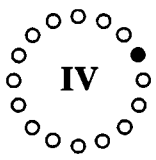
YEAST VOL. 12: 1085–1090 (1996)



Yeast Sequencing Reports

Sequencing and Analysis of a 35·4 kb Region on the Left Arm of Chromosome IV from *Saccharomyces cerevisiae* Reveal 23 Open Reading Frames

LIV GUNN EIDE, CHRIS SANDER AND HANS PRYDZ*

Biotechnology Centre of Oslo, PO Box 1125, Blindern, N-0316 Oslo, Norway

Received 15 March 1996; accepted 22 April 1996

The complete DNA sequence of cosmid clone 31A5 containing a 35 452 bp segment from the left arm of chromosome IV from *Saccharomyces cerevisiae*, was determined from an ordered set of subclones in combination with primer walking on the cosmid. The sequence contains 23 open reading frames (ORFs) of more than 100 amino acid residues and the tRNA-Val2a gene. Five ORFs corresponded to the known yeast genes *SNQ2*, *SES1*, *GCV1*, *RPL2B* and *RPS18A*. The DNA sequence for *RPS18A* is interrupted by an intron. One ORF corresponded to a part of the yeast gene *HEX2* at the end of the cosmid insert. Four ORFs encoded putative proteins which showed strong homologies to other previously known proteins, three of yeast origin and one of non-yeast origin. Two ORFs were classified as having borderline homologies: one had similarity to two protein families and another to two protein products of unknown function from other species. The remaining 11 ORFs bore no significant similarity to any published protein. The complete DNA sequence has been submitted to the EMBL data library, Accession Number X95966.

KEY WORDS — *Saccharomyces cerevisiae*; chromosome IV; *SNQ2*; *SES1*; *GCV1*; *RPL2B*; *HEX2/SRN1*; *RPS18A*; *tRNA-Val12a*

INTRODUCTION

Within the framework of the European Union BIOTECH programme of sequencing the *Saccharomyces cerevisiae* genome, we have sequenced a 35 452 bp fragment (cosmid clone 31A5) on the left arm of chromosome IV. This fragment is located near the centromere, and flanked by the yeast insert from cosmid clone 31D12 (Urrestarazu) and 2M15 (Arnold).

MATERIALS AND METHODS

Cosmid, vectors, strain and oligonucleotides

Cosmid 31A5, containing a 35·5 kb insert, obtained by a partial *Mbo*I digest of yeast chromosome IV DNA and cloned into the *Bam*HI site of vector pWE15, was received from Dr C. Jacq (Laboratoire de Génétique Moléculaire, Ecole Normale Superieure, Paris). Subcloning was performed in pUC18 or pBluescript vectors. *Escherichia coli* strain DH5α was used to obtain all subclones and deletion clones. Synthetic oligodeoxynucleotide primers were made at the Biotechnology Centre of Oslo (Dr E. Babaie).

Sequencing strategies and methods

An ordered sequencing strategy was used to determine the sequence of 31A5. The cosmid DNA 31A5 was digested with *Eco*RI and the fragments were separated on an agarose gel. Eight restriction fragments in the range of 200–7400 bp were isolated and subcloned. Another three smaller

*Corresponding author.

CCC 0749–503X/96/101085–06

fragments (67–147 bp) were detected by sequencing.

Restriction maps of the subclones were generated to construct second-order restriction clones. All clones were sequenced on both strands with universal and/or reverse primers in combination with primer walking.

The sequences of the remaining first-order *Eco*RI fragments, shorter than 200 bp or longer than 7400 bp, were obtained by using primer walking strategy directly on the cosmid. To establish the order of the *Eco*RI fragments, primers were designed and sequences obtained with the entire cosmid as a template. DNA templates for sequence reactions were prepared using Tip20 columns (Qiagen) or by CsCl-gradient purification. Double-stranded dideoxy DNA sequencing was carried out using the Autoread Kit (Pharmacia) with simplified denaturation (Zimmerman *et al.*, 1990) and T7 DNA polymerase, and with the modifications necessary for using internal labelling with fluorescein-15-dATP (Voss *et al.*, 1992). For the direct sequencing on cosmid DNA, we modified the protocol further by using only 5 μg template, 40 pmol primer, denaturing at 85°C and increasing the amount of T7 DNA-polymerase to 10 U per reaction.

Walking primers were designed so that they either ended with a dATP or the first nucleotide to be incorporated in the sequencing reactions was a (fluorescent) dATP (Wiemann *et al.*, 1995).

