YEAST VOL. 13: 655–672 (1997)

•°°°° xv °° °°°°° Yeast Sequencing Reports

DNA Sequencing and Analysis of 130 kb from Yeast Chromosome XV

HARTMUT VOSS¹*, VLADIMIR BENES¹, MIGUEL A. ANDRADE^{2,3}, ALFONSO VALENCIA³, STEFANIE RECHMANN¹, CRISTINA TEODORU¹, CHRISTIAN SCHWAGER¹, VACLAV PACES⁴, CHRIS SANDER² AND WILHELM ANSORGE¹

¹Biochemical Instrumentation Programme, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany ²Biological Structures and Biocomputing Programme, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany

³Protein Design Group, CNB-CSIC, Cantoblanco, Madrid 28049, Spain

⁴Institute of Molecular Genetics, Czech Academy of Sciences, Prague, Czech Republic

Received 11 April 1996; accepted 10 October 1996

We have determined the nucleotide sequence of 129 524 bases of yeast (*Saccharomyces cerevisiae*) chromosome XV. Sequence analysis revealed the presence of 59 non-overlapping open reading frames (ORFs) of length >300 bp, three tRNA genes, four delta elements and one Ty-element. Among the 21 previously known yeast genes (36% of all ORFs in this fragment) were nucleoporin (NUP1), ras protein (RAS1), RNA polymerase III (RPC1) and elongation factor 2 (EF2). Further, 31 ORFs (53% of the total) were found to be homologous to known protein or DNA sequences, or sequence patterns. For seven ORFs (11% of the total) no homology was found. Among the most interesting protein identifications in this DNA fragment are an inositol polyphosphatase, the second gene of this type found in yeast (homologous to the human OCRL gene involved in Lowe's syndrome), a new ADP ribosylation factor of the arf6 subfamily, the first protein containing three C2 domains, and an ORF similar to a *Bacillus subtilis* cell-cycle related protein. For each ORF detailed sequence analysis was carried out, with a full consideration of its biological function and pointing out key regions of interest for further functional analysis. The sequence has been submitted to the EMBL data library under Accession Number X94335. © 1997 by John Wiley & Sons, Ltd.

Yeast 13: 655–672, 1997. No. of Figures: 12. No. of Tables: 1. No. of References: 70.

KEY WORDS - genome sequencing; yeast-human homolog; genequiz

INTRODUCTION

Chromosome XV with an estimated size of 1.108 megabases is the third largest chromosome of the budding yeast *Saccharomyces cerevisiae*. In the frame of the European Union yeast genome

*Correspondence to: Harmut Voss. Contract grant sponsor: European Union yeast genome sequencing programme. Contract grant sponsor: CICYT, Spain.

CCC 0749–503X/97/070655–18 \$17.50 © 1997 by John Wiley & Sons Ltd project we have sequenced and analysed a cluster of nine overlapping cosmids covering the central region of the chromosome. The sequence of 129 524 bases has been submitted to the EMBL data library under accession number X94335. Here we discuss the structural features of this chromosomal region, base composition, density, distribution and orientation of genes. The detailed analysis carried out contributes to the current knowledge of the yeast genome in several aspects since it shows: (i) sequences highly homologous to other yeast sequences, indicating genome duplication; (ii) new yeast sequences in already known protein families, suggesting new connections within the family and new perspectives for the function of the family; (iii) first occurrences in yeast of sequences in already known families, showing new biological or evolutionary aspects; (iv) new yeast sequences defining a new protein family and establishing cross-relations between species. Biological information is gathered by deciphering the conserved regions in protein families, and information about protein evolution is gained from phylogenetic interpretations.

MATERIALS AND METHODS

Cosmids

A cluster of nine ordered overlapping cosmids (pEOA347, pUOA522, pEOA246, pEOA273, pEOA306, pEOA265, pEOA106, pEOA986 and pEOA1081) covering the chromosomal region from 485 000 to 615 000 of chromosome XV was obtained in the EU project from the chromosome co-ordinator B. Dujon, Institut Pasteur, Paris. The cosmids were isolated and mapped as described by Thierry *et al.* (1995).

Subcloning strategy

Escherichia coli strain XL1-Blue[®] (Stratagene) was used for all subcloning steps. In general, all *Eco*RI fragments of cosmid inserts were cloned into plasmid vector pUC18. One *Eco*RI fragment of about 1.7 kb (position 17 875–19 580) was not clonable into plasmids, therefore templates were prepared as biotinylated polymerase chain reaction (PCR) products for solid phase sequencing on magnetic beads (Hultman *et al.*, 1992) using neighbouring sequence information to design PCR primers. Another region from position 1140 to 6339 turned out to be unclonable in high copy number plasmids, but could be successfully cloned into low copy number plasmid pBR322.

DNA sequencing

The entire sequence of 129 524 bp was determined on both strands mainly by directed primer walking strategy and T7 DNA polymerase with unlabelled primers and fluorescein-15*dATP as internal label as described previously (Voss *et al.*, 1992). Sequences were analysed on two commercial ALF DNA sequencers (Pharmacia, Uppsala). After sequencing of plasmid subclones, linking of adjacent *Eco*RI fragments was performed by direct cycle sequencing on cosmid DNA (Zimmermann *et al.*, 1994). Raw data collection and evaluation were performed using the ALF manager software; sequence assembly, data evaluation and presentation were performed with the EMBL GeneSkipper sequence analysis software (Schwager *et al.*, 1995).

Definition of open reading frames

All open reading frames (ORFs) larger than 300 bp were translated using the standard genetic code, and independent database searches were carried out for each one. Names of the ORFs correspond to the general notation rule: YORnW stands for the Watson strand and YORnC for the Crick strand.

Data analysis

The database searches for homologous sequences have been carried out using 'GENEQUIZ', a project management, browsing and visualization tool developed by the EMBL protein design group (Scharf et al., 1994). The following databases were searched: protein sequence: PDB (Abola et al., 1987), SwissProt (Bairoch and Apweiler, 1996), PIR-NBRF (fraction not overlapping with Swiss-Prot; George et al., 1996), GENPEPT (a direct translation of the DNA sequences in GenBank; Benson et al., 1993), TREMBL (Bairoch and 1996); DNA Apweiler, sequence: EMBL (Rodriguez-Tome et al., 1996), GenBank (Benson et al., 1996), expressed sequence tags (ESTs) in dbEST (Boguski, 1995). Updated versions of the databases from 10 January 1996 were used. A continuous update of the results using the latest database versions is available through world wide web at http://gredos. cnb. uam. es/yeast130. html. Prior to the database scanning, sequences were masked using an algorithm to avoid spurious hits in regions of obvious composition bias (G. Casari et al., unpublished).

The scan of the database was done using the BLAST (Altschul *et al.*, 1990) and FASTA (Pearson and Lipman, 1988) programs (parameters: BLOSUM62 matrix for BLAST; and Ktup=2 for FASTA). Multiple-sequence alignments were obtained using the programs MAXHOM (Sander and Schneider, 1991), CLUSTALW (Higgins *et al.*, 1992) or PILEUP (GCG package). Protein secondary structure was predicted from multiple sequence alignments using the PHD neural network method (Rost and Sander, 1994), as

implemented on the PredictProtein network server (Rost *et al.*, 1994). Phylogenetic trees based on the neighbour-joining method (Saitou and Nei, 1987) were calculated using the CLUSTALW package (Higgins *et al.*, 1992). Corrections for multiple replacements were applied (Kimura, 1983). The stability of trees with respect to different choices of subsets of residue positions was checked by bootstrapping experiments (Felsenstein, 1985). Profile searches were made using PROFILESEARCH (GCG) or MAXHOM (Sander and Schneider, 1991).

RESULTS AND DISCUSSION

DNA analysis

We report here sequencing and analysis of 129 524 bases of yeast (*S. cerevisiae*) chromosome XV (accession no. X94335). A schematic presentation of the distribution of 59 ORFs (plus one case of two overlapping ORFs of significant length), three tRNA genes, four delta elements, seven perfect ARS consensus sequences and one Ty-element is shown in Figure 1.

The average GC-content of the sequenced part is 38.5%, very similar to the GC-content of other known yeast chromosomes. A plot of the GC-content calculated over 10 kb windows every 100 bp shows two minima around positions 20 000 and 120 000 (data not shown). Whether this finding reflects any periodicity in GC-content over the whole chromosome as described for chromosomes II and XI (Feldmann *et al.*, 1994; Dujon *et al.*, 1994) will be confirmed when the complete chromosome sequence becomes available. Three of the four delta elements flank the Ty-element, two of the three tRNA genes are found in the proximity of delta elements, a phenomenon frequently observed in yeast.

Among the seven perfect ARS consensus sequences, the elements at positions 6679 and 6704 are the most probable active elements according to the observations from yeast chromosome VI (Murakami *et al.*, 1995). The density of coding regions in this chromosomal segment (one every $2 \cdot 2$ kb) is lower than that found on other known yeast chromosomes. On the other hand the average ORF size (550 codons) is larger than on all other chromosomes reported so far (457–503 codons), reflecting the fact that the sequence contains seven ORFs larger than 1000 codons. An unusual clustering on the Watson strand is observed over ten ORFs within a stretch of 20 kb in the region from position 41 165 to 61 975.

In yeast, clustering of ORFs on one strand seems to occur in general more frequently than statistically expected, which raises the question whether it reflects polycistronic transcription, as recently observed in *Caenorhabditis elegans* (Spieth et al., 1993) or whether it reflects a preferred arrangement to prevent collisions between the transcription and replication complexes (Brewer, 1988). Even more interestingly, the preferred number of ORFs in a cluster is in general five to seven; if a cluster contains more than five to seven ORFs on one strand, it is frequently interrupted in the middle by a delta element to form two units of clustered ORFs (chromosome I: position 180 000 to 194 000; chromosome III: position 154 000 to 174 000; chromosome VIII: position 81 000 to 99 000 and 451 000 to 473 000). Besides the ten uninterrupted ORFs found here in this fragment from chromosome XV, a comparable cluster has been found so far only in chromosome II in the region between position 345 000 and 375 000 (Feldmann et al., 1994).

Analysis of ORFs

The data analysis involved two steps: exhaustive search in databases and in-depth protein family analysis. In the first step, database scanning was performed using GENEQUIZ, a tool for the analysis of massive sequence data (Scharf et al., 1994). GENEQUIZ uses daily updates of different databases, an integrated database search system, a rule-based engine for interpreting the results of homology searches, and an advanced humanmachine interface (Casari et al., 1995). The fraction of ORFs for which it was possible to assign a function in this fragment is relatively large (59%), larger than that for any other yeast chromosome. This is partly due to significantly improved searching strategies, as has been demonstrated in other cases (Casari et al., 1995; Ouzounis et al., 1996), but also due to the rapidly growing information in databases. The recent rapid increase in the number of database entries lacking primary functional annotation leads to an increasing number of cases where a sequence family emerges, yet no functional characterization is possible (corresponding to class (iv) described in the analysis below). A similarity search between the ORFs identified in this project and those in the public databases is summarized in Table 1.

