YEAST VOL. **13:** 373–377 (1997)

Yeast Sequencing Reports

Analysis of a 17·9 kb Region from *Saccharomyces cerevisiae* Chromosome VII Reveals the Presence of Eight Open Reading Frames, Including *BRF1* (TFIIIB70) and *GCN5* Genes

FIORELLA FEROLI[1], GIOVANNA CARIGNANI[1], ANNA PAVANELLO[1], PAULO GUERREIRO[2], DULCE AZEVEDO[2], CLAUDINA RODRIGUES-POUSADA[2], PASQUALE MELCHIORETTO[3], LUCIA PANZERI[3] AND MARIA LUISA AGOSTONI CARBONE[3]*

[1]*Dipartimento di Chimica Biologica, Università di Padova, via Trieste 75, 35121 Padova, Italy*
[2]*Laboratorio de Genética Molecular, Instituto Gulbenkian de Ciência, Apartado 14, P-2781 Oeiras Codex, Portugal*
[3]*Dipartimento di Genetica e di Biologia dei Microrganismi, Università di Milano, via Celoria 26, 20133 Milano, Italy*

Received 14 August 1996; accepted 27 August 1996

We report the nucleotide sequence of a 17 898 bp DNA segment from the right arm of *Saccharomyces cerevisiae* chromosome VII. This fragment begins at 482 kb from the centromere. The sequence includes the *BRF1* gene, encoding TFIIIB70, the 5′ portion of the *GCN5* gene, an open reading frame (ORF) previously identified as ORF *MGA1*, whose translation product shows similarity to heat-shock transcription factors and five new ORFs. Among these, YGR250 encodes a polypeptide that harbours a domain present in several polyA binding proteins, YGR245 is similar to a putative *Schizosaccharomyces pombe* gene, YGR248 shows significant similarity with three ORFs of *S. cerevisiae* situated on different chromosomes, while the remaining two ORFs, YGR247 and YGR251, do not show significant similarity to sequences present in databases. This sequence has been submitted to the EMBL data library under Accession Number Y07703.

Yeast 13: 373–377, 1997.
No. of Figures: 1. No. of Tables: 1. No. of References: 17.

KEY WORDS — *Saccharomyces cerevisiae*; chromosome VII; *BRF1*; *MGA1*; *SOL1*; *GCN5*

INTRODUCTION

As part of the European Union BIOTECH project for systematic sequencing of the yeast genome, we have determined the nucleotide sequence of a DNA fragment of 17 898 bp from the right arm of chromosome VII. The fragment was part of two partially overlapping cosmid clones, pUGH1273 and pEGH301, and is situated in a centromere-proximal position with respect to the *RAD2* gene. The overlapping regions (15 723 bp) were independently sequenced in two different laboratories with a 100% identity in the resulting sequences.

*Correspondence to: Maria Luisa Agostoni Carbone.
Contract grant sponsor: European Union (BIOTECH II programme).

CCC 0749–503X/97/040373–05 $17.50

MATERIALS AND METHODS

Strains, cosmids and vectors

The two cosmid clones pEGH301 and pUGH1273 were provided by H. Tettelin

(Université Catholique de Louvain). pEGH301 (sequenced in Lisboa) consists of a 43 kb DNA fragment cloned into the vector pWE15. pUGH1273 (sequenced in Milano and Padova) contains a DNA fragment of about 36 kb cloned into the vector pOU61cos. The fragments derive from chromosome VII of the *Saccharomyces cerevisiae* strain FY1679, isogenic with S288C (Winston *et al.*, 1995). A restriction map of the cosmid inserts was determined by using several restriction enzymes. Plasmids pUC19 and pGEM-7Zf(+) were used for subcloning experiments. *Escherichia coli* XL1-Blue, JM109 and INVaF′ were used for transformation and amplification of the plasmids. Recombinant DNA methods followed standard protocols (Sambrook *et al.*, 1989).

Sequencing strategy

Cosmid pEGH301: random libraries were generated by digestion with the *Bam*HI, *Bgl*II, *Cla*I, *Eco*RI, *Hin*dIII and *Sfu*I restriction enzymes. Overlapping clones that cover the entire DNA fragment were chosen to generate subclones suitable for sequencing. These were generated either by nested deletions (using the Pharmacia Exonuclease III nested deletion kit) or by direct deletions with suitable restriction enzymes. Specific oligonucleotide-directed sequencing filled a few gaps.