Hardware and software

DNA sequencing was carried out using a standard automated ALF DNA sequencer (Pharmacia). Raw data collection and evaluation were performed with the ALF Manager software (Pharmacia). Fragment assembly and primer design were done with the GeneSkipper program (C. Schwager, EMBL). Open reading frames (ORFs), codon preferences and the codon usage profile were calculated with GeneSkipper and the UWGCG program package in parallel.

RESULTS AND DISCUSSION

Sequence determination

The final sequence of 35 452 bp was determined by a total number of 266 overlapping fragments and a redundancy of 3.0. A total of 175 walking primers were designed, and the average reading length was 406 bases.

Sequence analysis

Analysis of the sequence revealed 23 ORFs of more than 100 amino acid residues and a gene for tRNA-Val2a. The ORFs were distributed in all six reading frames and also evenly over the whole cosmid, except for a 2 kb fragment where no ORF was detected. All ORFs were provisionally named by MIPS: pz, followed by a letter for the frame and the number of amino acids deduced from the nucleotide sequence (Table 1). Sequences of putative genes were then aligned with updated protein databases. The deduced amino acid sequences were also analysed using the GeneQuiz automated search program (Casari *et al.*, 1995). The codon preference values (Table 1) were calculated using a yeast codon usage table made out of 435 genes from *S. cerevisiae* found in GenBank 63 (produced by J. Michael Cherry with the GCG program CodonFrequency). The lower the value, the greater the probability that the ORF represents a functional gene. Table 1, in addition, shows the best optimized FASTA scores and the corresponding homologous proteins.

Open reading frames

Five of the ORFs encoded genes identical or highly similar to known yeast genes. In addition, one ORF encoded the end of a known yeast gene. Four ORFs encoded putative proteins which showed strong homologies to other known proteins, three of yeast origin and one of non-yeast origin (Table 1). Two ORFs were defined to be borderline cases with FASTA scores of 201 and 169. No significant similarities were observed for the other 11 ORFs. They showed optimized FASTA scores below 125 and less than 29% identity over maximum 174 amino acid residues to known proteins.

Five ORFs encoded yeast genes already known. ORF pzA462 encoded *S. cerevisiae* seryl-tRNA synthetase, Ses1p (Weygand-Durasevic *et al.*, 1987). The ATP-dependent permease conferring hyper-resistance to certain chemicals, Snq2p (Servos *et al.*, 1993), was encoded by ORF pzB1501. ORF pzD400 encoded the *S. cerevisiae* glycine cleavage T protein, Gcv1p (McNeil *et al.*, AC L41522). Our cosmid sequence for the above three genes differed in each gene by one single nucleotide from that already published. These differences also caused differences in the amino acid sequence (Table 2). The calculated codon preferences for ORFs pzB362 and pzA156i gave rather

Table 1. Analysis of ORFs in cosmid 31a5. Provisional names of ORFs assigned by MIPS. Codon preference values calculated from yeast codon usage table based on 435 genes from *S. cerevisiae* in GenBank 63. The deduced amino acid sequences of the ORFs were compared with available data in relevant libraries. Two cases of borderline similarity are in brackets.

ORF	MIPS	Codon preference	FASTA score	% Identity and length of overlap		Homologous protein	Organism
1	pzA520	0·91	2082	74%	517 aa	Galactokinase, Gal1p	*S. cerevisiae*
2	pzA208	1·69	201	27%	190 aa	['KIAA0186 protein', unknown function]	*H. sapiens*
3	pzA114	2·88	77	25%	80 aa	No homology observed	
4	pzA462	0·7	2328	99%	462 aa	Seryl-tRNA synthetase, Ses1p	*S. cerevisiae*
5	pzA161	2·43	72	26%	97 aa	No homology observed	
6	pzA109	4·16	83	25%	98 aa	No homology observed	
7	pzA105	2·8	84	25%	94 aa	No homology observed	
8	pzB1501	0·33	7593	99%	1501 aa	Snq2p protein	*S. cerevisiae*
9	pzB362	4·26	1643	99%	361 aa	60S ribosomal protein, Rpl2Bp	*S. cerevisiae*
10	pzB647	1·81	93	25%	94 aa	No homology observed	
11	pzC399	1·13	1313	62%	392 aa	Initiation factor 4A-3	*N. plumbaginifolia*
12	pzC109	3·43	70	27%	44 aa	No homology observed	
13	pzF240	2·53	1001	100%	240 aa	End of Hex2p protein	*S. cerevisiae*
14	pzF889	1·45	102	18%	117 aa	No homology observed	
15	pzF396	1·19	811	41%	362 aa	45·5 kDa protein in ATP3-RPS18b intergenic region	*S. cerevisiae*
16	pzF1050	1·24	125	21%	174 aa	No homology observed	
17	pzF129	3·13	95	29%	86 aa	No homology observed	
18	pzD196	3·7	85	27%	82 aa	No homology observed	
19	pzD400	1·45	1960	99%	400 aa	Glycine cleavage T protein, Gcv1p	*S. cerevisiae*
20	pzE570	1·47	650	33%	389 aa	DNA binding protein, Reb1p,	*S. cerevisiae*
21	pzE232	2·74	169	25%	171 aa	[uridine kinase, Urk1p]	*S. cerevisiae*
22	pzE110	4·63	84	26%	73 aa	No homology observed	
23i	pzA156i	4·54	761	100%	156 aa	40S ribosomal protein, Rps18Ap	*S. cerevisiae*