NameFonToaIdentityProtein/DNADescriptionaSoreFaturesYOB31064652142										
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	Name	From	To	аа	Identity	Protein/DNA	Description	аа	Score	Features of the ORFs
Non-state Sindar Mindörd LinASE-SER Son to the state 24 1-16-1 LInASE-SER Son the state 256-7 LinASE-7 Three CJ dom ABS-cons 1391 1391 1392 432 Similar RYN-JY 100 3-56-7 Three CJ dom VOR3157c 1481 1321 1332 10 dentical RYN-JY 100 12-55 Transmuth-transferrace 12-56-7 Transmuth-transferrace 12-56-7 Transmuth-transferrace 12-56-7 Transmuth-transferrace 12-56-7 12-56-7 ADP-holyalia 12-57-56 ADP-holyalia <td< td=""><td>YOR2964c YOR3116w</td><td>465 3059</td><td>2714 3946</td><td>749 295</td><td>Similar Similar</td><td>YK69_YEAST YK71_YEAST</td><td>Hypothetical protein Hypothetical protein</td><td>910 152</td><td>2·7e-247 1·4e-4</td><td></td></td<>	YOR2964c YOR3116w	465 3059	2714 3946	749 295	Similar Similar	YK69_YEAST YK71_YEAST	Hypothetical protein Hypothetical protein	910 152	2·7e-247 1·4e-4	
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	YOR3120w YOR3124w	4113 5559 6270	5276 6611 6220	387 350	Similar Identical	Mm0361_1 OSTG_YEAST	Lipase-esterase operon product Oligosaccharyl transferase γ precursor	264	1.1e-1	LIPASE_SER PROSITE
ORX1416 Gr35 Diss Similar SYT1_CAEL Rynongamin I 441 35-7 Three C3 dom RNS.com 1066 1033 210 1031 1035 1035 1035 1035 1035 1035 1035 23e-7 Three C3 dom YOR310A 17314 1731 1031 1035 210 Horizabi 1745 23e-7 Tansmemb+4 YOR310A 1730 1705 1433 210 Kenting Portini Portaportini Portini Portini Portini Portini Portini Portini Portini	ARS-cons ARS-cons	6704 6704	0089 6714							
ARS-ons 3043 1323 43 Similar TRP_DROME Transfer treeptor potential protein 1275 2:3e-7 Transfer treeptor potential protein ARS-ons 1943 1353 21 dentical YF:1_YEAT Gtp-Binding Protein YT 31 70e-19 Protein phosp YOR3165 1644 1724 35 similar dentical YF:1_YEAT Gtp-Binding Protein YT 31 32-57 Lentin-Linbola Differin Drop YOR3165 1644 1724 35 similar dentical Nt Differin Drop 36 signature YOR3172 2618 Similar RPL_ECULI Rth-ibosylation factor 6 175 7-2-66 ADP-ibosylation YOR3174 2075 3513 329 Similar Rth_ibosonal 57 35 5 35 35 35 35 35 35 35 35 35 35 35 35 35 36 36 36 36 36 36 36 <td>YOR3141c tRNA</td> <td>6745 10965</td> <td>10305 11038</td> <td>1186</td> <td>Similar</td> <td>SYT1_CAEEL</td> <td>Synaptotagmin I tRNA-Asn</td> <td>441</td> <td>3·5e-7</td> <td>Three C2 domains</td>	YOR3141c tRNA	6745 10965	10305 11038	1186	Similar	SYT1_CAEEL	Synaptotagmin I tRNA-Asn	441	3·5e-7	Three C2 domains
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	YOR3151w ARS-cons	11811	13259	482	Similar	TRP_DROME	Transient receptor potential protein	1275	2·3e-7	Transmemb+coiled-coil
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $	YOR3157c	13721 13721 14648	14353	210 572	Identical Similar	YP51_YEAST PDP_BOVIN	Gtp-Binding Protein YPT_51 Pyr DH (lipoamide)-phosphatase mecursor	538	7.0e-19	Protein phosphatase 2C signature
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	YOR3160w VOR3162c	16789 16944	17994	401 376	No homologue Similar	dhest-anl_73646	4 thaliana asne moduct		n7.5e-7	contains ORF YOR3162c
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	YOR3165w VOB 3170c	18651	20114 20114	487 1648	Similar Similar	Sctrnaorf_2	S. cerevisiae ORF	642	nl·7e-177	ATPase α-β
YOR3174 2705 2785 Similar RPIA_ECOLI Ribosophate isomerase A 219 2-9e-12 Riposomal S7 YOR3177w 29317 3020 190 Similar Vorsilar Similar to ribosomal S7 190 6-9e-113 Ribosomal S7 YOR318Ac 3101 1076 Identical NUP1_YEAST Nucleoporin NUP1 190 6-9e-113 Ribosomal S7 YOR318Ac 3103 3103 107 Identical NUP1_YEAST Nucleoporin NUP1 307 4-4e-13 Ribosomal S7 YOR318A 3563 3663 393 Identical KTR1_YEAST Nucleoporin NUP1 307 4-4e-13 Ribosomal S7 YOR318A 3653 3663 393 Identical KTR1_YEAST Probable mamosyltransferase Ktrl 307 4-4e-13 Ripton YOR3205A 3663 3663 100 GSTE_YEAST Mitochondrial carrier protein Ymc1 307 4-4e-13 Ripton YOR321A 4105 301 Identical OSTE_YEAST Vacuolar prot	YOR3172w	26318	26869	183	Similar	ARF6_CHICK	ADP-ribosylation factor 6	175	7·2e-66	ADP-ribosylation factors
YOR 3180: 3050 3102 T/S No homologue NUPL YEAST Nucleoporin NUP1 ARS-cons 3471 3470 1076 Identical NUP1_YEAST Nucleoporin NUP1 ARS-cons 3432 3539 3534 3529 393 Identical NUP1_YEAST Probable mamosyltransferase Ktrl ARS-cons 3662 36642 373 3612 3701 $44e-13$ Mitochondrial ARS-cons 3663 36642 3990 3091 Identical $KTR1_YEAST$ Probable mamosyltransferase Ktrl 307 $44e-13$ Mitochondrial VOR3205w 38767 39906 390 Identical $RAS1_YEAST$ $Ras-like protein 1$ 307 $44e-13$ $Mitochondrial VOR321c 39972 4106 43013 250 3500 3500 3500 3500 3600 307 44e-3 94e-3 94e-3 94e-3 94e-3 94e-3 94e-3 9400 9400 $	YOR3174c YOR3177w	27075 29317	27851 30290	258 190	Similar Similar	RPIA_ECOLI Sctrnaorf_1	Ribose 5-phosphate isomerase A Similar to ribosomal S7	219 190	2·9e-12 6·9e-113	Ribosomal S7e blocks+one
WTM 2000 355.9 353 Identical KTR1_YEAST Probable mannosyltransferase Ktr1 ARS-cons 36632 36642 300 Identical YMC1_YEAST Mitochondrial carrier protein Ymc1 307 4-4e-13 Mitochondrial signature YOR 3105c 36818 37801 327 Similar YMC1_YEAST Mitochondrial carrier protein Ymc1 307 4-4e-13 Signature YOR 3205w 38767 39696 300 Identical RAS1_YEAST Mitochondrial carrier protein Ymc1 307 4-4e-13 Signature YOR 3204w 43095 309 Identical OSTE_YEAST Oigosaccharyltransferase 16 kDa subunit Signature YOR 3204w 4306 4300 Similar VKO7_CAEEL Hypothetical protein 221 94e-2 Aipm_Homoc YOR 3204w 4550 435 Similar YKO7_CAEEL Hypothetical protein 221 94e-2 Aipm_Homoc YOR 3214w 4550 435 Similar YKO7_CAEEL Hypothetical protein 221 94e-2 <td< td=""><td>YOR3180c YOR3182c</td><td>30501 31471 31471</td><td>31028 34701 34036</td><td>$\begin{array}{c} 175\\1076\end{array}$</td><td>No homologue Identical</td><td>NUP1_YEAST</td><td>Nucleoporin NUP1</td><td></td><td></td><td></td></td<>	YOR3180c YOR3182c	30501 31471 31471	31028 34701 34036	$\begin{array}{c} 175\\1076\end{array}$	No homologue Identical	NUP1_YEAST	Nucleoporin NUP1			
ANS-0015 30032 30042 Similar YMC1_YEAST Mitochondrial carrier protein Ymc1 307 4-4e-13 Mitochondrial signature YOR33105 36818 37801 327 Similar YMC1_YEAST Mitochondrial carrier protein I 307 4-4e-13 Mitochondrial signature YOR32105 3767 3996 309 Identical OSTE_YEAST Oligosaccharyltransferase 16 kDa subunit signature ARS-cons 43976 393 Similar OSTE_YEAST Oligosaccharyltransferase 16 kDa subunit 4-4e-9 PROSITE of YOR3214w 41856 4395 283 Similar VEVOT_CAEEL Hypothetical protein 221 9-4e-2 Ate-9 PROSITE of YOR3224w 48706 4395 309 Similar YKOT_CAEEL Hypothetical protein 221 9-4e-2 Ate-9 PROSITE of YOR3234w 5210 4355 Similar YKOT_CAEEL Hypothetical protein 221 9-4e-2 Ate-9 PROSITE of YOR3234w 5305 5448	YOR3189w	35348	36529	393	Identical	KTR1_YEAST	Probable mannosyltransferase Ktrl			
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	Arcs-colls YOR3193c	36818	37801	327	Similar	YMC1_YEAST	Mitochondrial carrier protein Ymc1	307	4·4e-13	Mitochondrial energy carrier siona ture
Arxs-cous 40000 4000 4000	YOR3205w YOR3211c	38767 39972 10006	39696 40373	309 130	Identical Identical	RASI_YEAST OSTE_YEAST	Ras-like protein 1 Oligosaccharyltransferase 16 kDa subunit			
TOR3224w 45876 303 Similar FTL2 Vacuoiat proteases sorting 200 44-9 FKO311E 010 YOR3224w 44876 45805 309 Similar YKO7_CAEEL Hypothetical protein 221 9:4e-2 FKO311E 010 YOR3224w 44876 45805 309 Similar YKO7_CAEEL Hypothetical protein 221 9:4e-2 FKO311E 010 YOR3231w 4550 435 No homologue LEU1_YEAST 2-lsopropylmalate synthase 619 0:0 Apm_Homoc YOR3234w 53950 54648 232 Similar RSD1_YEAST Recessive suppressor of secretory defect 623 1:le-50 YOR3240w 55029 57314 761 Similar Cev07a12_5 C. elegans product 1183 3:7e-18 AA transfer cl YOR3240w 57366 60340 914 Identical Z26253 S. cerevisiae AZFI gene for zinc finger NA transfer cl YOR3248w 61091 61975 294 No homologue T38532 S. cerevis	YOR3214w	41165	42013	282	No homologue			000		
YOR3227w 46550 48238 562 Similar LEU1_YEAST 2-Isopropylmalate synthase 619 0.0 Aipm_Homoc YOR3231w 48799 53122 1107 Similar RSD1_YEAST Recessive suppressor of secretory defect 623 1·1e-50 patterns YOR3234w 53769 435 No homologue MAF_BACSU Hypothetical protein 189 7·0e-11 YOR3234w 53505 54648 232 Similar MAF_BACSU Hypothetical protein 189 7·0e-11 YOR324w 53050 57314 761 Similar Cew07a12_5 C. elegans product 1183 3·7e-18 AA transfer cl YOR324w 5796 60340 914 Identical Z26253 S. cerevisiae AZFI gene for zinc finger 3·7e-18 AA transfer cl YOR3248w 61091 61975 294 No homologue YOR322 S. cerevisiae EST n1·3e-66 638 1·3e-68 YOR YOR 374s 6.384 1/3e-61 RO1 YOR YOR YA n	YOR3224w	42044 44876	45805	309 309	Similar	YK07_CAEEL	v acuolar proteases sorting Hypothetical protein	288 221	4:4e-9 9.4e-2	PROSULE of epimorphines
YOR 3231w 48799 52122 1107 Similar RSD1_YEAST Recessive suppressor of secretory defect 623 1·1e-50 YOR 3234w 53769 435 No homologue MAF_BACSU Hypothetical protein 189 7·0e-11 YOR 3237w 53950 54648 232 Similar MAF_BACSU Hypothetical protein 189 7·0e-11 YOR 3240w 55029 57314 761 Similar Cew07a12_5 C. elegans product 1183 3·7e-18 AA transfer cl YOR 3248w 51091 61975 294 No homologue Z26253 S. cerevisiae AZFI gene for zinc finger 1183 3·7e-18 AA transfer cl YOR 3248w 61091 61975 294 No homologue Z26253 S. cerevisiae AZFI gene for zinc finger n1·3e-48 YOR 3254c 62186 6298 268 Similar T38532 S. cerevisiae EST n1·3e-48 YOR 3754c 63784 67666 1460 Identical RPC1 YA Andinected RNA nolymerase III	YOR3227w	46550	48238	562	Similar	LEU1_YEAST	2-Isopropylmalate synthase	619	0.0	Aipm_Homocit_Synth 1 & 2 patterns
YOR 3237w 53950 54648 232 Similar MAF_BACSU Hypothetical protein 189 7-0e-11 YOR 3237w 55309 57314 761 Similar Cew07a12_5 C. elegans product 1183 3-7e-18 AA transfer cl YOR 3244w 55506 60340 914 Identical Z26253 S. cerevisiae AZFI gene for zinc finger 1183 3-7e-18 AA transfer cl YOR 3244w 55506 60340 914 Identical Z26253 S. cerevisiae AZFI gene for zinc finger 1183 3-7e-18 AA transfer cl YOR 3244w 51091 61975 294 No homologue No totin No totin NO totin n1·3e-48 YOR 3251c 62180 62986 268 Similar T38532 S. cerevisiae EST n1·3e-48 YOR 3754c 63784 67666 1460 Identical RPC1 YEAST DNA driverted RNA nolymerase III	YOR 3231w VOB 3734w	48799 52462	52122 53760	1107	Similar No homologue	RSD1_YEAST	Recessive suppressor of secretory defect	623	1·1e-50	4
YOR3240w 55029 57314 761 Similar Cew07a12_5 <i>C. elegans</i> product 1183 3.7e-18 AA transfer cl YOR3244w 57596 60340 914 Identical Z26253 <i>S. cerevisiae AZF1</i> gene for zinc finger YOR3248w 61091 61975 294 No homologue YOR3251c 62180 62986 268 Similar T38532 <i>S. cerevisiae</i> EST n1.3e48 YOR3354 63784 67666 1460 Identical RPC1 YEAST DNA.directed <i>RNA</i> nolymerase III	YOR3237w	53950	54648	232	Similar	MAF_BACSU	Hypothetical protein	189	7·0e-11	
YOR3248w 61091 61975 294 No homologue YOR3251c 62180 62986 268 Similar T38532 S. cerevisiae EST YOR3754c 63784 67666 1460 Identical RPC1 YEAST DNA-directed RN4 nolymerase III	YOR3240w YOR3244w	55029 57596	57314 60340	761 914	Similar Identical	Cew07a12_5 Z26253	C. elegans product S. cerevisiae AZFI gene for zinc finger	1183	3·7e-18	AA transfer class PROSITE
YOR3251c 62180 62986 268 Similar T38532 S. cerevisiae EST YOR3354c 63384 67666 1460 Identical RPC1 YEAST DNA-directed RN4 nolymerase III	YOR3248w	61091	61975	294	No homologue					
YOR3258% 68550 69854 434 Identical TBPI YEAST Tar-binding homolog I	YOR3251c YOR3254c YOR3258w	62180 63284 68550	62986 67666 69854	268 1460 434	Similar Identical Identical	T38532 RPC1_YEAST TBP1_YEAST	S. cerevisiae EST DNA-directed RNA polymerase III Tat-binding homolog 1		n1·3e-48	