Cosmid pUGH1273: fragments derived from digestion with *Eco*RI or from double digestion with *Bam*HI and *Kpn*I were subcloned into the pGEM7Zf(+) plasmid vector. The sequence of the two strands was determined on two independent subclones of each fragment, by using either primer walking or nested ExoIII deletions.

DNA sequencing reactions were performed by the dideoxy-chain-termination method (Sanger *et al.*, 1977) using either the T7 DNA polymerase kit (Pharmacia) or the Taq polymerase 'cycle sequencing' system (Applied Biosystem). Both manual and automated (ABI373 DNA sequencer, Applied Biosystem) procedures were used to analyse the sequencing reactions.

Computer analysis

Sequences were assembled and analysed using the following software: DNASIS 5.0 (LKB-Hitachi), DNA Strider 1.1 (Marck, 1988) and the GCG software package from Wisconsin University (Devereux *et al.*, 1984). Searches for similarity of proteins to entries in the databases were carried out using FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.*, 1990). The results were complemented with the analysis performed by K. Kleine at MIPS (Martinsried, Germany): the codon adaptation index (CAI) was calculated with the program CODONS (Lloyd and Sharp, 1992); prediction of transmembrane spans was performed with the program ALOM (Klein *et al.*, 1985), motifs were calculated with the GCG package and PROSITE (Bairoch *et al.*, 1996).

RESULTS AND DISCUSSION

General organization of the sequence

The 17 898 bp sequence of the left part of the cosmid clone pUGH1273, which includes the 15 723 bp sequence of the right part of the cosmid clone pEGH301, was submitted to the EMBL database (accession no. Y07703). The complete sequence includes 11 open reading frames (ORFs) longer than 300 bp. Figure 1 shows the *Eco*RI restriction map of the fragment (A) and the localization of the ORFs (B). The three ORFs G9105 (position 439–993 W), G9115 (position 3066–3503 W) and G9150 (position 17421–17756 C) are internal ORFs; therefore they probably do not represent expressed genes and will not be described.

Three large intergenic regions of 1·3 kb (position 7257–8568), 1·76 kb (position 9937–11698) and 2·1 kb (position 14042–16158) respectively do not contain ORFs longer than 99 codons. The 1·76 kb region includes an ARS-consensus sequence (position 11620–11630 C).

The strand orientation and base positions of the ORFs on the 17 898 bp fragment, the length of the deduced proteins and their CAI values are shown in Table 1. Homologies with known proteins in databases are reported only when FASTA scores are higher than 100. The sequence of three ORFs was already present in databases: two of them represent *S. cerevisiae* genes functionally characterized, although not mapped on chromosome VII.

YGR246 corresponds to the essential *BRF1* gene (synonyms *TDS4*, *PCF4*; Colbert and Hahn, 1992; Lopez-de-Leon *et al.*, 1992; Buratowski and Zhou, 1992), encoding the 70 kDa subunit of transcription factor IIIB (TFIIIB70). TFIIIB plays a central role in transcription initiation by RNA polymerase III on genes encoding tRNAs, 5S rRNA and other small structural RNAs, and TFIIIB70 interacts

