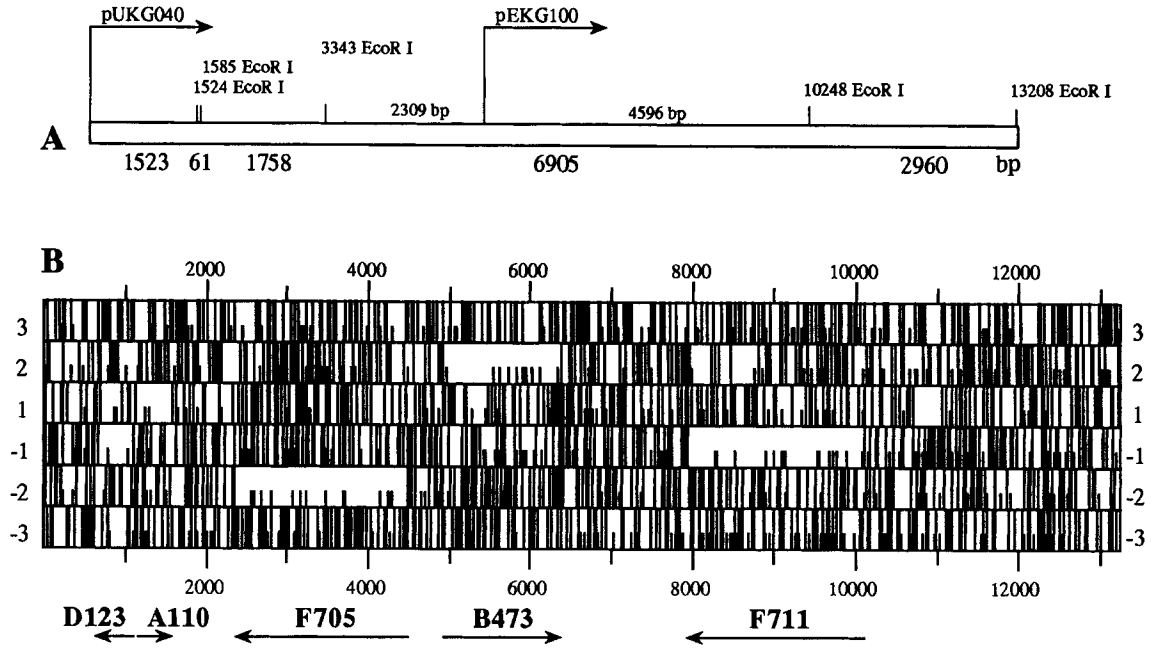


Figure 1. (a) *Eco*RI restriction map of the 13 213 base pair segment. The arrows indicate the beginning (*Sau*3AI sites) of the sequences included in the two overlapping cosmids. The numbers below the bar indicate the size of each *Eco*RI fragment. (b) 6-phase ORF map of the 13 213 base pairs. Small bars indicate initiation codons and full bars indicate stop codons. The location and the direction of five ORFs are indicated by arrows. The number in the name of each ORF indicates its size in amino acids. The ORF names were assigned by MIPS.

The order of the *Eco*RI fragments is shown on the map of Figure 1A. The sequences of the two 5′ *Eco*RI fragments and 2 kb of the third one have been determined from cosmid pUKG040. The rest of the reported sequence derives from cosmid pEKG100. The additional sequence of 24·6 kb included in the pEKG100 insert has been reported separately (Tzermia *et al.*, 1994).

Double stranded template DNAs were prepared by alkaline lysis followed by Qiagen-tip selection (Qiagen Inc.) or PEG precipitation (Ausubel *et al.*, 1987). They were subsequently sequenced using [^{35}S]dATP and the Sequenase kit (United States Biochemical Corp.) following the supplier's protocols. Sequencing of both strands of the *Eco*RI fragments, subcloned in both orientations, was performed by the 'universal' or the 'reverse' M13 primers on nested ExoIII-S1 nuclease (Ausubel *et al.*, 1987) deletions. Synthetic oligonucleotides corresponding to internal sequences (prepared on an Applied Biosystems synthesizer by the Department of Microchemistry at I.M.B.B.-Crete) were used as primers to fill in the gaps. The junctions between the sequenced *Eco*RI fragments have been established by PCR sequencing of PCR products, which were synthesized from oligonucleotide primers near the ends of the fragments, using cosmid DNA as template, [^{32}P]ATP-labelled primers and the *fmol* DNA sequencing kit (Promega Corp.). Samples of sequenced DNAs were analysed on 40 cm long 6% or 4% polyacrylamide gels with single or double loadings.

Sequence analysis software

Endonuclease restriction, 6-phase ORF mapping and hydrophobicity profiles of the sequences were accomplished by the DNA Strider software (Marck, 1988). Comparisons of nucleotide and amino acid sequences were made to the GenBank, EMBL, SWISS-Prot and NBRF libraries using the GCG package software by us on the I.M.B.B. MicroVAX and by the staff at MIPS.