Table 1. Position of the protein and DNA features found in the sequence reported.

© 1997 by John Wiley & Sons, Ltd

YEAST VOL. 13: 655–672 (1997)

658

H. VOSS ET AL.

Name	From	To	аа	Identity	Protein/DNA	Description	аа	Score	Features of the ORFs
YOR 3263w YOR 3266c YOR 3269w YOR 3275c	70378 72313 74635 75819	72081 73767 75573 76408	567 484 312 126	No homologue Similar Identical Identical	YOT3_CAEEL GCY_YEAST PROF_YEAST	Hypothetical protein GCY protein of unknown function Profilin prevents the polymerization of	510	1·5e-32	One intron
YOR 3278c YOR 3281c YOR 3284c YOR 3287c YOR 3290w	76697 78345 82551 83482 84691 87007	78091 82163 83369 84198 87714 87713	464 1272 272 238 238 1007	Identical Identical Similar Similar Identical	LEO1_YEAST UBP2_YEAST U13642 M94674 SC07421 BUD6_VEAST	actin Ubikuown function Ubikuown function C. elegans gene product C. albicans a-glucosidase (maltase) mRNA (non-translated) S. cerevisiae S288C rho-type GTPase activating protein D Eihondania contranslated		n2-9e-26 n1-2e-13	
YOR3296c YOR3299c	90398 93450	93079 94328 94328	201 893 292	No homologue Similar	YMC1_YEAST	r-ruocojamnonnuazore carooxuase catalytic subunit Mitochondrial carrier protein YMC1	307	3·2e-12	Mitochondrial energy carrier signature
ARS-cons tRNA YOR3311c YOR3314w	95163 95479 95703 96696	95173 95550 96359 98351	218 551	Similar Identical	YHFE_ECOLI VP17_YEAST	tRNA-Asp Hypothetical protein Vacuolar protein sorting-associated	252	3.2e-8	
YOR 3317w YOR 3320w	98619 102085	101147 103314	842 409	Identical Similar	EF2_YEAST SAC7_YEAST	protein VPS1/ Elongation factor 2 SACT protein involved in	274	8·6e-18	Probable rho/racGAP domain
YOR 3326w YOR 3329c VOR 3337c	103771 105334 107830	104880 107202 109845	369 622 671	Identical Similar No homologue	IDH2_YEAST Scl8093_3	assemptyrumetrou of acture IDH Mitochondrial sub 2 precursor Yeast chromosome XII cosmid	578	3.1e-24	Leucine_Zipper PROSITE
YOR 3339w YOR 3339w	110502 110502 113693	112802 1112802 116107	766 804	Identical Similar	SFL1_YEAST ACT2_YEAST	Flocculation suppression protein Actin-like	391	3.0e-6	Actin proteins block,
YOR3352w tRNA ô-rennant s	116577 117875 118033 118033	117566 117945 118302 118677	329	Similar	SUCA_RAT	Succinyl CoA ligase tRNA-Gly-sup	333	2·8e-124	
YOR3367w §	118632 118632 123923	123900 124254	1755	Similar	Sc8229_23	Transposon peptide	1755	0.0	Frameshift (by homology)
YOR3373c YOR3510c YOR3513c	124903 126237 128867	125862 128612 129523	319 791 218	Identical Identical Similar	TH80_YEAST SC0612 D28195	Thiamin pyrophosphokinase EST Translated cDNA (rice)	I	n8•0e-29	Fragment
The protein/ throughout database of underscore a EMBL entry scores are lis	DNA colt the paper. protein tra and a nun /). The EM	Imn indic: SwissPro anslation I nber that ABL acces the closesi	ates the ot ident product indicat ssion nu t homo	closest homologuifiers are presente ifiers are presente i from the EMBL es the order of the mber is given for logue is a protein	ae in the database d with two word DNA database. 1 he translation pro • DNA, one or two sequence, BLAST	. Unless stated otherwise, only SwissProt an s in capital letters joined by an underscort The TREMBL identifiers are composed of t oduct (since many consecutive translation to capital letters followed by a number. The IX scores in case of a nucleotide sequence (nd TREN b, the sec he corres products score in indicated	1BL prot cond refe ponding are freq dicates the by an 'n	ein sequence identifiers are used rring to species. TREMBL is a EMBL identifier followed by an uently reported from the same he degree of homology, BLAST).

130 kb FROM CHROMOSOME XV

© 1997 by John Wiley & Sons, Ltd

Table 1. Continued

YEAST VOL. 13: 655-672 (1997)

659



© 1997 by John Wiley & Sons, Ltd

YEAST VOL. 13: 655-672 (1997)

660

Among the 59 ORFs identified here, 21 (36%) are identical to previously described genes, including nucleoporin (NUP1; Davis and Fink, 1990), ras protein (RAS1; Powers et al., 1984), 16-kDa subunit of oligosaccharyl transferase (OST2; Silberstein et al., 1995), RNA polymerase III (RPC1; Allison et al., 1985) and elongation factor 2 (EF2; Perentesis et al., 1992). The region from 74 600 to 82 000 has previously been found to contain the yeast genes GCY (Oechsner et al., 1988), PFY (Magdolen et al., 1988), LEOI (Magdolen et al., 1994) and UBP2 (Baker et al., 1992). Thirty-one ORFs (53%) were found to be homologous to known protein or DNA sequence or patterns. For seven ORFs (11%) no homologous pattern was found. Eight small internal ORFs and eight small partially overlapping ORFs were excluded; these were in the size range from 100 to 150 codons and did not show any homology to database entries. However, we included the case of two overlapping ORFs (YOR3160w, 401 codons, ATG at position 16 789; YOR3162c, 326 codons, ATG at position 17 924) in our study. Both ORFs have significant length, the shorter ORF, YOR3162c, shows high homology to an Arabiclopsis thaliana EST including a leucine zipper motif. In the case of the YOR3170c, a human EST was found 100% identical to the yeast sequence.

Two ORFs, YOR3177w and YOR3275c (identical to *PFY*), were predicted as intron-containing genes. Coding sequence for transposon peptide (YOR3367w) displays a +1 frameshift as is common for this type of sequence (Farabaugh, 1995). Among the most interesting protein identifications in this DNA fragment are (a) YOR3231w, an inositol polyphosphatase, the second found in yeast (homologous to the human OCRL gene involved in Lowe's syndrome); (b) YOR3172w, a new ADP ribosylation factor of the arf6 subfamily; (c) YOR3141c, the first protein containing three C2 domains; and (d) YOR3237w, an ORF similar to a *Bacillus subtilis* cell-cycle related protein.

Detailed analysis of selected ORFs

In the second step of the analysis, a detailed study of similarities between the different ORFs and sequences in databases was performed. In contrast to the high efficiency of the first step of database screening, the step of detailed analysis is not yet optimized and requires expert human intervention. Different cases, such as the ones described under (i) to (v), require a variety of sophisticated strategies. Analysed ORFs were divided into: (i) sequences highly homologous to other yeast sequences indicating duplication in the genome; (ii) sequences belonging to already known protein families; (iii) first reported yeast sequences in already known families; (iv) yeast sequences facilitating establishment of new protein families and (v) sequences without homologues.

(i) Sequences highly homologous to other yeast sequences indicating genome duplication Case 1: YOR3227w (46 551-48 365; 604 amino acids, aa) is a potential isoform of the yeast enzyme α -isopropyl-malate synthase. YOR3227w shows significant homology to the previously described yeast gene LEU4 (Beltzer et al., 1988), located on chromosome XIV, which encodes enzyme α -isopropyl-malate synthase. The *LEU4* gene has two alternative in-frame translation initiation sites, leading to two proteins with different lengths (619 and 589 aa). The larger form is imported into the mitochondria due to an 18-residue amphiphilic helix on its N-terminus whereas the other form remains in the cytoplasm. The existence of at least one other gene encoding isopropyl-malate synthase in yeast has been reported (Chang et al., 1985). Figure 2 shows an alignment of YOR3227w to the LEU1 YEAST sequence. The assumption that YOR3227w corresponds to the isoform of the yeast α -isopropyl-malate synthase is further supported by the following two observations: there is a methionine at a position where the alternative translation initiation starts in the LEU4 gene, and the presence and organization of many directed and inverted repeats in the 5'-flanking region of the YOR3227w are similar to those in the same region of the LEU4 gene (data not shown).