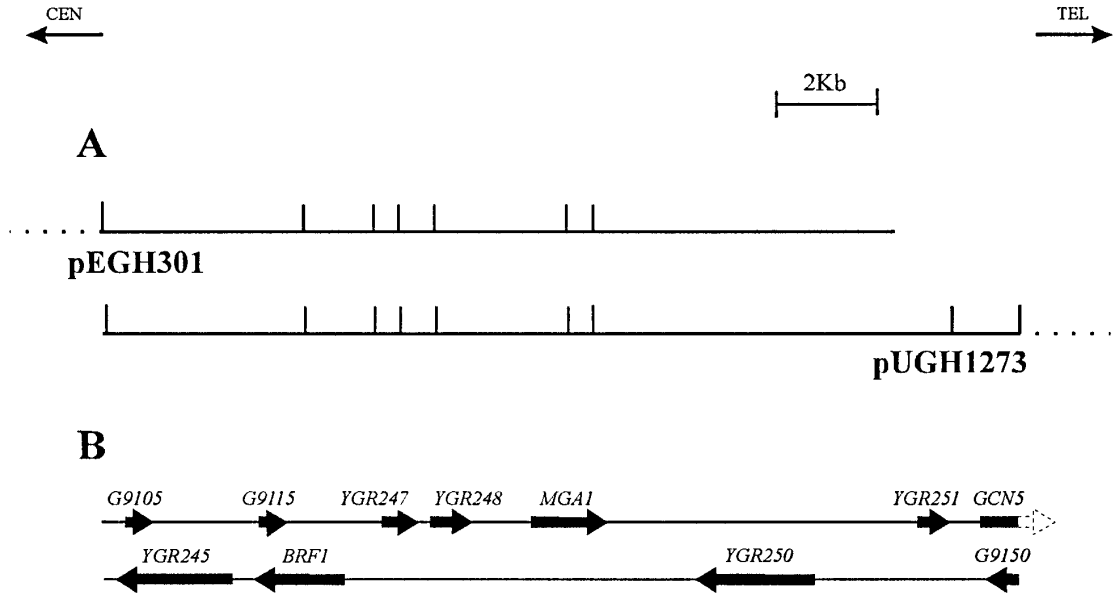

Figure 1. (A) *Eco*RI restriction map of the DNA fragment analysed. The portion of each cosmid whose sequence is presented here is indicated by a continuous line; the upper line shows the 3′ end of cosmid pEGH301 insert; the lower line shows the 5′ end of cosmid pUGH1273 insert. (B) Position and orientation of the ORFs on the 17·9 kb fragment.

Table 1. Characteristics of the open reading frames (ORFs) included in the reported fragment of chromosome VII.

Name of ORF	Position[a]		Size (aa)	CAI[b]	Identical or similar to	FASTA score
YGR245	288–2588	(C)	767	0·20	Hypothetical protein (*S. pombe*)	514/3600
YGR246	3005–4792	(C)	596	0·18	Is *BRF1* (transcription factor TFIIIB70 kDa)	2739/2739
YGR247	5486–6202	(W)	239	0·11		
YGR248	6492–7256	(W)	255	0·12	Hypothetical protein YHR163 (*S. cerevisiae*)	749/1281
YGR249	8569–9936	(W)	456	0·11	Is *MGA1*	2084/2084
YGR250	11699–14041	(C)	781	0·14	64 K polyadenylation factor (human)	174/3737
YGR251	16159–16746	(W)	196	0·12		
YGR252	17389–17898	(W)	[c]	0·12	Is *GCN5*	

[a](W) denotes location on the Watson strand, (C) on the Crick strand.
[b]Codon adaptation index according to Lloyd and Sharp (1992).
[c]N-terminal fragment of 170 amino acids.

both with the TATA-binding protein and a subunit of polymerase III (Chaussivert *et al.*, 1995). The protein is evolutionarily related to TFIIB, showing two repeats of the TFIIB signature (positions 119–134 and 217–232). Nucleotide sequence alignment with two sequences present in the databases (accession no. L00630 and M91073) showed one base substitution in the 5′ upstream region, in positions -175 and -194 respectively. The protein also includes a cytochrome *c* family heme-binding site signature.

YGR252 The N-terminal portion (170 amino acids) of the ORF is included in the reported fragment, while the 3′ portion was sequenced elsewhere (L. Frontali, personal communication). YGR252 is identical to the *GCN5* gene, which encodes a general transcriptional regulator that

cooperates with transcriptional activators such as *GCN4* and the HAP2-HAP3-HAP4 complex to promote high levels of transcription. The gene had been sequenced, but not mapped (Georgakopoulos and Thireos, 1992).

YGR249 The nucleotide sequence of ORF YGR249 is identical to that submitted to the PIR database as gene *MGA1* (accession no. D29626). YGR249 encodes a 456 amino-acids long protein of unknown function displaying weak similarity to heat-shock transcription factors from a variety of organisms, from yeast to man (FASTA scores ranging from 226 to 116). The region from amino acid 6 to 116 has the features of a heat-shock factor DNA binding domain. The putative protein has a molecular weight of 50·7 kDa and is rich (14·3%) in serine residues. The CAI (0·11) suggests a low expression rate.

Five putative proteins are encoded by novel genes

YGR245 This ORF shows similarity to an unknown protein from *Schizosaccharomyces pombe* (accession no. Z69909). The putative protein is rich in acidic residues and shows in its C-terminal third (position 538–657) an acidic region in which 50 of 120 residues are either glutamate or aspartate. A putative transmembrane domain was found (Klein *et al.*, 1985) in position 374–390.