RESULTS AND DISCUSSION

Sequence determination

The reported sequence was determined from overlapping Exo III produced deletions and

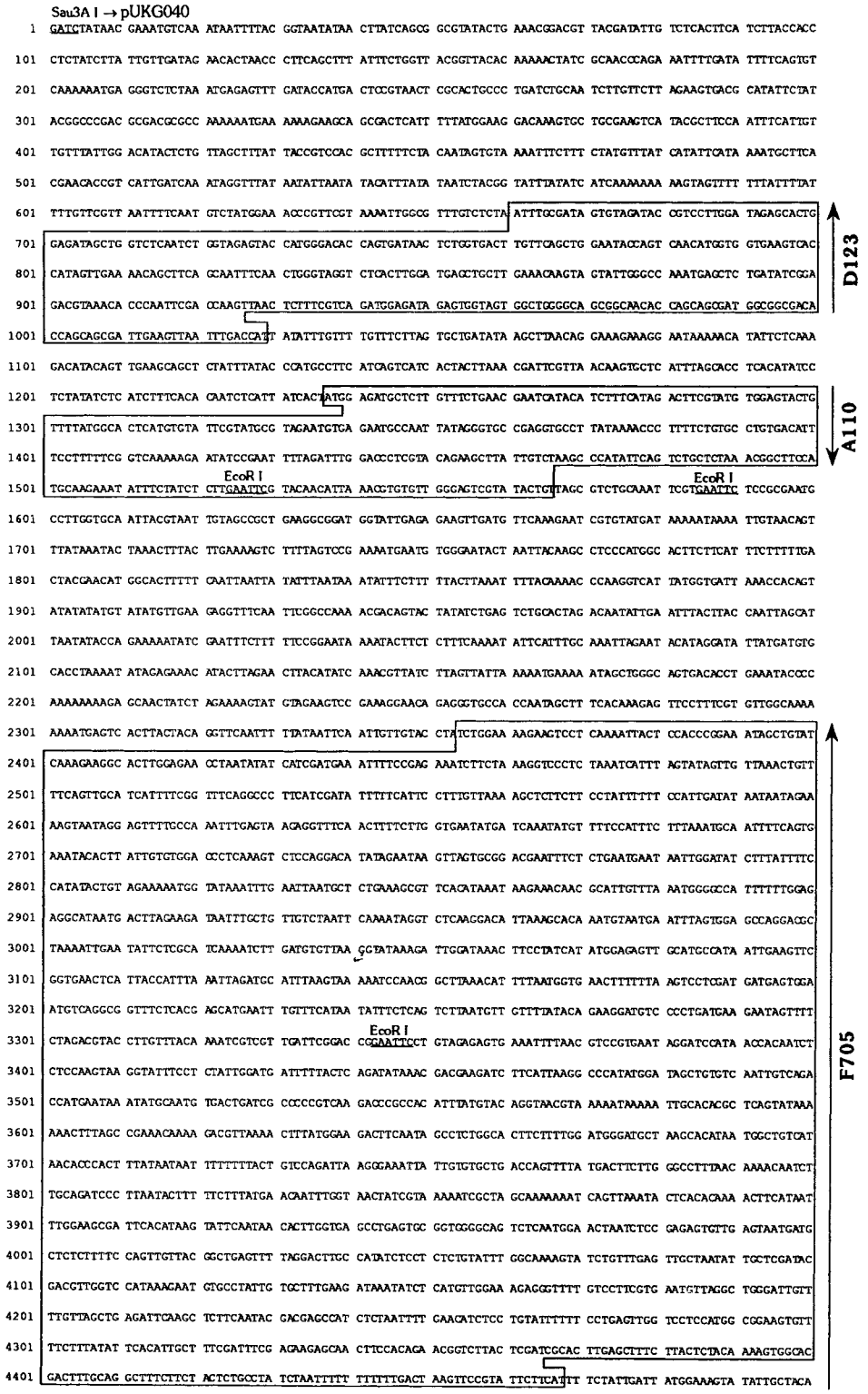


Figure 2.