Case 2: YOR3177w (29 317-29 460+29 862-30 290; 190 aa) and YOR3165w (18 651–20 492; 613 aa) show homology to two adjacent ORFs on chromosome XIV. Dot-plot comparison between a 5.5 kb segment from chromosome XIV (accession no. X85811; Garcia-Cantalejo et al., 1994) containing Sctrnaorf_1 and Sctrnaorf_2, and the area between 17 500 and 31 000 from accession no. X94335 identified two homologous stretches around the region of YOR3177w and YOR3165w (data not shown). Although the ORFs Sctrnaorf 1 and Sctrnaorf 2 on chromosome XIV are adjacent to each other, the homologues on chromosome XV are interrupted by three ORFs over a distance of about 10 kb. The coding sequence of YOR3165w is homologous to Sctrnaorf_2. YOR3177w is

YOR3227w	# MVKHSPILIERA-SKIRRSIPOVKI.TYKNMIRDPSVKYREFEPKMVKRIWPDKTICKA
LEU1_YEAST	MVKESIIALAEHAASRASRVIPPVKLAYKNMLKDPSSKYKPPNAPKLSNRKWPDNRITRA
YOR3227w	PRWLSTDLRDGNQSLPDPMSVAQKKEYFHKLINIGFKEIEVSFPSASQTDFDFTRYAVEN
LEU1_YEAST	PRWLSTDLRDGNQSLPDPMSVEQKKEYFHKLVNIGFKEIEVSFPSASQTDFDFTRYAVEN
YOR3227w	APDDVGIQCLVQSREHLIKRTVEALTGAKRATIHTYLATSDMFREIVFNMSREEAISKAV
LEU1_YEAST	APDDVSIQCLVQSREHLIKRTVEALTGAKKATIHTYLATSDMFREIVFNMSREEAISKAV
YOR3227w	EATKLVRKLTKDDPSQQATRWSYEFSPECFSDTPGEFAVEICEAVKKAWEPTEENPIIFN
LEU1_YEAST	EATKLVRKLTKDDPSQQATRWSYEFSPECFSDTPGEFAVEICEAVKKAWEPTEENPIIFN
YOR3227w	LPATVEVASPNVYADQIEYFSTHITEREKVCISTHCHNDRGCGVAATELGMLAGADRVEG
LEU1_YEAST	LPATVEVASPNVYADQIEYFATHITEREKVCISTHCHNDRGCGVAATELGMLAGADRVEG
YOR3227w	CLFGNGERTGNVDLVTVAMNMYTQGVSPNLDFSDLTSISEIVHRCNKIPIPPRAPYGGEL
LEU1_YEAST	CLFGNGERTGNVDLVTVAMNMYTQGVSPNLDFSDLTSVLDVVERCNKIPVSQRAPYGGDL
YOR3227w	VVSAFSGSHQDAIKKGFAIQNKKQAQGETRWRIPYLPLDPKDIGRDYEAVIRVNSQSGKG
LEU1_YEAST	VVCAFSGSHQDAIKKGFNLQNKKRAQGETQWRIPYLPLDPKDIGRDYEAVIRVNSQSGKG
YOR3227w	GAAWVIMRSLGLDVPRPMQVDFSNTLQKNADALGRELKSEEITKLFKETYNYNNNEHIYV
LEU1_YEAST	GAAWVILRSLGLDLPRNMQIEFSSAVQDHADSLGRELKSDEISKLFKEAYNYNDEQYQAI
YOR3227w	TLLNYEVKKLNPERRALVGQVEINDKVVNIEGYGNGPISSLVDALSNLLNVKLSVQNYSE
LEU1_YEAST	SLVNYNVEKFGTERRVFTGQVKVGDQIVDIEGTGNGPISSLVDALSNLLNVRFAVANYTE
YOR3227w	HSLGSGSATQAASFINLSYIKDINNHATSNMWGVGVSEDTGDASIKAVFATVNNIIHSGD
LEU1_YEAST	HSLGSGSSTQAASYIHLSYRRNADNEK-AYKWGVGVSEDVGDSSVRAIFATINNIIHSGD
YOR3227w LEU1_YEAST	VLLAE VSIPSLAEVEGKNAAASGSA *



highly homologous to Sctrnaorf_1 as shown in the multiple alignment (Figure 3). The sequences belong to the S7, 40S ribosomal subunit protein family. Both yeast sequences are interrupted by an intron of 401 bp (YOR3177w) or 345 bp (Sctrnaorf 1), and have two exons of identical sizes of 48 and 142 aa, which indicates duplication in the yeast genome.

Case 3: ORFs YOR3193c (37 801–36 818; 327 aa) and YOR3299c (94 328–93 450; 292 aa) are similar to the yeast mitochondrial carrier protein YMC1. YOR3193c and YOR3299c belong to the diverged family of mitochondrial carrier proteins of several different substances (e.g. inorganic phosphate transporters, dicarboxylate exchangers etc.), which so far contains nine yeast proteins. The sequences of YOR3193c and YOR3299c display remarkable similarity to YMC1_YEAST (located on chromosome XVI; Graf *et al.*, 1993). All three ORFs form a new sub-family and point to a more recent duplication within the YMC1 branch. From the bootstrapping values it is possible to speculate that duplication of YOR3193c and YOR3299c might have occurred after the duplication which led to the YMC1 sequence. The presence of a delta element in the proximity of YOR3299c points to a possible involvement of transposition in this duplication process.

Case 4: YOR3116w (3059–3946; 295 aa) and YOR2964c (2714–465; 749 aa). Adjacent ORFs YOR3116w and YOR2964c are homologous to hypothetical yeast proteins YK71_YEAST (152 aa) and YK69_YEAST (910 aa), which are also adjacent on chromosome XI (Bou *et al.*, 1993; Garcia-Cantalejo *et al.*, 1994). BLAST scores are 2·7e-247 for YOR2964c/YK69_YEAST and 1·4e-4 for YOR3116w/YK71_YEAST. The orientation of the ORFs is maintained on both chromosomes. Since YOR3116w and YOR2964c are the first ORFs of the chromosomal fragment sequenced here, we cannot exclude that the region of duplication is extended beyond this point.

YOR3177w Sctrnaorf_1 RS7_HUMAN RS7_ANOGA Msrps7a_1 RS8_XENLA Rnrps8mr_1	MSAPQAKILSQAPTELELQVAQAFVELENSSPELKAELRPLQFKSIREIDVAGGKKALAIP MSSVQSKILSQAPSELELQVARTFIDLESSSPELKADLRPLQIKSIREIDVTGGKKALVLF MSSVAKIVKNGKEPPEREGSISQLLELENNS-DLKAQLRELNITAAKEIEVGGGRKAIIF WSTGKIKAGOTEADSFETSIGALVELENNS-DLKAQLRELYITKAESVEFNI-KKAIII MSTGIKLAGOTEADSFETSIGALVELENNS-DLKAQLRELYITKAESVEFNI-KKAIIIF MFSTSAKIVKPNGEKPDEFESGISQLLELENNS-DLKAQLRELNITAAKEIEVGAGRKAIIF MSSSAKIVKPNGEKPDEFESGISQLLLELENNS-ELKTQLWELNITAAKEIEVGAGRKAIIF	VP VP VP VP VP VP
YOR3177w Sctrnaorf_1 RS7_HUMAN RS7_ANOGA Msrps7a_1 RS8_XENLA Rnrps8mr_1	VPSLAGFHKVQTK-LTRELEKKFQDRHVIFLAERRILPKPSRTSRQVQKRPRSRTLTAVHD- VPALSAVHKVQTK-LTRELEKKFPGDRHVIFLAERRILPKPSRTSRQVQKRPRSRTLTAVHD- VPQLSFQKUQVFLVRELEKKF9GKHVVFJAQRILPKPRKSRTKNKQKRPRSRTLTAVHD- VPCQSAFQKUQTR-LVRELEKKF9GKHVVFJAQRILPKPHKSRTKNKQKRPSSRTLTAVHD- VPQLSFQKIQTR-LVRELEKKF9GKHVVFJAQRILPKPTRKSRTKNKQKRPSSRTLTAVHD- VPQLSFQKIQVR-LVRELEKKF9GKHVVFJAQRILPKPTRKSRTKNKQ-KAQKPTLTSSADR	KI AI AI AI RI
YOR3177w Sctrnaorf_1 RS7_HUMAN RS7_ANOGA Msrps7a_1 RS8_XENLA Rnrps8mr_1	LEDLVFPTEIVGKRVRYLVGGNKIQKVLLDSKDVQQIDYKLESFQAVYNKLTGKQIVFEIPSET LEDMVFPTEIVGKRVRYLVGGNKIQKVLLDSKDVQQIDYKLESFQAVYNKLTGKQIVFEIPSET LEDMVFPSEIVGKRIRKLDGSLIXVHLDKAQQNNVEHKVETFSGVYKKLTGRDVTFEPFEN LEDLVFPSEIVGKRIRKKLDGSLIXVHLDKAQQNNVEHKVETFSGVYKKLTGRDVTFEPFEN LEDLVFPSEIVGRRIRKKLDGSLIXVHLDKAQQNNVEHKVETFSGVYKKLTGRDVTFEPFEN LEDLVFPSEIVGRRIRKKLDGSLIXVHLDKAQQNNVEHKVETFSGVYKKLTGRDVTFEPFEN LEDLVFPSEIVGRRIRKKLDGSLIXVHLDKAQQNNVEHKVETFSGVYKKLTGRDVTFEPFEN LEDLVFPSEIVGRRIRKLDGSLIXVHLDKAQQNNVEHKVETFSGVYKKLTGRDVFFEPFEN	H L L V

Figure 3. Case 2. Multiple alignment of the S7, 40S ribosomal subunit protein family. The alignment includes YOR3177w, its homologue Sctrnaorf_1, a translation product from yeast (X85811), and examples from human (RS7 Human), *Xenopus laevis* (RS8_XENLA), *A. gambiae* (RS7_ANOGA), a ribosomal S7 protein from an insect *M. sexta* (Msrps7a 1) and rat (Rnrps8mr_1). The sequences labeled as S8 (RS8_XENLA and Rnrps8mr_1) were annotated in the database by mistake as S8 sequences but belong to the S7 family. Completely conserved residues are indicated with (*), highly conserved ones with (.).

(ii) Sequences belonging to a protein family that sequences Case contains other veast 5: YOR3231w (48 799-52 122; 1107 aa) belongs to the OCRL-inositol polyphosphatase family. We have previously identified a yeast ORF in the centromeric region of chromosome IX (YIA2 YEAST) as a member of the OCRLinositol polyphosphate-5-phosphatase family (Voss et al., 1995). Inositol polyphosphate-5phosphatases catalyse the conversion of inositoltrisphosphate to inositol-bisphosphate. Defects of the human homologue OCRL_HUMAN have been reported to be responsible for Lowe's oculocerebrorenal syndrome, a nervous system disorder that causes mental retardation in addition to other symptoms (Attree et al., 1992; Leahey et al., 1993). Here we describe YOR3231w as the second yeast sequence that belongs to this family. Figure 4a shows that the C-terminal half of YOR3231w aligns well with the other members of the family. In addition, the N-terminal regions of both YOR3231w (aa 60-530) and YIA2_YEAST (aa 128-575) show clear homology to another yeast sequence, RSD1 YEAST, a recessive suppressor of secretory defect (623 aa long; Cleves et al., 1989; Figure 4b). YOR3231w and YIA2 YEAST may constitute natural links between the inositol polyphosphatase function in the N-terminal region and a C-terminal function related to the secretory pathway.