Table 2. Comparison of present and previously published DNA sequences of highly homologous genes. Numbering in square brackets is based on complete cDNAs and their deduced amino acid sequences.

ORF	Gene	Nucleotide difference 31a5/published	Amino acid difference 31a5/published
pzA462	*SES1*	[671] T/C	[224] Leu/Pro
pzB1501	*SNQ2*	[233] T/A	[78] Val/Glu
pzD400	*GCV1*	[366] C/G	[122] Asp/Glu

high values (4·26 and 4·54), which normally might suggest non-functional genes; both these ORFs encoded ribosomal proteins. ORFs pzB362 and pzA156i coded for the *S. cerevisiae* ribosomal proteins Rp12Bp (Presutti *et al.*, 1988) and Rps18Ap (Folley and Fox, 1994), respectively. These genes have both been described previously in two versions. The latter was interrupted by an intron as described by Folley and Fox (1994). The ORF pzA156i on chromosome IV is identical to the one designated *RPS18A*, whereas *RPS18B* is located on yeast chromosome II. *RPS18A* differed from *RPS18B* by 21 silent mutations and by having an intron of 339 bp instead of 511 bp. The designation of the A- and B-version of gene *RPL2* varies among sequences submitted to data libraries. We suggest that the designation proposed by Presutti *et al.* (1988) should be used. The ORF pzB362 sequence was highly homologous to the partial sequence of *RPL2B* (Lucioli *et al.*, 1988). The *RPL2* gene on yeast chromosome II is then *RPL2A* (Smits *et al.*, 1994). The *RPL2B* gene differs from the *RPL2A* gene by 11 silent mutations and by one (G1066A) which results in an amino acid difference (Ala356Thr).

A fifth ORF, pzF240, at the end of the cosmid insert corresponded to one-quarter of the *S. cerevisiae HEX2/REG1/SRN1* (Niederacher and Entian, 1991), whose product is a negative regulatory element in glucose repression. *HEX2* mutants are deficient in glucose repression, and defective in RNA processing (*SRN1*), two very different phenotypes given by the same gene.

Nucleotides 25690–25763 were identical to those coding for the tRNA-Val2a previously described on chromosome II (André *et al.*, AC Z35914), carrying the anticodon sequence TAC. It is interesting to note that this valine codon, apparently rarely used in yeast (Sharp and Li, 1987), has at least two identical genes.

Four ORFs had deduced amino acid sequences which showed similarities to other genes that were clearly significant. The ORF pzA520 amino acid sequence had 74·3% identity over 517 amino acids with *S. cerevisiae* galactokinase, Gal1p (Smits *et al.*, 1994) and may encode another galactokinase. The ORF pzC399 had 62·0% identity over 392 amino acids with the eukaryotic initiation factor 4A-3 in *Nicotiana plumbaginifolia* (Owttrim *et al.*, 1991) and showed high similarity to several other proteins in this eukaryotic initiation factor family. The codon preference value for this ORF is 1.13. The ORF pzE570 amino acid sequence was 39·7% identical over its full length with a DNA binding protein Reb1p from *Kluyveromyces lactis* (Morrow *et al.*, 1993) and 33·4% identical over 389 amino acids with the corresponding protein from *S. cerevisiae* (Ju *et al.*, 1990) located on chromosome II. These proteins have two areas with high similarity to Myb, but there is not much Myb similarity in pzE570. The ORF pzF396 amino acid sequence had 41·7% identity over 362 amino acids with the *S. cerevisiae* putative 45·5 kDa protein in the *ATP3-RPS18b* intergenic region (André *et al.*, AC P38226). The closest similarity to any protein of known function for this ORF was with 1-acylglycerol-3-phosphate acyltransferase from *Zea mays* (Brown *et al.*, 1994), with an identity of 36% over 76 amino acids residues and an optimized FASTA score of 173. The codon preference value was 1.19. The sequence revealed a hydrophobic segment corresponding to residues 27–54, which may indicate a membrane-spanning domain.