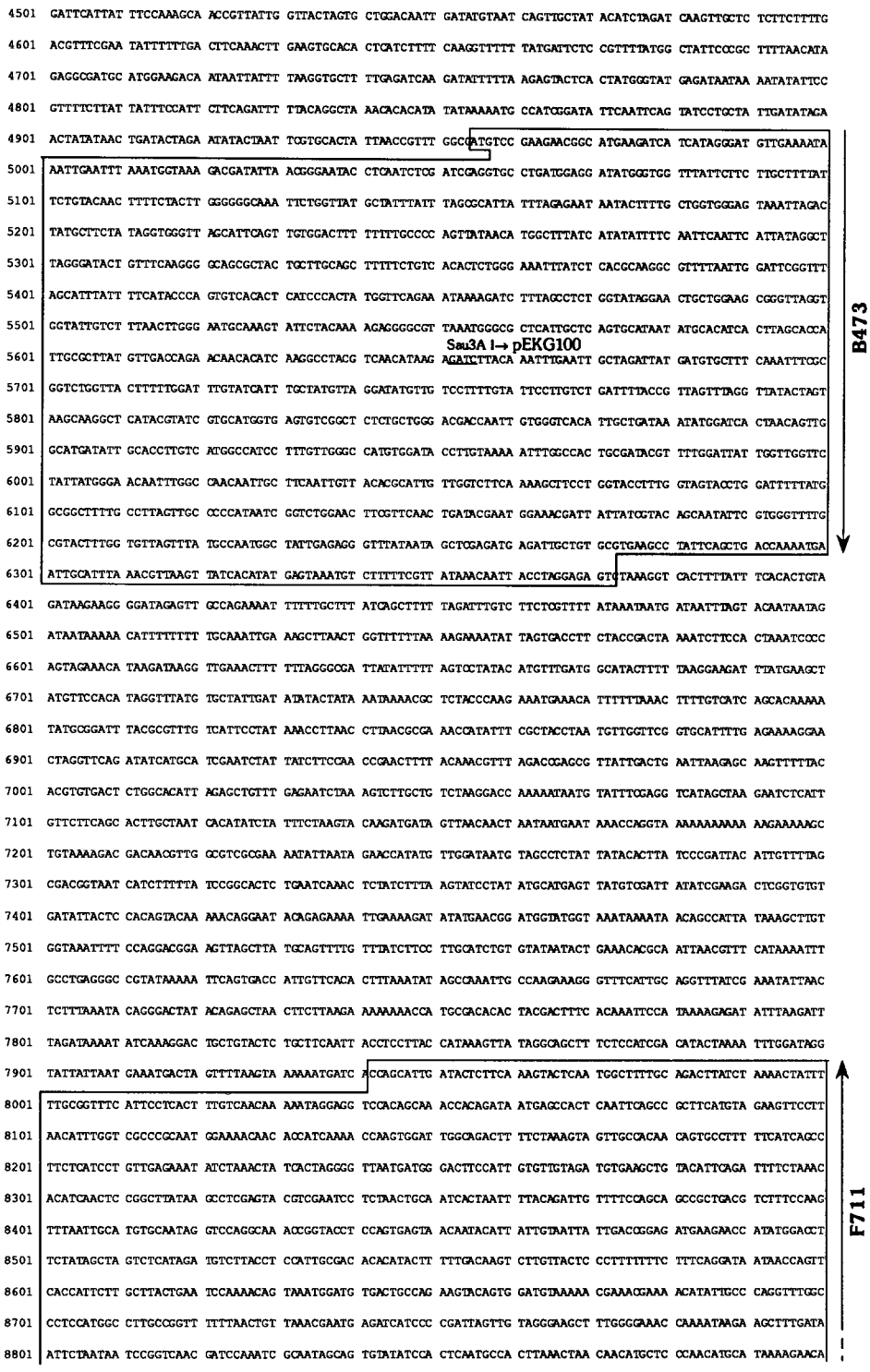


Figure 2. (*Continued*).