Case 6: YOR3348c (116 111–113 466; 881 aa) belongs to the actin family. Eukaryotic actin genes usually contain a single intron in the coding region (Weber and Kabsch, 1994). Two yeast actin genes are known: ACT_YEAST and ACT2_YEAST (both 391 aa long, with introns of 305 bp and 124 bp, respectively). ORF YOR3348c shows significant homology to both yeast actin genes. Interestingly, if BLASTX is used for comparison of the DNA sequence, the actin hits appear in the middle of the 881 aa long ORF YOR3348c. However, raw sequencing data are unambiguous, and Northern Blot analysis supports the existence of a transcript with the length of YOR3348c (data not shown).

Case 7: YOR3172w (26 318-26 869; 183 aa) is similar to the ADP-ribosylation factors (ARFs). ADP-ribosylation factors are essential and ubiquitous in eukaryotes. They are involved in vesicular transport and specifically function as activators of phospholipase D and of cholera toxin. The functions of ARF proteins in membrane traffic and organelle integrity are tied to their reversible association with membranes and specific interactions with membrane phospholipids. A common feature of ARFs is their regulation by the binding and hydrolysis of GTP. The arf family now includes 18 sequences from different species, including previously characterized highly homologous yeast genes ARF1 YEAST and ARF2 YEAST. Interestingly, the new yeast sequence YOR3172w

H. VOSS ET AL.

YOR3231w 559	KFTSTSNINLLIGSFNVNGATK-KVDLS-WLFPIGEKFKPHTVVLGLOEVIE
YIA2_YEAST 520	KTIFERDISIFAGTFNISGKIP-KDDIKDWIFFKSWSKEDEMADLYVIGLEEVVE
OCRL_HUMAN 308	EYVNIQTFFFFVGTNNVNGQSP-DSGLEPWLNCDPNPPDIYCIGFQE-LD
IT5P_HUMAN 55	DYTYIQNFRFFAGTYNVNGQSPKECLRL-WLSNGIQAPDVYCVGFQE-LD
YLJ7_CAEEL 18	DSEAVENNINGMIDDETEVAIGLQEV
YOR3231w	LSAGSILNADYSKSSFWENLVGCLMYDDKYLLLRVEQMTSLLILFVKADK
YIA2_YEAST	LTPGHMLATDFYKQFWEKKILTLNGPGRKKYIRLWSTQIGGILLLFMNTTE
OCRL_HUMAN	LSTEAFFFFEVKEQUENAVERGHKAKYKKUUVLVGUMLLFARKD
IT5P_HUMAN	LSKEAFFFEDTFKEEEMFKAVSEGLHPDAKYAKVKLIRLVGIMLLIYAKCD
YLJ7_CAEEL	ABSETIGGAVLTWATTIASWMTNGRWVLLAKTFQATNQVLIFGRKQL
YOR3231w	AKYVKQVEGATKKTGFRGHAGNKA-VSIRFEYGA-TSFCFVNSELAAGATNVEER
YIA2_YEAST	YSKVRIEGUVKKYGFGHASNKGAVAVSFKYSA-TSFCLVSELAAGLENVEOR
OCRL_HUMAN	CRVIRDIATETVGTGINGKNGNKGVAVRFVFHN-TTFCIVNSELAAHUSDFERA
IT5P_HUMAN	AAVISVEAETVGGTGINGKNGKNGGVAIRFOFEN-TSICVVNSELAAHIEFYER
YLJ7_CAREL	IGQIKRIDYRFORNTMGGLTGERGSIGVRLOLASFYSIVFVDSFHEGPENYGKR
YOR3231w YIA2_YEAST OCRL_HUMAN IT5P_HUMAN YLJ7_CAEEL	RSDYESIVRGITFTRTKMIPHEDSIFWLGDMNYRINLPNEDVRRELLNO HNDYKTIAKSIRFSKGIRIKDEDAIIMAGPFNYRILMSNEDVRRKIVSK NODYKDICSRMSFVORDO-TPOLNINKENVVINGLNYRIEELDVRKVKKLIEE VEQYETN-RNCSFP-EDKSVRAAFFGDFNYRVEEDVNTVIRKIKNG
YOR3231w YIA2_YEAST OCRL_HUMAN IT5P_HUMAN YLJ7_CAEEL	$\begin{split} \textbf{EEGYIDKLLHFDQ1}-LGINSGSVFEGFKEPTLKFRPTTKYDPGTGTYDSSEKE\\ \textbf{EY}ASLFEKDQ1NQQMIAGSSFYFEFAIDFPTTKFDPGTKNTDTSEKM\\ KDLGRLKHPQL-IQFRCKAFVDFMEBEIKFIFTKYDSKTKTDTSEK-SKKDFQMLXATDQKIQVAAKTVFSGPTEBEIKFIFTKYDSKTKTTGIFVRSAVTHLELLDTKSQLKRALVERDAFIGFEDPVTFEPTWTVGTGTEQDGK**$
YOR3231w YIA2_YEAST OCRL_HUMAN IT5P_HUMAN YLJ7_CAEEL	RTPSWTDRIIY-GENLLPLSY-SDAPIMISDHRPVYAAYRAKITFVDDK 866 RIPAWTDRILSRGSVLEQLEYKCCEDILFSDHRPVYAIFRAKVTVVDEQ 830 RVPAWCDRILW-GYNVNGUNYSSHUELKTSDHRPVSAIFHIGVKVVDE 364 RVPSWTDRILYKGGIGLSYNNKKAVAJDHLPVVAHPKVTAPAAPKP 289
YOR3231w	60 FGVIGLIEVNGLLFVGAITGKSKVAQFCP
RSD1_YEAST	2 TGFIYYVQNADGIFFKLAEGKGTNDAVIHLANQDQGVRVLGAEEFPVQ
YOR3231w	GETVNKIFAVDFFCLNDNSWDFIEIDSSG 118
RSD1_YEAST	GEVVKIASLMGFIKLKLNRYAIIANTVEETGRFNGHVF 87
YOR3231w	119 YPVLPETASTEYQDALPKHPCYELKKLLSNGSFYYSSDFDLTSTQ
RSD1_YEAST	88 YRVLQHSIVSTKFNSRIDSEEAEYIKLLELHLKNSTFYFSYTYDLTNSQ
YIA2_YEAST	128 DYLLCERSEQNYDKLIHEHPCGPLKKLFSDGTFYYSRDFDISNIK
YOR3231w	LHRGY-GQHSLSTDTYEEEYMWNSFLMQEMITYRDHLDTNLKQILDDEGFLTTVI
RSD1_YEAST	LRNEK-VGPAASWKTADERFFWNHYLTEDLRNF-AHQDPRIDSFIQPVI
YIA2_YEAST	VNHGLSHNLEYTVDNQDLSFIWNANLASEVINWRSKISNEEKQLFANAGFLTFVI
YOR3231w	RGFAETFVSYVKKLKVALTIISKQSWKRAGTRFNARGVDDEANVANFVETEFIMY
RSD1_YEAST	YGYAKTVDAVLNATPIVLGLITRRSIFRAGTRVFRGVDKDGNVGNFNETEQILL
YIA2_YEAST	RGYCKTALIEDGPNTASITIISRISTESKQDTLELEGISEDGRVSLFVETEIVVT
YOR3231w	SSOFYCYATQIRGSIPVFWEQCTSLINPRVQITRSFEATQPVFDKHI
RSD1_YEAST	AENPEPSEKIHVFSLOTRGSVPIYWABINNLKYKENLVLGENSLDATKKHF
YIA2_YEAST	TEKYFIFSTQVNGSIPLFWESVESOLLYGKKIKVTKDSIEAQGAFDRHF
YOR3231w	MKSVEKYGPVHVVNLLSTKSSEIELSKRYKEHLTHSKKLNFNKDIFLTEPDPHKE
RSD1_YEAST	DQQKELYGDNYLVNLVNQKGHELPVKEGY-ESVVHALNDPKIHYVYFDFHHE
YIA2_YEAST	DNLTSKYGVVSIVNIIKPKSESQEKLALTYKDCAESKGIKITNIEYSS
YOR3231w	TSQEGFSGVRKLIPLILDSLLSSGVYSYDVREKKNISEOHGIFRINCLDC
RSD1_YEAST	CREMQWHRVKLLIDHLEKLGLSNEDFFHKVIDSNGNTVEIVNEOHSVVRINCMDC
YIA2_YEAST	VLIKSPHKLLVLLKODIYEFGAFAYDISRGIYFAKOTCV_RISAFDS
YOR3231w	LDRINLAQQIISLAAFRIFLEDFRLISSNSFIDDDD-FVSKHNTLWAD 471
RSD1_YEAST	LDRINVVQSVLAQWVLQKEFESADVVATGSTWEDNAPLLISYQNLWAD 443
YIA2_YEAST	IEKPNIVERLVSKEVLELTINEIDVFELISFFLDAHDKLWSENYYWLD 475
YOR3231w	472 HGDQISQIYTGTNALKSSPSRKGKMSLAGALSDA
RSD1_YEAST	444 NADAVSVAYSGTGALKTDFTRTGKRTRLGAFNDF
YOR3231w	TKSVSRIYINNFMDKEKQQNIDTLLG 531
RSD1_YEAST	LNSASRYYQNNWTDGPRQDSYDLFLG 503

Figure 4. Case 5. YOR3231w. Alignment of sequences: (a) C-terminal against OCRL_HUMAN and its homologues and (b) N-terminal against RSD1_YEAST and YIA2_YEAST.

is more closely related to the arf6 sub-family than to the previously known ARF1_YEAST and ARF2_YEAST sequences. Comparison of the new sequence with the whole family points to 14 different residues in positions previously found to be conserved within the family (data not shown). The reduction of the number of conserved residues helps to shape the key functional regions of this



Figure 5. Case 7. The conservation levels in the arf family are mapped in the three-dimensional structure of the human ADP-ribosylation factor 1 complexed with GDP. The backbone is coloured in orange for those residues conserved in the whole arf family. White colour and side chains indicate residues conserved in the family except in YOR3172w. The substrate, GDP, and the Mg^{2+} ion are shown in ballmodel.

family. Figure 5 shows conservation levels in the arf family displayed in the three-dimensional structure of the human ARF1_HUMAN complexed with GDP (Amor *et al.*, 1994). For more detailed protein sequence analysis of the arf6 family, see Valencia and Sander (1995). Non-conserved residues in YOR3172w involve the GDP binding site (Asp-Cys exchange in position 159, Lys-Pro exchange in position 131), the opposite face of the active centre (Val-Ala exchange in position 119, Val-Tyr exchange in position 167) and residues shown to interact with Mg²⁺ ions (Ser-Thr exchange in position 45).