Two ORFs with similarities of borderline significance were observed. In spite of relatively low sequence similarity between the deduced amino acid sequence of ORF pzE232 and the *S. cerevisiae* uridine kinase, Urk1p (Kern, 1990), the pattern conserved between mouse-, *E. coli*- and yeast-uridine kinases is conserved also in the ORF sequence (Figure 1). Phosphoribulokinases appear to have the same conserved functional pattern as the uridine kinases (Figure 1). These observations may define an enzymatic family, probably of similar enzymatic mechanism and 3D structure, of which pzE232 most likely is a member, although the identity to *Synechocystis sp.* phosphoribulokinase (Su and Bogorad, 1991) was only 16% over 186 amino acids. In the other borderline case, an interesting homology, suggesting an undescribed protein family, was observed. The deduced sequence from the *S. cerevisiae* ORF pzA208 showed similar homology to two 'protein products', of different origin, with unknown function. An identity of 27% over 190 amino acid residues is shared with *Homo sapiens* 'KIAA0186 protein' (Nagase *et al.*, AC D80008) and an identity of 26% over 208 amino acid residues is shared with *Caenorhabditis elegans* 'R53.6 product' (Wilkinson, AC Z66515). In addition, a fragment of this protein was sequenced by the Institute for Genomic Research as an Expressed Sequence Tag (EST) from *S. cerevisiae* (Weinstock, AC T38583), showing that pzA208 is indeed expressed. The EST shared an identity of 98% over 51 amino acid residues with the deduced sequence of ORF pzA208. The alignment of the four sequences showed strong conserved pattern homology (Figure 2). The relevance of these borderline cases remains to be confirmed experimentally.

Of the remaining ORFs, pzA114, pzA109, pzA105 and pzC109 overlapped with ORFs on the opposite strand and were very short (105–114 amino acids). None of these four ORFs showed significant similarities with any protein in the searched databases and their codon preference values were generally high. They are most likely not functional genes. Seven ORFs remain for further analysis in the future.

The part of chromosome IV reported here does not deviate significantly from previously described cosmids (Dujon *et al.*, 1994). It most likely contains 18 genes (not counting the *HEX2* fragment), i.e. one gene per 1·9 kb. ORF lengths vary between 110 and 1501 amino acids, one gene out of 18 has an intron. The base composition of the cosmid was A 30·5%, C 19·2%. G 18·5% and T 31·8%. Our analysis reveals two ribosomal protein genes on chromosome IV which are close homologues of genes on chromosome II, suggesting a fairly recent duplication. There was no evidence of sequence