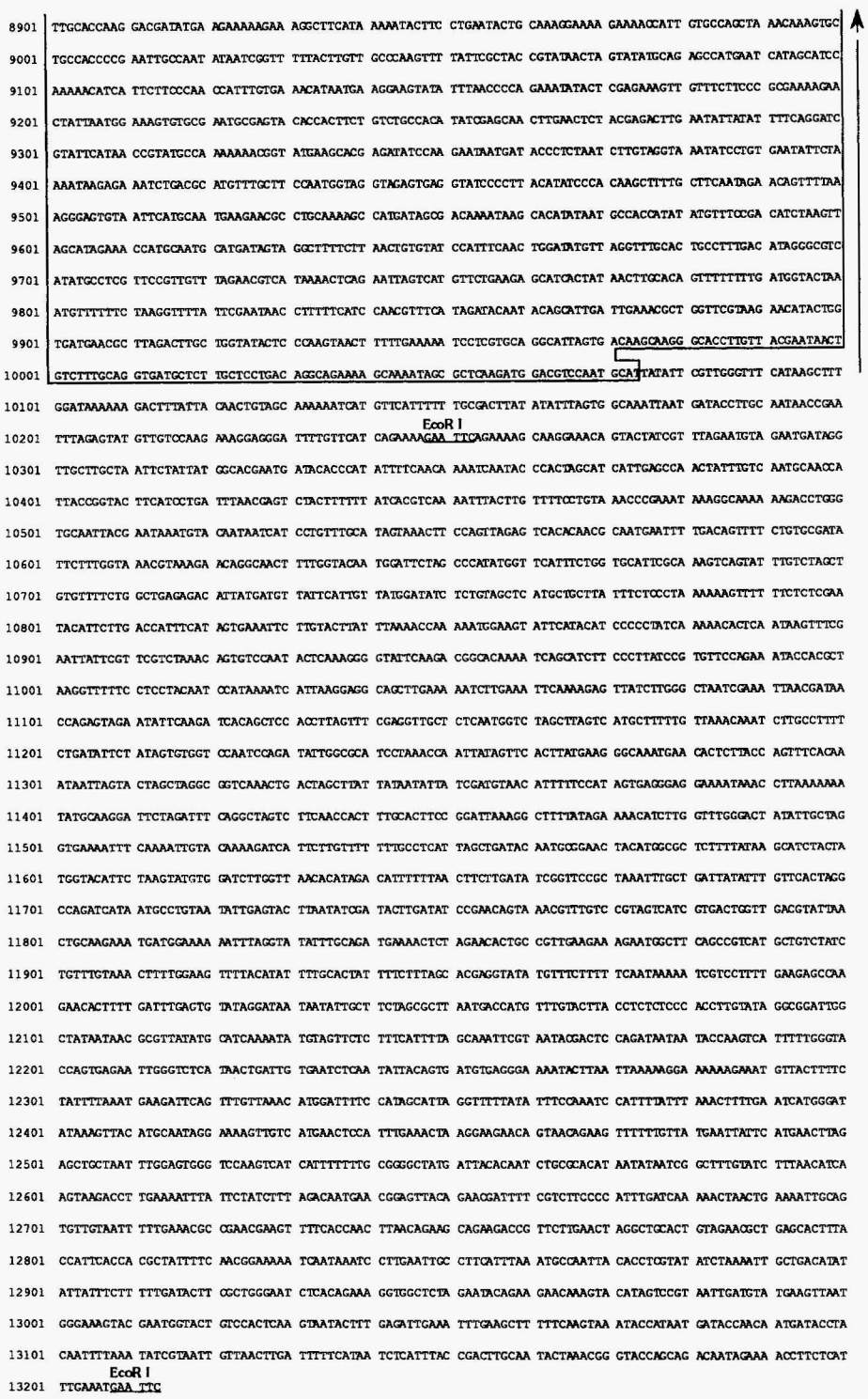


Figure 2. (*Continued*).

Figure 2. Complete sequence of the 13 213 bases of chromosome XI. The sequence reads 5′ to 3′ from the left telomere to the centromere. *Eco*RI sites are underlined. ORFs are boxed. The direction of each ORF is shown by arrow.

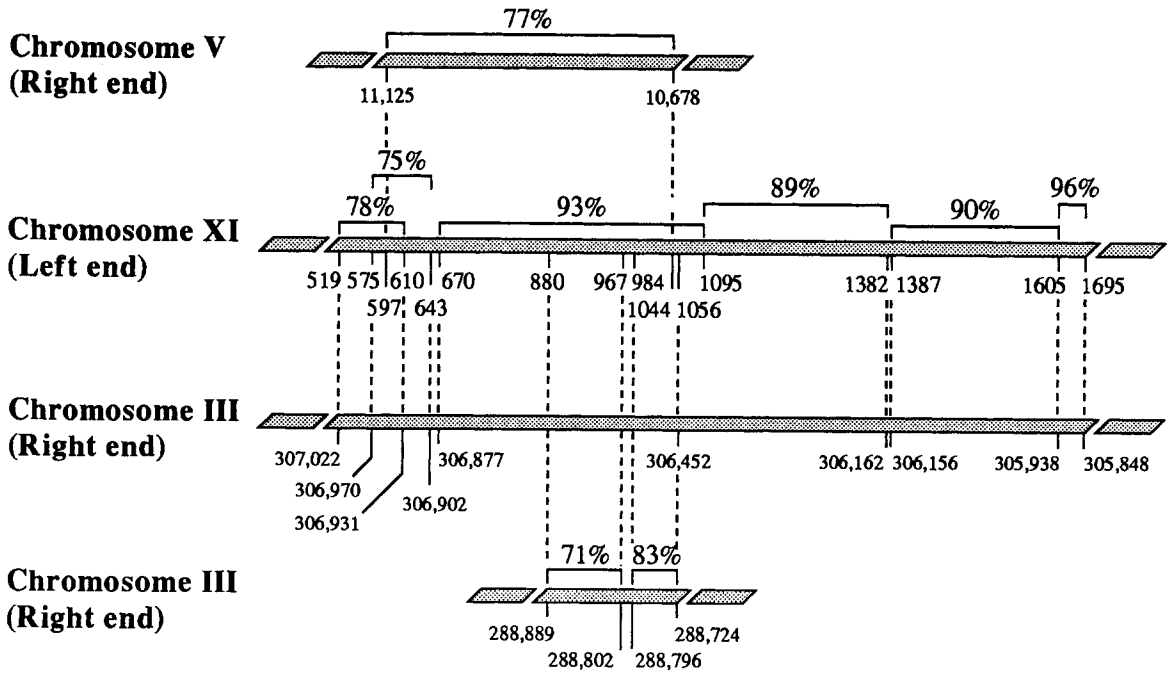


Figure 3. Diagram of the homologies between different chromosomes. Shaded bars indicate the chomosomes. Numbers at the bottom of each bar indicate the base coordinates as given in the BlastA analysis. Numbers on top of each bar indicate the percentage of base pair identities.

internal oligo-priming to fill in the gaps. An average length of 315 nucleotides was read from each sequencing reaction. Readings up to 400 bases were achieved on 4% polyacrylamide gels. Compressions seen at several specific positions were solved by repeating the sequencing reactions using dITP (5 different instances). Sequence assembly was performed manually according to restriction maps and the sequences obtained from PCR connecting fragments. Sequence alignments of both strands were done using the GCG program. The final sequence contained an additional 61-base *Eco*RI fragment following the first (5')*Eco*RI site, which had not been detected at the original gel electrophoretic analysis of the cosmid DNA (Figure 1a).

Sequence analysis

Six phase ORF map analysis of the 13·2 kb fragment revealed five ORFs>100 codons (Figure 1b). Their sizes range from 110 to 711 codons and they constitute 48·2% of the entire sequence (6366 bases). This percentage is a low compared to the average chromosome content in coding sequences and it is probably due to the location of this fragment near the end region of the chromosome (Oliver *et al.*, 1992).