Case 8: YOR3141c (10 305–6745; 1186 aa) contains three C2 domains. C2 domains, probably involved in Ca²⁺ and phospholipid binding, have been described in different protein families such as Ca²⁺-dependent protein kinases C (Clark *et al.*, 1991); synaptotagmines, which are related to syn-

aptic vesicle traffic control (Sossin and Schwartz, 1993), and in C. elegans phorbol ester/DAG binding protein unc-13 (Bork and Sudol, 1994). Usually one or two C2 domains are present in these proteins. Analysis of YOR3141c revealed for the first time the presence of three C2 domains (termed YOR3141c1, c2 and c3) in a protein. Figure 6a shows a multiple alignment of the C2 domains from this ORF with their counterparts from other proteins (RSP5_YEAST, KPC2_HUMAN, SYT1_RAT, PIPA_DICDI, UN13_CAEEL and PIR:A42142). The beta sheets derived from the published three-dimensional structure of the first C2 domain of SYT1 RAT (Sutton et al., 1995) are shown with boxes. The proteins did not align for the first beta sheet of the structure calculated. This first beta sheet is not even present in the case of RSP5 YEAST. Thus only the last seven are shown (denoted as $\beta 2-\beta 8$). The phylogenetic analysis of

H. VOSS ET AL.

OR3141C/1	385	IGILEITVKNAKGLKRTSS-ILNESIDPYLSFEFNDISIAKTRTVRD-TLNPVWDETLYVLL
OR3141C/2	657	IGAIRVFIEKANDLRNLEKFGTIDPYCKVLVNGLSKGRTDFKSQ-TLNPVWNQVIYVAVT-
OR3141C/3	991	SGDLTIMSRSAENLIASDLNGYSDPYLKYYINNEEDCAYKTKVVKK-TLNPKWNDEGTIQIN-
RSP5_YEAST	2	PSSISVKLVAAESLYKRDVFRSPDPFAVLTIDGYQTKSTSAAKK-TLNPYWNETFKFD
SYT1_RAT/1	155	NNCLLVGIIQAAELPALDMGGTSDPYVKVFLLPDKKKKFETKVHRK-TLNPVFNEQFTFKVP-
SYT1_RAT/2	285	AGKLTVVILEAKNLKKMDVGGLSDFXVKIHLMQNGKRLKKKKTTIKKN-TLNFYYNESFSFEVP-
PIPA_DICDI	673	YSRLIVNVISARQLPKYTKSTKGEVIDPYVTLSIVGTHFDQKVEKTKVIDNNGFNPHWGEEFEFPLYN
JN13_CAEEL	736	SAKITLTVLCAQGLIAKDKTGKSDPYVTAQVGKTKRRTRTIHQ-ELNPVWNEKFHFECH-
42142	275	HGRFVGVTIKVPACVDLAKKQGTCDPFVVCTAHYSNKHQVTRRTKQRKK-TVDPEFDEAMYFDLHI
PC2_HUMAN	170	RDVLIVLVRDAKNLVPMDPNGLSDPYVKLKLIPDPKSESKQKTKTIKC-SLNPEWNETFRFQLK-
		1000 000 1000 1000 1000 1000 1000 1000

COR3141C/1	NSFTDP-LTISVYDKRAKLKDKVLGRIQYNLNTLHDKT	481
/OR3141C/2	SPNQRITLQCMDVETVNKDRSLGEFNVNVQDLFKKDENDKYEETI	760
COR3141C/3	NRLNDVL-RIKVMDWDSTSADDTIGTAEIPLNKVKVEGTTELDVPVEGL	1099
RSP5_YEAST	DINENSILTIQVFDQKKFKKKDDGFLGVVNVRVGDVLGHLDED	101
SYT1_RAT/1	-YSELGGKTLVMAVYDFDRFSK-HDIIGEFKVPMNTVDFGHVTEEWRDLQ	263
SYT1_RAT/2	-FEQIQKVQVVVTVLDYDKIGKNDAIDKVFVGYNSTGAELRHWSDML	394
PIPA_DICDI	SQLSMLLIRVDDKDKVGHNRIGHHCIRVENIRPGYRILKLKNNF	784
JN13_CAEEL	NSTDRIKVRVWDEDNDLKSKLRQKLTRESDDFLGQTVIEVRTLSGEMDVWYNLE	847
42142	DADAGSTNTTGSNKSAGSLESSANKGYSIYPVGGADLVE-IVVSVWHDAH	388
CPC2_HUMAN	ESDKDRRLSVEIWDWDLTSRNDFMGSLSFGISELQKASVDGWFKLL	278



Figure 6. (a) Case 8. Alignment of the three C2 domains from YOR3141c with the C2 domains of other representative proteins. The alignment extends around 120 aa. The proteins are indicated by their SwissProt identifiers: KPC2_HUMAN, protein kinase C; PIPA_DICDI, phosphalidyl inositol phosphodiesterase; SYT1_RAT, phospholipase c (their two C2 domains marked /1 or /2); A42142 (PIR database identifier), gap protein from *Drosophila*; RSP5_YEAST, translation product from yeast; UN13_CAEEL, phorbol ester/DAG binding protein. YOR3141c/1, 2 and 3 are the C2 domains deduced from YOR3141c. Note that the only yeast sequences are the RSP5_YEAST and the new ORF. The beta sheets as derived from the three-dimensional structure of the first C2 domain of rat synaptotagmin (Sutton *et al.*, 1995) are shown with open rectangles. The most conserved residues are shown in grey boxes. The importance of the 'G' in the centre of $\beta7$ is pointed out by the fact that it is mutated into a 'D' in the second C2 domain of c2 domains. The low bootstrapping values indicate the high divergence of the domain.

the C2 domains is shown in Figure 6b. YOR3141c domains 1 and 2 are closely related to each other while domain 3 is more related to synaptotagmin C2 domains (SYT_RAT1, 2).

(*iii*) First yeast sequences in already known protein families In several cases ORFs deduced from the accession no. X94335 could be identified as the

© 1997 by John Wiley & Sons, Ltd

first yeast member in an existing known protein family. These new findings could have interesting consequences for further biological and phylogenetic characterization of the protein families involved.

Case 9: YOR3352w (116 577–117 566; 329 aa) belongs to a family of CoA-ligases. YOR3352w is the first yeast member belonging to the family of



Figure 7. Case 10. YOR3224w is similar to two ORFs from human Hsg0s8pp_1 (Hong *et al.*, 1993) and Hs1r20rna_1 (Siderovski *et al.*, 1994), and to one ORF from *C. elegans* (YK07_CAEEL). Multiple sequence alignment with secondary structure prediction using PHD (Rost, 1994). Completely conserved residues are indicated by a (*), highly conserved residues are marked by a (.). The two helices previously predicted for Hs1r20rna 1 by Siderovski *et al.* are shown as black boxes.

CoA-ligases with known members from animals, plants and bacteria. It shows very high homology (identity approx. 60%) over the whole sequence.

Case 10: YOR3224w (44 876-45 805; 309 aa) is the fourth member in a family of proteins containing a previously incorrectly assigned helix-loophelix (HLH) motif. YOR3224w is similar to human ORFs Hsg0s8pp 1 (Hong et al., 1993) and Hs1r20rna 1 (Siderovski et al., 1994) and to a C. elegans ORF YKO7 CAEEL. In this case we propose a new protein family whose members are homologous and share common features in their secondary structure prediction. A previous comparison between the human ORF Hs1r20rna 1 and HLH proteins (e.g. transcription factors) was based on a weak homology and on secondary structure prediction. The analysis presented here, obtained from alignment with mutually highly homologous sequences, allows a more precise definition of the family based on a more accurate secondary structure prediction deduced from general properties of the family rather than from individual sequences. The previous assignation of Hs1r20rna 1 to HLH proteins was based on a similarity in two regions, the QTK and EAxKE motifs. However, Figure 7 (alignment obtained for four homologous sequences) clearly shows that the these motifs are not conserved within the family. Furthermore, Siderovski et al. (1994) have suggested the similarity of the ORF Hs1r20rna 1 with HLH proteins based on a secondary structure prediction of an HLH motif achieved by the Chou-Fasman method (Gribskov and Devereux, 1991). A new prediction performed for the newly established family indeed indicates the presence of two alpha helices for all members, but at different positions than in the classical HLH motif.

Case 11: YOR3120w (4113–5276; 387 aa) shares conserved motifs with prokaryotic members of the lipase-esterases family. ORF YOR3120w matches the PROSITE motif for lipases for the serine active site (PS00120) and represents the first eukaryotic sequence found with this motif. YOR3120w matches with a subset of the whole lipase-esterases family. The alignment shown in Figure 8 indicates an extension of the consensus sequence around the PROSITE pattern.

(*iv*) Definition of a new protein family facilitated by the new sequence With the progress of the genome sequencing projects, protein sequences without known function accumulate in the databases. Definition of new families where only sequences are available, but no biological information, can give important hints for the search for protein functions.

Case 12: YOR3237w (53 950-54 648; 232 aa) is the first eukaryotic member of a protein family presumably related to prokaryotic cell cycle proteins. Figure 9 shows that YOR3237w is similar to three proteins from B. subtilis and E. coli. However, only little functional information is available for the B. subtilis ORF MAF_BACSU, which is coded by the *spoIIB* gene. This gene is flanked by many cell-cycle-related genes on the bacterial chromosome. It has been shown experimentally that MAF BACSU is involved in the cell cycle and particularly in septum formation: mutations in the spoIIB gene do not usually lead to a significant alteration of the spore formation, but if mutations in this gene are combined with inactivation of another sporulation gene (spoVG), the joint effect of the defective genes is an interruption of sporulation at the stage of septum formation (Margolis

H. VOSS ET AL.

		hhhhHG .	
YOR3120w	53	3 LNLVFLHGSGMSKVVWEYYLPRLVAADAEGNYAIDKVLL1	IDQVNHGDSAVRNRGRL
Mm0361_1	19	ENLIFVHGYNSSPRTFEYL-KNIQQDQIIMHYNFQDQ-IYVH	KPVKDH-KVTVEGFAQL
Mm0361_2	20) LNIIYIHGFNSSYKAFEIFEK-YWTKTNYYSIQFPGSQ-LVH	KPVKNH-KVSVEGFAQL
Mm0361_3	20) ENTIFCHGFNASINVFNIF-KNYWTKSNYYALQSPGNNN-VI	KPIKDD-EISVLQFAKL
Sgnonr_2	39	PGLALTHGAGADHRMFDPQIPALVGAGYRTLRWDVRYHGA	ASVSGTGRFRTVYAARD
Mgrpobc_1	27	7 HKLVFLHGFGENFKIKRRLWEYYDNCSFYALNLPGHGES	SKIQDPNQLSIAYFAQI
Ppacox_4	134	TPLVLVHGFGGDLNNWLFNHPALAAERRVIALDLPGHGH	ESAKALQRGDLDELSET
BPHD_PSES1	34	ERVIMLHGGGPGAGGWSNYYRNIGPFVEAGYRVLLPDAPGFNKSD	TVVMDEQRGLVNARSVK
		. ** .	
		hhG SMGG	PhL
YOR3120w		GINFNWIDGARDVLKIATCELGSIDSHPAL-NV-VIGHSMGGFOAI	LACDVLOPNLFHLLILIE 169
Mm0361_1		LIHFIEONOIK-NVVAIGHSMGGGVIS	SIRYKMRPDLFKKLIFIT 116
Mm0361_2		LIDFIEONOIK-NVVAVGKSMGGGTL	AIAYKMEPDLEKKLIFIT 117
Mm0361_3		VVEFVKNNNLK-NVTLIGHSMGGGTI	SLAYOLAPELFKKLVYVC 117
Sgnonr_2		LAALLDAAGMPRPVVLLGOSMGGNIA	OEYLERRPDEVAALVVIG 138
Mgrpobc_1		IKAYFEKHDLK-DVILLGHSMGGGLA	AIMNSLIPERIK-LSVLE 123
Ppacox_4		VLALLDHLDIAKAHLGHSMGGAVS	LNVAGLAPORVASLSLIA 231
BPHD_PSES1		GMMDVLGIEKAHLVGNSMGGAGA	LNFALEYPERTGKLILMG 136
		* ****	* *

Figure 8. Case 11. YOR3120w. Alignment against products of the lipaseesterase operon from Mycoplasm (Mm0361_1, Mm0361_2 and Mm0361_3), Sgnonr_2 (antibiotic-resistance protein, 279 aa, *Streptomyces griseus*), Mgrpobc_1 (unknown product, Mycoplasm), Ppacox_4 (dihydrolipoamide acetyltransferase, Pseudomonas) and BHPD_PSES1 (2-hydroxy-6-oxo-6phenylhexa-2,4-dienoate hydrolase, Pseudomonas). The three conserved boxes are marked: <u>hhhh</u>HGx[G/N], <u>hh</u>GxSMGG, and Pxx<u>h</u>xxL (<u>h</u> is a hydrophobic amino acid).