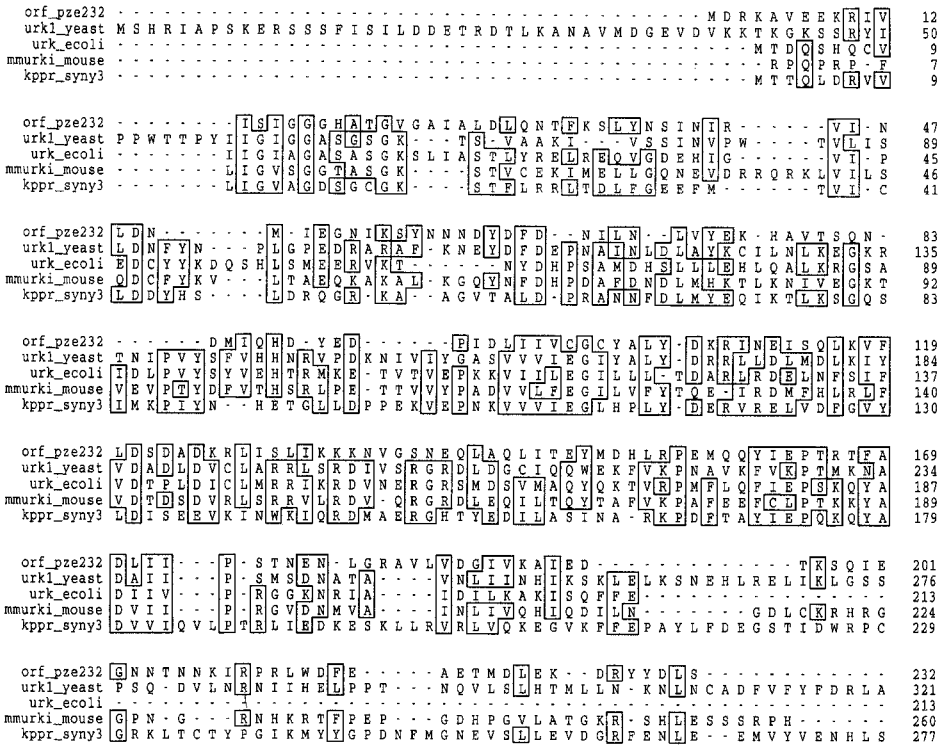


Figure 1. Alignment of the deduced amino acid sequence of pzE232 with uridine kinases from yeast (urk1), *E. coli* (urk) and mouse (mmurki), and the *Synechocystis sp.* phosphoribulokinase (kppr_syny3).

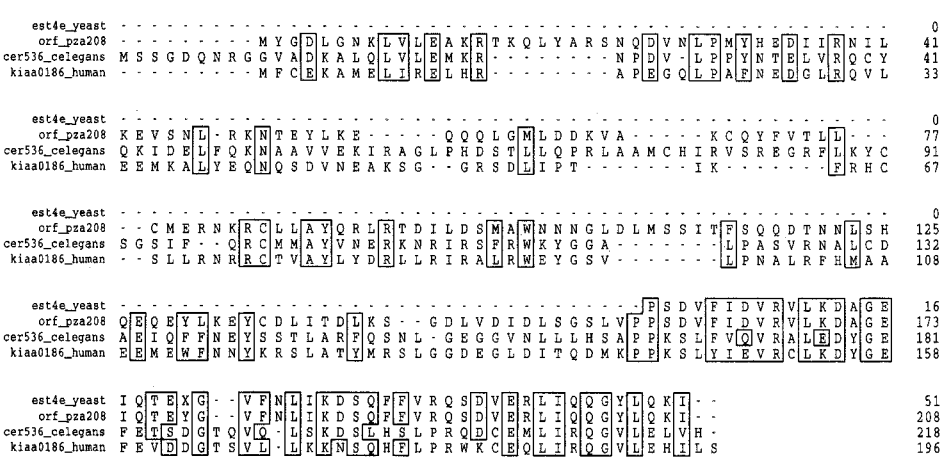


Figure 2. Alignment of the deduced amino acid sequence of pzA208 with a *C. elegans* protein product (cer53.6), a human protein (kiaa 0186), and a yeast EST (est4e).

homology beyond the limits of the genes, neither when the flanking sequences of *RPL2A* and *B* were compared, nor when *RPS18A* and *B* were compared.

ACKNOWLEDGEMENT

The assistance of Dr Marianne Wright, Jorun Solheim and Janne Røe in establishing some

subclones is gratefully acknowledged. This work was supported by grants from the European Union (contract no. B102-CT94-2071) and the Research Council of Norway to H.P.

REFERENCES

André, B., Cziepluch, C., Hein, C., Jauniaux, J. C., Urrestarazu, A. and Visser, S. *EMBL AC Z35914.*

André, B., Cziepluch, C., Hein, C., Jauniaux, J. C., Urrestarazu, A. and Vissers, S. *SwissProt AC P38226.*

Brown, A. P., Coleman, J., Tommey, A. M., Watson, M. D. and Slabas, A. R. (1994). Isolation and characterization of a maize cDNA that complements a 1-acyl sn-glycerol-3-phosphate acyltransferase mutant of *E. coli* and encodes a protein which has similarities to other acyltransferases. *Plant. Mol. Biol.* **26,** 211–223.

Casari, G., Andrade, M., Bork, P., *et al.* (1995). Challenging times for bioinformatics. *Nature* **376,** 647–648.

Dujon, B., Alexandraki, D., Andre, B. *et al.* (1994). Complete DNA sequence of yeast chromosome XI. *Nature* **369,** 371–378.