The complete sequence of the 13 213 bases is given in Figure 2. FastA (Pearson and Lipman, 1988) and BlastA (Altschul *et al.*, 1990) analyses of the sequenced segment revealed extensive homologies to known sequences on different yeast chromosomes (Figure 3). More specifically: a) a region of 1182 bases showed an overall identity of 90·7% (ranging from 75% to 100%) to the right arm of chromosome III; b) 159 bases of that showed 77% identity to a second region of chromosome III; c) 638 bases were found 77% identical to the right arm of chromosome V and d) homology has been detected to the right arm of chromosome II (Becker, personal communication). Obviously, these are due to recombination events between fragments near the ends of the mentioned chromosomes, as has been previously reported (Oliver *et al.*, 1992).

Analysis of the putative ORF products

The putative translation products of the identified ORFs have been compared to protein databases using FastA (Table 1). For better evaluation

Table 1. Best optimized FastA scores obtained by the comparison of the putative translation product of each ORF with the protein databases

ORF	Homologous or Identical protein	Optimized score	Highest score	Reference
D123	*S. cerevisiae* Hypothetical protein YCR104W (124aa) 98·3% identity in 120aa	544	554	van der Linden *et al.*, 1992 EMBL: X59720
	S. cerevisiae SYGP-ORF12 (120aa) 86·2% identity in 123aa	485	554	Mulligan *et al.*, 1993, unpublished EMBL: L10830
A110	*S. cerevisiae* Hypothetical protein YCR103C (111aa) 79·3% identity in 111aa	496	655	van er Linden *et al.*, 1992 EMBL: X59720
F705	*S. cerevisiae* CYP1 (HAP1) regulatory protein (1483aa) 28% identity in 130aa	155	3624	Verdiere, 1988 EMBL: X13793
B473	Chinese hamster (*Mev*) mevalonate transporter (494aa) 28·1% identity in 153aa	182	2515	Kim *et al.*, 1992 EMBL: S48888
F711	*S. cerevisiae* (*FRE1*) Ferric reductase (686aa) 24·5% identity in 693aa	542	3723	Dancis *et al.*, 1992 EMBL: M86908

of the significance of each obtained score, we have also included the highest FastA score, obtained by the comparison of each ORF to itself. Optimum scores higher than 200 have been considered as significant. Lower scores due to homologies in restricted areas of the protein sequences indicated conservation of specific domains. Protein patterns (motifs) have been identified by the ProSite program (Bairoch, 1991) of the GCG package. Our findings on each individual ORF are discussed below.

ORF D123 is included in the region which is closely related to the right arms of yeast chromosomes III and V. It is almost identical to YCR104W a hypothetical protein in the HMR 3′ region on chromosome III, and very similar to the STGP-ORF12 encoded by a gene contained in a region of 36 772 base pairs of chromosome V between the known genes MAK10, AFG18 on its 5′ site and CYC7 on its 3′ site (Mulligan *et al.*, unpublished, Mortimer *et al.*, 1989) (Figure 4a). The function of both of these hypothetical proteins remains unknown. As has been already reported for YCR104W (Bork *et al.*, 1992), D123 showed also similarities to the yeast temperature-shock inducible protein TIP1 (Kondo and Inouye, 1991) (27·3% identity in 99 overlapping amino acids, FastA score: 144) and to the yeast serine rich, glucose induced protein SRP1 (Marguet *et al.*, 1988) (26·3% identity in 99 overlapping amino acids, FastA score: 121), the function of which is similarly unknown. All of these proteins, including D123, start with a putative hydrophobic signal sequence, which is followed by a conserved domain of about 90 residues including the stress-induced protein motif (P-W-Y-[ST](2)-R-L). This domain is followed, in SRP1 (total length of 254 amino acids) and TIP1 (total length of 210 amino acids) proteins only, by a repetitive serine and alanine rich region (Figure 4b). According to Kondo and Inouye (1991) and Marguet *et al.* (1988), there is a family of several genes in different chromosomes which cross-hybridize to both *TIP1* and *SRP1* sequences but some of these may not be highly expressed genes, since only three distinct transcripts have been detected. We have not