YOR3237w YHDE_ECOLI MAF_BACSU YCEF_ECOLI	LASXXPXR 21 ILLASTSPRRVEILHDINGITDLKTNVSTFEENLDKMNYSTDPIGYVCDTSWHKAQNIIEIL/DYEDEN 4 LYLASGSPRRVEILAQLGVTFERIVTGIEEQRQPQESAQQVVVRLAREKARA 5 LILASQSPRRKELLDLLQLPYSIIVSEVEEKLNRNFSPEENVQMLAKQKAKA 17 LILASTSPWRKALLEKLQISFECAAPEVDETPRSDESPRQLVLRLAQEKAQS *****	
YOR3237w YHDE_ECOLI MAF_BACSU YCEF_ECOLI	- PNEIDKPKLIICADTIIIDKSGRIYEKPKTKEVQKKFLMKFCYEDDEPVNVVTAVTLIKWYGRENFEL GVAQTAKDLPVLGADTIVIL-NGEVLEKPRDAEHAAQMLAKLSGTHQVMTAVALADSQHIL -VADLHPHAIVIGADTMVCL-DGECLGKPQDQEEAASMLRLSGRSHSVITAVSIQAENHSE -LASRYPDHLIIGSDQVCVL-DGEITGKFLTEENARLQLRKASGNIVTFYTGLAFNSANGHL	
YOR3237W YHDE_ECOLI MAF_BACSU YCEF_ECOLI	GLP VPFRDETKVYFDNKIPLRILEEYVESGDGLEVGGGFKIQGQGAILIEKIEG-DYYNVGLPLKNKFKGL 2: DCL-UVTDVTFRTLTDEDIA-GYVASDDPLDKAGAYGIQGLGCCPVRKING-SYHAVVGLPLVETYELL 1: TFY-DKTEVAFWSLSEEEIW-TYIETKEPMDKAGAYGIQGRGALFVKKIDG-DYYSVMGLPISKTMRAL 1: QTEVEPFDVHFRHLSEAEID-NYVRKEHPLHCAGSFKSEGFGTTLFERLEGRDPNTLVGLPLALCOML 1	25 82 82 97

Figure 9. Case 12. YOR3237w is similar to *B. subtilis* and *E. coli* proteins with possible roles in the cell cycle. Multiple sequence alignment with main motifs boxed. Asterisks indicate residues conserved throughout the whole family; dots mark similar residues.

et al., 1993). This finding suggests, though indirectly, the possible involvement of YOR3237w in the yeast cell cycle. Alignment of MAF_BACSU with the deduced protein sequence of YOR3237w revealed the presence of two conserved regions at the N- and C-termini. Searches with the conserved sequence patterns LASxSPxR and GLP expanded the alignment by YHDE_ECOLI and YCFE_ECOLI, which are hypothetical proteins without known function. We propose the inclusion of these two sequences as additional representatives of this new family.

Case 13: YOR3174c (27 851–27 075; 258 aa) is the first homologue of E. coli ribose-5P isomerase (RPIA_ECOLI). Figure 10 shows a sequence alignment of YOR3174c to RPIA_ECOLI. The presence of the second sequence allows us to pin-point conserved residues which may facilitate determination of the active site. From the 71

YOR3174c	MAAGVPKIDALESLGNPLEDAKRAAAYRAVDENLKPDDHKIIGIGSGSTVVY
RPIA_ECOLI	
YOR3174c RPIA_ECOLI	VAERIGQYLHDPKFYEVASKFICIPTGFQSRNLILDNKLQLGSIEQYPRIDI FIDALGTMKGQIEGAVSSSDASTEKLKSLGIHVFDLNEVDSLGI *
YOR3174c	AFDGADEVDENIQLIKGGGACLFQEKLVSTSAKTPIVVADSRKKSPKHLGKN
RPIA_ECOLI	YVDGADEINGHMOMIKGGGALTREKIIASVAEKFICIADAS-KQVDILGK-
YOR3174c	WRQGVPIEIVPSSYVRVKNDLLEQLHAEKVDIRQGGSAKAGPVVTDNNNFII
RPIA_ECOLI	PPLPVEVIPMARSAVARQLVKLGGRPEYRQGVVTDNGNVIL
YOR3174c	DADFGEISDPKLHREIKLLUGVVETGLFIDN-ASKAYFGNSDGSVEVTEK
RPIA_ECOLI	DVHGMEILDPIAMENAINAIPGVVTVGLFANGADVALIGTPDGVKTIVK-

Figure 10. Case 13. Full sequence alignment of YOR3174c to RPIA_ECOLI. Highly conserved motifs are boxed, conserved residues are marked by an (*), similar residues are indicated as a (.).

residues conserved between these two sequences the best candidates for active site functions are four clusters with characteristic patterns of Gly

YOR3510c	321	GEANSSLSEFVPL	4ILHG-	NSIGKK	TLIC	TIMRE	AGDDNS	YQIYEV	NSNMNRS	KKDLL	DIL	
RFC1_YEAST	336	KHAGKDGSGVFRA	MLYGP	PGIGKT	TAAF	ILVAQEI	GYDILE	QNASDV	RSKT-LL	NAGVK	NAL	
GNF1_DROME	470	PWAKNDDGSFYKA	ALLSGP	PGIGKT	TTAT	LVVKE	GFDAVE	FNASDT	RSKR-LI	KDEVS'	ГLL	
AC15_HUMAN	634	KFSGKDDNSSFKA	ALLSGP	PGVGKT	TTAS	SLVCQE	GYSYVE	LNRSDT	RSKS-SL	KAIVA	ESL	
			.* *	.**	*	*			*		*	
							_					
YOR3510c		LDFTTTHYVK!	OSSKRK	SDYGLA	/LFN	DVDVLFI	KEHDRGY	WAMISK	LCEFSRF	PLVLT	CKDL	441
RFC1_YEAST		DNMSVVGYFKHNE	EAQNLN	IGKHFV1	IMD	VDGMS	GG-DRGG	VGQLAQ	FCRKTST	PLILI	CNER	458
GNF1_DROME		SNKSLSGYFT	GQGQAV	SRKHVI	лиц	EVDAMA	GNEDRGG	MQELIA	LIKDSSI	PIICM	CNDR	590
AC15_HUMAN		NNTSIKGFYSN	GAASSV	STKHAI	IML	EVDGMA	GNEDRGG	IQELIG	LIKHTKI	PIICM	CNDR	755
						* *	***			*	*	

Figure 11. Case 14. Multiple alignment of YOR3510c with DNA binding proteins GNF1_DROME, AC15_MOUSE, AC15_HUMAN and RFC1_YEAST. Homologous profiles are boxed, conserved residues are marked by a (*), similar residues are indicated as a (.).

YOR3513c	45		VPPHRMTPLRNSWTKIYPPLVEHLKLQVRMNLKTKSVELRT 85
D22835	252 1	bp	VPQHAFAPLKKAWMDIYNPVYEHMKIDIRMNLKARRVELKT 374 bp
T10779	140 1	bp	VPANRYTPLKENWMKIFTPIVEH 208 bp
		-	** ** * * * ** * **** ****
YOR3513c	148		RIAGKDGKTKFAIENATRTRIVLADSKIHILGGFT
D28195	3 1	ad	RLSGKGGKXKYAIENSTRTRIVIADTKIHILGSFV
T40124	2 1	h'n	RIAGKGKTKETTENVTRTRIVLADVKVHILGSFO
R03754	274 1	5p	TI GAYO
T38107	167 1	hp.	*Dout &
150107	107 3	υp	* ** ** * *** ****** ** * ****
VOP35130			UTEMADECIAICI TI CEBBOULVOM DINIA CDI VERV. 019
D20105			NIKUARCESVISEIDOSFFORVIGNERIVASKERERI 210
D20190			WINVARDSDCSDIDGSPAGRVISKARAVSARDAERI 212 DD
140124			NIKMARTAICNLILGNPPSKVYGNIRAVASRSADRF 211 bp
R03754			NLKLARNAVCSLILGSNPSKVYGNLRXMASRGAER- 150 bp
T38107			VSLILGSPPGKVYGDXRTVASRLKERY 90 bp
			** **** *** * *

Figure 12. Case 15. Alignment of ORF YOR3513c against several ESTs (see text for details). Asterisks indicate positions where all the fragments show identical residues. No gaps were allowed for the alignment.

residues: GxGxGST, DGADE, QxIKGGGA and GxxDG (boxed).

Case 14: YOR3510c (128 612–126 237; 791 aa). Figure 11 shows that YOR3510c can be aligned to a small set of DNA binding proteins coding for transcription and replication factors. The alignment allows the definition of profiles like LxGxxx-<u>hGKxTxxxhxEh</u> or <u>aVDxhxxxDRG</u>, where × is any amino acid, <u>h</u> is a hydrophobic residue, and a denotes an acid residue.

Case 15: YOR3513c (129 523–128 867; 175 aa). It is very likely that YOR3513c is expressed since its EST (T39061) is already known. Figure 12 shows that YOR3513c is highly homologous in the N-terminus to human D22835 and rice T10779 ESTs and in the C-terminal region to rice D28195, yeast T38107, R03754 (*C. briggsae*) and human T40124 ESTs.

(v) No homologue We identified ORFs without clear homology to any database entry, but still likely to code for protein since they are in compliance with the following criteria: they are long enough to be coding, do not overlap with other ORFs with assigned function, and have an aa

© 1997 by John Wiley & Sons, Ltd

composition and GC content typical for coding regions in yeast. YOR3170c (25 975–21 029; 1648 aa) has no significant hit against the protein databases but it shows an almost 100% identity with a human EST dbest-gln-4055. Other large ORFs without clear homologues are YOR3296c (893 aa) and YOR3329c (622 aa). An overview of ORFs larger than 100 aa and non-overlapping ORFs with clear homologues is shown in Table 1 and Figure 1.

Since the yeast genome sequence is now complete and several bacterial genome sequences are available, linking large-scale DNA sequencing with database searches and detailed case-to-case analysis is increasingly profitable. Data analysis of the type discussed here is an increasingly important bridge between the accumulation of raw sequence data and the planning of functional analysis experiments aiming at detailed elucidation of gene function.

ACKNOWLEDGEMENTS

DNA sequencing was supported by the European Union yeast genome sequencing programme. The support of the GENEQUIZ consortium and especially of Georg Casari are gratefully acknowledged. The protein design group of CNB-CSIC is financed by grant BIO94–1067 from CICYT, Spain.

REFERENCES

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. and Weng, J. (1987). In Allen, F. H., Bergerhoff, G. and Sievers, R. (Eds), *Crystallographic Databases* — *Information Content, Software Systems, Scientific Applications*. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.