YGR247 No significant similarity was found in databank searches (FASTA scores of less than 100). The CAI value (0·11) suggests a low level of expression.

YGR248 The putative protein shows similarity to the *S. cerevisiae* hypothetical protein YHR163 (chromosome VIII, accession no. S48903), as well as to the proteins of unknown function encoded by the YCRX13 ORF (chromosome III, accession no. S53589, FASTA 384) and by the *SOL1* gene (chromosome XIV, accession no. S62015, FASTA 391). A putative transmembrane domain is present near the C-terminus (position 189–205).

YGR250 This ORF encodes a 781 amino acids long putative protein which shows similarity to several RNA-associated proteins from various organisms. The alignment with several poly(A)-binding proteins showed that the region situated between positions 196 and 275 has the characteristics of an RNA-binding domain, with the two typical consensus sequences RNP1 and RNP2 fairly conserved (Burd *et al.*, 1991; Query *et al.*, 1989). This protein is also characterized by the presence of another RNP1 signature, at position 600–607, and of a proline-rich carboxyl-terminal domain, which is typical of most RNA-binding proteins (Query *et al.*, 1989).

YGR251 No significant similarity was found with proteins present in the databases (FASTA scores lower than 100).

Summary

In summary, the 17 898 bp analysed contain a putative replication origin (ARS-consensus sequence), seven complete non-internal ORFs longer than 100 codons and part of an eighth ORF. No introns were found associated with ORFs. The size of the ORFs ranges from 196 to 781 amino acids. In this fragment approximately 64% of the DNA is potentially coding and the gene density amounts to one gene per 2·2 kb. This value is similar to the average gene density found in complete yeast chromosomes. In this region, the (G+C) content is 38·9%, which corresponds to the average value for the yeast genome.

ACKNOWLEDGEMENTS

We gratefully acknowledge H. Tettelin (Université Catholique de Louvain, Belgium) for providing cosmids pUGH1273 and pEGH301 and K. Kleine (MIPS, Martinsried, Germany) for help in computer analysis. This work was supported by the European Union BIOTECH II programme.

REFERENCES

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

Bairoch, A., Bucher, P. and Hofmann, K. (1996). The PROSITE database, its status in 1995. *Nucl. Acids Res.* **24,** 189–196.

Buratowski, S. and Zhou, H. (1992). A suppressor of TBP mutations encodes an RNA polymerase III transcription factor with homology to TFIIB. *Cell* **71,** 221–230.

Burd, C. G., Matunis, E. L. and Dreyfuss, G. (1991). The multiple RNA-binding domains of the mRNA poly(A)-binding protein have different RNA-binding activities. *Mol. Cell. Biol.* **11,** 3419–3424.

Chaussivert, N., Conesa, C., Shaaban, S. and Sentenac, A. (1995). Complex interaction between yeast TFIIIB and TFIIIC. *J. Biol. Chem.* **270,** 15353–15358.

Colbert, T. and Hahn, S. (1992). A yeast TFIIB-related factor involved in RNA polymerase III transcription. *Genes Dev.* **6,** 1940–1949.

Devereux, J., Haeberli, P. and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **12,** 387–395.

Georgakopoulos, T. and Thireos, G. (1992). Two distinct yeast transcriptional activators require the function of the GCN5 protein to promote normal levels of transcription. *EMBO J.* **11,** 4145–4152.

Klein, P., Kanehisa, M. and De Lisi, C. (1985). The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta* **815,** 468–476.

Lloyd, A. T. and Sharp, P. M. (1992). CODONS: a microcomputer program for codon usage analysis. *Y. Heredity* **83,** 238–240.

Lopez-de-Leon, A., Librizzi, M., Puglia, K. and Willis, I. M. (1992). PCF4 encodes an RNA polymerase III transcription factor with homology to TFIIB. *Cell* **71,** 211–220.

Marck, C. (1988). 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple MacIntosh family of computers. *Nucl. Acids Res.* **16,** 1829–1836.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.

Query, C., Bentley, R. C. and Keene, J. D. (1989). A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. *Cell* **57,** 89–101.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). *Molecular Cloning*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.

Winston, F., Dollard, C. and Ricupero-Hovasse, S. L. (1995). Construction of a set of convenient *Saccharomyces cerevisiae* strains that are isogenic to S288C. *Yeast* **11,** 53–55.