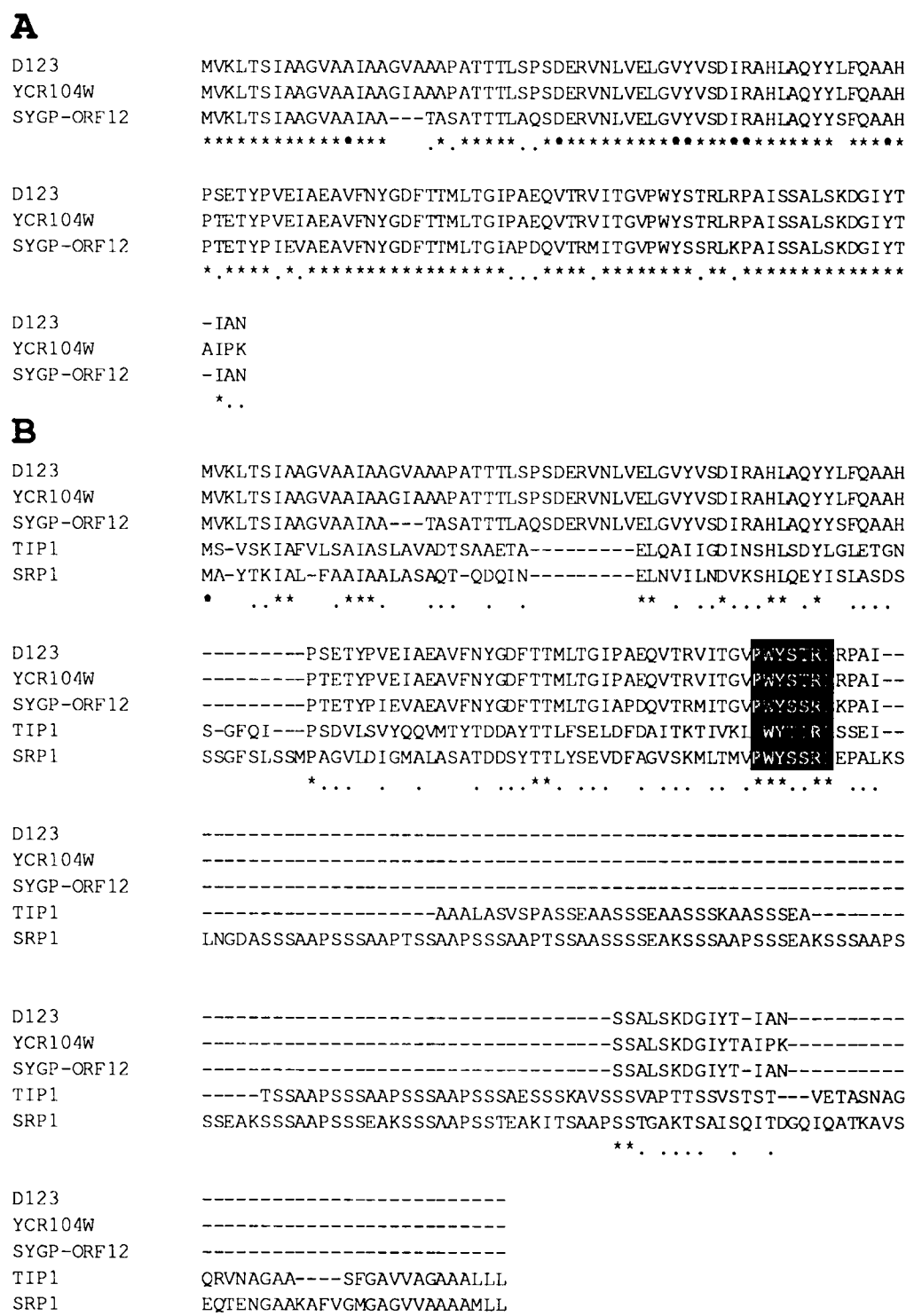


Figure 4. (a) Multiple alignment of the sequences of the D123 ORF, the YCR104W hypothetical protein and SYGP-ORF12 using the CLUSTAL program. (b) Multiple alignment of the sequences of the D123 ORF, YCR104W, SYGP-ORF12, TIP1 and SRP1 proteins. The stress-induced protein motif is shadowed. Asterisks indicate residue identities and dots indicate conservative substitutions.

```
A110        MEMLLFLNESYIFHRLRMWSTVLWHSCVFVCVECENANYRVPR-CLIKPF-SVPVTFPFS
YCR103C     MEMLLFLNESYIFHRFRMWSIVLWHSCVFVCAECGNANYRGAG-VPCKTLLRAPVKFPLS
DYC101      MEMLLFLNESYIFHRFRMWSIVLWHSCVFVCAECGNAYYRGAGGCLEKPF-CAPVKFPFS
            ***************.**** ********** ** ** ** .     * .   ** **.*

A110        VKKNIRILDLDPRTEAYCLSPYSVCSKRLPCKKYFYLLNSYNIKRVLGVVYC
YCR103C     VKKNIRILDLDPRSEAYCLSLNSVCFKRLPRKKYFHLLNSYNIKRVLGVVYC
DYC101      VKKNIRILDLDPRSEAYCLSHHLVCPKRFPCKATSLLL----------IPEG
            *************.******   ** **.* *    **          .
```

Figure 5. Multiple alignment of the sequences of the A110 ORF, YCR103C and DYC101 hypothetical proteins using the CLUSTAL program.

examined the expression of the gene encoding D123 protein. However, its proximity to the telomere could indicate that it is expressed at flow levels or under specific conditions (Sandell and Zakian, 1992). In fact, we have noticed, by FastA analyses, that the TIP1 and SRP1 proteins are more similar to the SYGP-ORF12 (30·6% identity in 108 overlapping amino acids and 29·3% identity in 99 overlapping amino acids, respectively) than they are to the D123 and YCR104W ORFs.

ORF A110 is also contained in the region of homology between chromosomes XI and III. It is similar to the hypothetical protein YCR103C. It contains several non conservative amino acid substitutions which may imply that the two products serve different functions. A third ORF of 101 codons (DYC101 sequence communicated by H. Domdey), that was identified on chromosome II, is also homologous to A110, exhibiting 76·3% identity in 97 overlapping amino acids (FastA score: 442). That ORF resembles YCR103C as much as the A110 does (79·4% identity in 97 overlapping amino acids). A multiple alignment of the three sequences revealed more residue substitutions between A110 and the other two proteins, except for the last 14 amino acids at the carboxy terminus which are identical in A110 and YCR103C ORFs and absent from the chromosome II ORF (Figure 5).