Folley, L. S. and Fox, T. D. (1994). Reduced dosage of genes encoding ribosomal protein S18 suppresses a mitochondrial initiation codon mutation in *Saccharomyces cerevisiae. Genetics* **137,** 369–379.

Ju, Q., Morrow, B. E. and Warner, J. R. (1990). REB1, a yeast DNA-binding protein with many targets, is essential for growth and bears some resemblance to the oncogene *myb. Mol. Cell. Biol.* **10,** 5226–5234.

Kern, L. (1990). The *URK1* gene of *Saccharomyces cerevisiae* encoding uridine kinase. *Nucl. Acids Res.* **18,** 5279–5279.

Lucioli, A., Presutti, C., Ciafre, S., Caffarelli, E., Fragapane, P. and Bossoni, I. (1988). Gene dosage alteration of L2 ribosomal protein genes in *Saccharomyces cerevisiae*: effects on ribosome synthesis. *Mol. Cell. Biol.* **8,** 4792–4798.

McNeil, J. B., Zhang, F. R., Taylor, B. V., Pearlman, R. E. and Bognar, A. L. *EMBL AC L41522.*

Morrow, B. E., Ju, Q. and Warner, J. R. (1993). A bipartite DNA-binding domain in yeast Reb1p. *Mol. Cell. Biol.* **13,** 1173–1182.

Nagase, T., Seki, N., Tanaka, A., Ishikawa, K. I. and Nomura, N. Prediction of the coding sequences of unidentified human genes. *Translated EMBL DataBank AC D80008.*

Niederacher, D. and Entian, K.-D. (1991). Characterization of Hex2 protein, a negative regulatory element necessary for glucose repression in yeast. *Eur. J. Biochem.* **200,** 311–319.

Owttrim, G. W., Hofmann, S. and Kuhlemeier, C. (1991). Divergent genes for translation initiation factor eIF-4A are coordinately expressed in tobacco. *Nucl. Acids Res.* **19,** 5491–5496.

Presutti, C., Lucioli, A. and Bozzoni, I. (1988). Ribosomal protein L2 in *Saccharomyces cerevisiae* is homologous to ribosomal protein L1 in *Xenopus laevis.* Isolation and characterization of the genes. *J. Biol. Chem.* **263,** 6188–6192.

Servos, J., Haase, E. and Brendel, M. (1993). Gene *SNQ2* of *Saccharomyces cerevisiae*, which confers resistance to 4-nitroquinoline-N-oxide and other chemicals, encodes a 169 kDa protein homologous to ATP-dependent permeases. *Mol. Gen. Genet.* **236,** 214–218.

Sharp, P. M. and Li, W.-W. (1987). The codon adaptation index—a measure of directional synonymous codon bias, and its potential applications. *Nucl. Acids Res.* **15,** 1281–1295.

Smits, P. H. M., De Haan, M., Maat, C. and Grivell, L. A. (1994). The complete sequence of a 33 kb fragment on the right arm of chromosome II from *Saccharomyces cerevisiae* reveals 16 open reading frames, including ten new open reading frames, five previously identified genes and a homologue of the *SCO1* gene. *Yeast* **10**(Suppl. A), S75–S80.

Su, X. and Bogorad, L. (1991). A residue substitution in phosphoribulokinase of Synechocystis PCC 6803 renders the mutant light-sensitive. *J. Biol. Chem.* **266,** 23698–23705.

Voss, H., Wiemann, S., Wirkner, U., *et al.* (1992). Automated DNA sequencing system resolving 1,000 bases with fluorescein-15*dATP as internal label. *Meth. Mol. Cell. Biol.* **3,** 153–155.

Weinstock, K. *GenBank AC T38583.*

Weygand-Durasevic, I., Johnson-Burke, D. and Soell, D. (1987). Cloning and characterization of the gene coding for cytoplasmic seryl-tRNA synthetase from *Saccharomyces cerevisiae. Nucl. Acids Res.* **15,** 1887–1904.

Wiemann, S., Rupp, T., Zimmerman, J., Voss, H., Schwager, C. and Ansorge, W. (1995). Primer design for automated DNA sequencing utilizing T7 DNA polymerase and internal labeling with fluorescein-15-dATP. *BioTechniques* **18,** 688–697.

Wilkinson, J. *Translated EMBL DataBank AC Z66515.*

Zimmerman, J., Voss, H., Schwager, C., *et al.* (1990). A simplified protocol for fast plasmid DNA sequencing. *Nucl. Acids Res.* **18,** 1067.