The FastA alignment of the F705 ORF product showed homology to the yeast protein CYP1 (HAP1) only in 130 overlapping residues of its amino terminus (residues 7 to 136 in F705 and 49 to 170 in CYP1). In fact, the alignment varied depending on the program used, except for a stretch of 36 residues, 30 of which make up the fungal Zn (2)-Cys (6) binuclear cluster domain. F705 was also regionally similar to several proteins that contain the same motif (FastA scores: 100–155). All of these proteins are transcriptional activators (GAL4, MAL63, LEU3, etc.) that bind DNA with their cysteine-rich amino terminus in a zinc dependent fashion (Coleman, 1992). The identified motif in F705 starts at residue 23 (SCHFCRVRKLKCDRVRPFCGSCSSRNRKQC) and follows the consensus sequence: [GAS]-C-x(2)-C-[RKH]-x(2)-[RK]-x-[RK]-C-x(5,9)-C-x(2)-C-x(6,8)-C F705 contains also, near its carboxy terminus, a second rare motif characterizing membrane proteins involved in sugar transport, starting at residue 629 (MEKIGRRAFNKG) and following the consensus sequence: [LIVMST]-[DE]-x-[LIVMFA]-G-R-[RK]-x(4,6)-G. However, its significance is questionable since the hydrophobicity profile of F705 protein does not reveal the typical transmembrane domains (data not shown), unless it is a novel type of protein that can mediate both interaction with a sugar and transcriptional regulation.

The FastA search for ORF B473 product revealed low similarities to a number of membrane proteins. The most significant similarity was between its amino terminus and a mammalian membrane protein, the putative transporter of mevalonate (Figure 6a). This similarity does not necessarily indicate a directly homologous molecule but some sort of membrane transporter. In fact, the hydrophobicity profile of ORF B473 showed 12 membrane spanning stretches typical of such proteins (Figure 6b). The profile also included one central and one carboxy terminal hydrophilic region which followed the transmembrane segments 6 and 12 respectively, similarly to the profile of MEV protein and other membrane transporters (Culham *et al.*, 1993). The main difference was at the amino terminus (~30 residues) or ORF B473 which appeared hydrophilic. A transcript corresponding to the ORF B473 was not detected in extracts of cells grown in standard YPD (Guthrie and Fink, 1991) growth medium (data not shown).

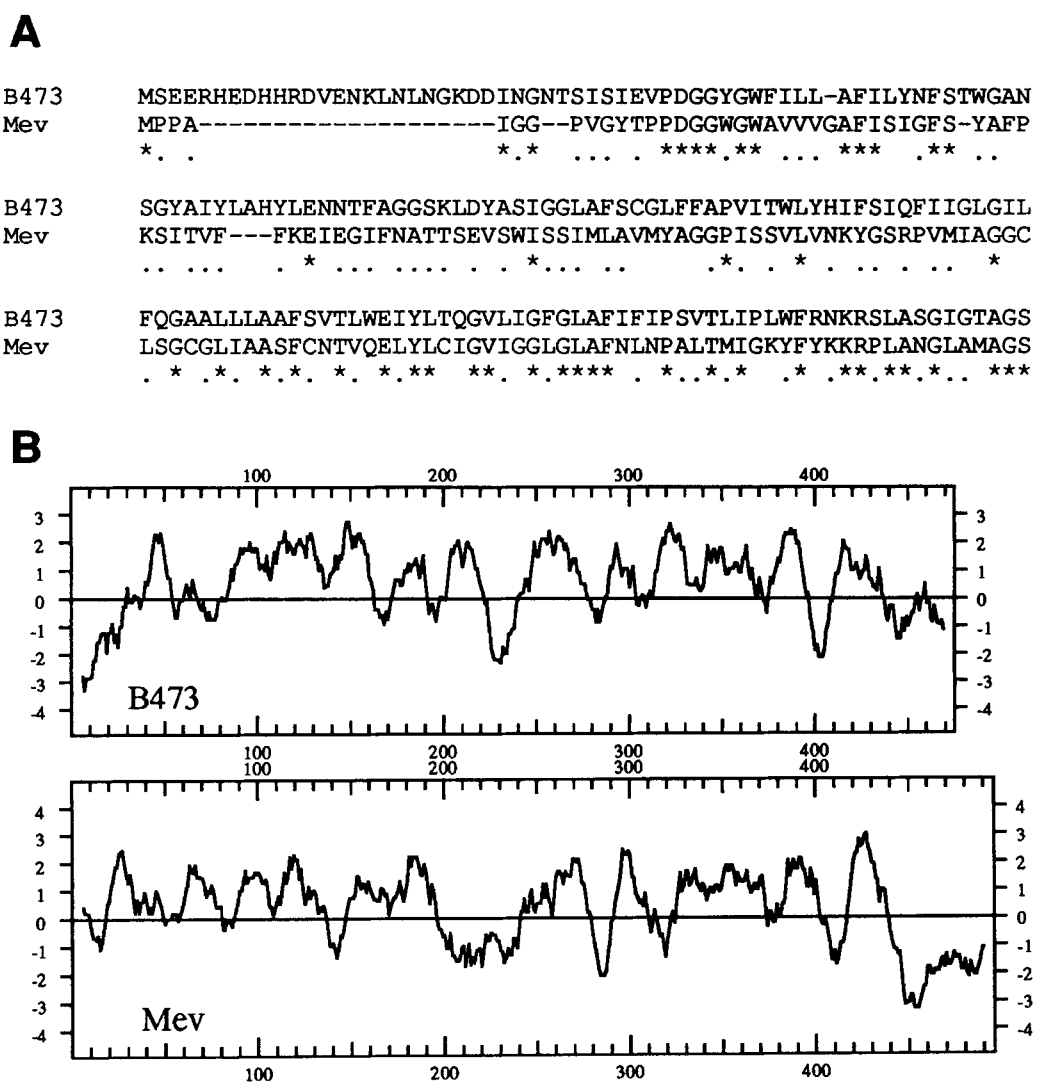


Figure 6. (a) Alignment of the 180 aminoterminal residues of ORF B473 with the 156 aminoterminal residues of the mammalian membrane transporter MEV. (b) Hydrophobicity profiles (Kyte and Doolittle, 1982) of ORF B473 and MEV protein.

This may indicate gene expression under specific conditions or gene repression related to its position near the telomere.

The product of ORF F711 showed a significant similarity to the known yeast FRE1 ferric reductase. We have proven by biochemical, genetic and structural analyses that it is also a membrane protein that can reduce the environmental ferric iron to its intracellular ferrous form. The expression of this non-essential gene is down-regulated by the presence of iron in the growth medium and its RNA is not detectable under standard YPD medium growth conditions (Georgatsou and Alexandraki, submitted for publication).

ACKNOWLEDGEMENTS

We thank Bernard Dujon for the excellent coordination of the Chromosome XI sequencing project. We thank Irmi Becker and all MIPS staff for help with the sequence analysis. We thank Horst Domdey for allowing us to use the unpublished

DYC101 hypothetical protein from chromosome II. We also thank Yannis Papanikolaou for help with the computer analyses at IMBB, Georgia Houlaki for help with the preparation of the figures and the co-contractor George Thireos for helpful discussions. This work was supported by the Commission of the European Communities under the BRIDGE program of the Division of Biotechnology and by the Greek Ministry of Industry, Energy and Technology.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. and Struhl, K. (Eds) (1987). *Current Protocols in Molecular Biology.* Green Publishing Associates and Wiley-Interscience.

Bairoch, A. (1991). A dictionary of sites and patterns in proteins. *Nucl. Acid Res.* **16,** 2241–2245.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E. (1992). Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Protein Sci.* **1,** 1677–1690.

Coleman, J. E. (1990). Zinc proteins: enzymes storage proteins, transcription factors and replication proteins. *Annu. Rev. Biochem.* **61,** 897–946.

Culham, D. E., Lasby, B., Marangoni, A. G., Milner, J. L., Steer, B. A., van Nues, R. W. and Wood, J. M. (1993). Isolation and sequencing of *Escherichia coli* gene *proP* reveals unusual structural features of the osmoregulatory proline/betaine transporter, ProP. *J. Mol. Biol.* **229,** 268–276.

Dancis, A., Roman, D. J., Anderson, G. J., Hinnenbush, A. G. and Klausner, R. D. (1992). Ferric reductase of Saccharomyces cerevisiae: Molecular characterization, role in iron uptake and transcriptional control by iron. *Proc. Natl. Acad. Sci. U.S.A.* **89,** 3869–3873.

Georgatsou, E. and Alexandraki, D. Two distinctly regulated genes are required for ferric reduction, the first step of iron uptake in *Saccharomyces cerevisiae.* (Submitted for publication)

Guthrie, C. and Fink, G. R. (Eds) (1991). Guide to yeast genetics and molecular biology. *Methods in Enzymology,* vol. **194.** Academic Press, New York.

Higgins, D. G. and Sharp, P. M. (1988). Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73,** 237–244.

Kim, C. M., Goldstein, J. L. and Brown, M. S. (1992). cDNA cloning of MEV, a mutant protein that facilitates cellular uptake of mevalonate and identification of the point mutation responsible for its gain of function. *J. Biol. Chem.* **267,** 23113–23121.

Kondo, K. and Inouye, M. (1991). *TIP1*, a cold shock inducible gene of *Saccharomyces cerevisiae. J. Biol. Chem.* **266,** 17537–17544.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157,** 105–132.

Marck, C. (1988). 'DNA Strider': a 'C ' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computer. *Nucl. Acid Res.* **16,** 829–1836.

Marguet, D., Guo, X. J. and Lauquin, G. J.-M. (1988). Yeast gene *SRP1* (serine rich protein)-Intragenic repeat structure and identification of a family of *SRP1* related DNA sequences. *J. Mol. Biol.* **202,** 455–470.

Mortimer, R. K., Schild, D., Contopoulou, C. R. and Kans, J. A. (1989). Genetic map of *Saccharomyces cerevisiae*, Edition 10. *Yeast* **5,** 321–403.

Oliver, S. G. *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357,** 38–46.

Pearson, V. R. and Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. U.S.A.* **85,** 2444–2448.

Sandell, L. L. and Zakian, V. A. (1992). Telomeric position effect in yeast. *Trends Cell Biol.* **2,** 10–14.

Tzermia, M., Horaitis, O. and Alexandraki, D. (1994). The complete sequencing of a 24·6 kb segment of yeast chromosome XI identified the known Loci *URA1, SAC1* and *TRP3*, and revealed 6 new open reading frames including homologues to the threonine dehydratases, membrane transporters, hydantoinases and the phospholipase A_2-activating protein. *Yeast* **10.**