- Allison, L. A., Moyle, M., Shales, M. and Ingles, C. J. (1985). Extensive homology among the largest subunits of eucaryotic and procaryotic RNA polymerases. *Cell* 42, 599–610.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
- Amor, J. C., Harrison, D. H., Klein, R. A. and Ringe, D. (1994). Structure of the human ADP-ribosylation factor 1 complexed with GDP. *Nature* 372, 704– 708.
- Ansorge, W., Voss, H., Wiemann, S., *et al.* (1992). High throughput automated DNA sequencing with fluorescent labels at the EMBL. *Electrophoresis* **13**, 616–619.
- Attree, O., Olivios, I. M., Okabe, I., *et al.* (1992). The Lowe's oculocerebrorenal syndrome gene encodes a protein highly homologous to inositol polyphosphate-5-phosphatase. *Nature* 358, 239–242.
- Bairoch, A. and Apweiler, R. (1996). The SWISS-PROT protein sequence databank and its new supplement TREMBL. Nucl. Acids Res. 24, 21–25.
- Bairoch, A. and Boeckmann, B. (1993) The SWISS-PROT protein sequence databank, recent developments. *Nucl. Acids Res.* 21, 3093–3096.
- Baker, R. T., Tobias, J. W. and Varshavsky, A. (1992). Ubiquitin-specific proteases of *Saccharomyces cerevisiae*. Cloning of UBP2 and UBP3, and functional analysis of the UBP gene family. *J. Biol. Chem.* 267, 23364–23375.
- Barrell, B., et al. (1994). Accession Numbers: L12980, L20215, L05146, L22015, L28920.
- Beltzer, J. P., Morris, S. R. and Kohlhaw, G. B. (1988). Yeast *LEU4* encodes mitochondrial and nonmitochondrial forms of alpha-isopropymalate synthase. J. Biol. Chem. 263, 368–374.
- Benson, D., Boguski, M., Lipman, D. J. and Ostell, J. (1996). GenBank. Nucl. Acids Res. 24, 1–5.
- Boguski, M. (1995). The turning point in genome research. *Trends Biochem. Sci.* 20, 295–296.
- Boguski, M. S., Lowe, T. M. J. and Tolstoshev, C. M. (1993). dbEST database for expressed sequence tags. *Nature Genetics* **4**, 332–333.
- Bork, P. and Sudol, M. (1994). The WW domain: a signalling site in distrophin? *Trends. Biochem. Sci.* **19**, 531–533.
- Bou, G., Esteban, P. F., Baladron, V., et al. (1993). The complete sequence of a 15 820 bp segment of *Saccharomyces cerevisiae* chromosome XI contains the *UBP12* and *MPL1* genes and three new open reading frames. *Yeast* 9, 1349–1354.
- Brewer, B. J. (1988). When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53**, 679–686.
- Bussey, H., et al. (1995). The nucleotide sequence of chromosome I from Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA 92, 3809–3813.
- © 1997 by John Wiley & Sons, Ltd

- Casari, G., Andrade, M. A., Bork, P., *et al.* (1995). Challenging times for bioinformatics. *Nature* **376**, 647–648.
- Chang, L. F., Gaztek, P. R. and Kohlhaw, G. B. (1985). Total deletion of yeast *LEU4*: further evidence for a second alpha-isopropyl malate synthase and evidence for tight *LEU4-MET4* linkage. *Gene* **33**, 333–339.
- Clark, J. D., Lin, L., Kriz, R. W., et al. (1991). A novel arachidonic acid-selective cytosolic PLA₂ contains a Ca²⁺-dependent translocation domain with homology to PKC and GAP. Cell 65, 1043–1051.
- Cleves, A. E., Novick, P. J. and Bankaitis, V. A. (1989). Mutations in the SAC1 gene suppress defects in yeast Golgi and yeast actin function. J. Cell. Biol. 109, 2939–2950.
- Davis, L. I. and Fink, G. R. (1990). The NUP1 gene encodes an essential component of the yeast nuclear pore complex. Cell 61, 965–978.
- Davis, R., et al. (1994). Accession Numbers: U18795, U18779, U18530, U18778, U18796, U18813, U18814, U18839, U18916,U18917, U18992.
- Dujon, B., et al. (1994). Complete DNA sequence of yeast chromosome XI. Nature 369, 371–378.
- Farabaugh, P. J. (1995). Post-transcriptional regulation of transposition by Ty retrotransposons of *Saccharomyces cerevisiae*. J. Biol. Chem. **270**, 10361–10364.
- Feldmann, H., et al. (1994). Complete DNA sequence of yeast chromosome II. EMBO J. 13, 5795–5809.
- Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783– 791.
- Garcia-Cantalejo, J., Baladron, V., Esteban, P. F., *et al.* (1994). The complete sequence of an 18 002 bp segment of *Saccharomyces cerevisiae* chromosome XI contains the *HSB1*, *MRP-L20* and *PRP16* genes and six new open reading frames. *Yeast* **10**, 231–245.
- George, D. G., Barker, W. C., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1996). The PIR-International protein sequence database. *Nucl. Acids Res.* 26, 17–20.
- Graf, R., Baum, B. and Braus, G. H. (1993). *YMC1*, a yeast gene encoding a new putative mitochondrial carrier protein. *Yeast* **9**, 301–305.
- Gribskov, M. and Devereux, J. (1991). Sequence Analysis Primer. Stockton Press, New York.
- Higgins, D. G., Bleasby, A. J. and Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.* 8, 189–191.
- Hong, J. X., Wilson, G. L., Fox, C. H. and Kehrl, K. H. (1993). Isolation and characterization of a novel B cell activation gene. J. Immunol. 150, 3895–3904.
- Hultman, T., Stahl, S., Hornes, E. and Uhlen, M. (1989). Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support. *Nucl. Acids Res.* 17, 4937–4946.
- Johnston, M., et al. (1994). Complete nucleotide sequence of Saccharomyces cerevisiae chromosome VIII. Science 265, 2078–2082.

130 kb FROM CHROMOSOME XV

- Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- Leahey, A. M., Charnas, L. R. and Nussbaum, R. L. (1993). Nonsense mutations in the OCRL-1 gene in patients with the oculocerebrorenal syndrome of Lowe. *Hum. Mol. Genet.* 2, 461–463.
- Magdolen, V., Lang, P., Mages, G., Hermann, H. and Bandlow, W. (1994). The gene *LEO1* on yeast chromosome XV encodes a non-essential, extremely hydrophilic protein. *Biochim. Biophys. Acta* **1218**, 205–209.
- Magdolen, V., Oechsner, U., Mueller, G. and Bandlow, W. (1988). The intron-containing gene for yeast profilin (PFY) encodes a vital function. *Mol. Cell. Biol.* 8, 5108–5115.
- Margolis, P. S., Driks, A. and Losick, R. (1993). Sporulation gene *spoIIB* from *Bacillus subtilis*. J. Bacteriol. 175, 528–540.
- Mariotinni, P., Bagni, C., Francesconi, A., Cecconi, F. and Serra, M. J. (1993). Sequence of the gene coding for ribosomal protein S8 of *Xenopus laevis. Gene* 132, 255–260.
- Mellor, J., Fulton, S. M., Dobson, M. J., Wilson, W., Kingsman, S. M. and Kingsman, A. J. (1985). A retrovirus-like strategy for expression of a fusion protein encoded by yeast transposon Ty1. *Nature* 313, 243–246.
- Murakami, Y., et al. (1995). Analysis of the nucleotide sequence of chromosome VI from *Saccharomyces cerevisiae*. *Nature Genetics* **10**, 261–268.
- Oechsner, U., Magdolen, V. and Bandlow, W. (1988). A nuclear yeast gene (GCY) encodes a polypeptide with high homology to a vertebrate eye lens protein. *FEBS Lett.* **238**, 123–128.
- Oliver, S. G., *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46.
- Ouzounis, C., Casari, G., Valencia, A. and Sander, C. (1996). Novelties from the complete genome of *Mycoplasma genitalium*. Mol. Microbiol., **20**, 898–900.
- Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.
- Perentesis, J. P., Phan, L. D., Gleason, W. B., LaPorte, D. C., Livingston, D. M. and Bodley, J. W. (1992). *Saccharomyces cerevisiae* elongation factor 2. Genetic cloning, characterization of expression, and Gdomain modeling. J. Biol. Chem. 267, 1190–1197.
- Powers, S., Kataoka, T., Fasano, O., Goldfarb, M., Strathern, J., Broach, J. and Wigler, M. (1984). Genes in *S. cerevisiae* encoding proteins with domains homologous to the mammalian ras proteins. *Cell* 36, 607–612.
- Rodriguez-Tome, P., Stoher, P. J., Cameron, G. N. and Flores, T. P. (1996). The European Bioinformatics Institute databases. *Nucl. Acids Res.* 24, 6–12.
- Rost, B., Sander, C. and Schneider, R. (1994). PHD an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.* 10, 53–60.
- © 1997 by John Wiley & Sons, Ltd

- Rost, B. and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19**, 55–72.
- Saitou, N. and Nei, M. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.
- Sander, C. and Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* 9, 56–68.
- Scharf, M., Schneider, R., Casari, G., et al. (1994). Intelligent Systems for Molecular Biology. AAAI Press, pp. 348–353.
- Schwager, C., Wiemann, S. and Ansorge, W. (1995). GeneSkipper: integrated software environment for DNA sequence assembly and alignment. *Genome Digest* 2, 8–9.
- Siderovski, D. P., Heximer, S. P. and Forsdyke, D. R. (1994). A human gene encoding a putative basic helix-loop-helix phosphoprotein whose mRNA increases rapidly in cycloheximide-treated blood mononuclear cells. *DNA Cell Biol.* 13, 125–147.
- Silberstein, S., Collins, P. G., Kelleher, D. J. and Gilmore, R. (1995). The essential OST2 gene encodes the 16-kD subunit of the yeast oligosaccharyl transferase, a highly conserved protein expressed in diverse eukaryotic organisms. J. Cell. Biol. 131, 371–383.
- Sossin, W. S. and Schwartz, J. H. (1993). Ca^{2+} independent protein kinase Cs contains an aminoterminal domain similar to the C2 consensus sequence. *Trends Biochem. Sci.* **18**, 207–208.
- Spieth, J., Brooke, G., Kuersten, S., Lea, K. and Blumenthal, T. (1993). Operons in *C. elegans*: polycistronic mRNA precursors are processed by transsplicing of SL2 to downstream coding regions. *Cell* 73, 521–532.
- Sutton, R. B., Davletov, B. A., Berghuis, A. M., Sudhof, T. C. and Sprang, S. R. (1995). Structure of the first C2 domain of synaptotagmin I: a novel Ca²⁺/ phospholipid binding fold. *Cell* 80, 929–938.
- Suzuki, K., Olvera, J. and Wool, I. G. (1990). The primary structure of rat ribosomal protein S7. *FEBS Lett.* 271, 51–53.
- Thierry, A., Gaillon, L., Galibert, F. and Dujon, B. (1995). Construction of a complete genomic library of *Saccharomyces cerevisiae* and physical mapping of chromosome XI at 3.7 kb resolution. *Yeast* **11**, 121–135.
- Valencia, A. and Sander, C. (1995). In Zerial, M., Huber, L. A. (Eds), *The ras Superfamily, A Practical Handbook*. 12–20.
- Voss, H., Wiemann, S., Wirkner, U., et al. (1992). Automated DNA sequencing system resolving 1000 bases with fluorescein-15-*dATP as internal label. *Meth. Mol. Cell. Biol.* 3, 153–155.
- Voss, H., Tamames, J., Teodoru, C., *et al.* (1995). Nucleotide sequence and analysis of the centromeric region of yeast chromosome IX. *Yeast* 11, 61–78.

- Weber, K. and Kabsch, W. (1994). Intron positions in actin genes seem unrelated to the secondary structure of the protein. *EMBO J.* 13, 1280–1286.Zimmermann, J., Dietrich, T., Voss, H., *et al.* (1992).
- Zimmermann, J., Dietrich, T., Voss, H., et al. (1992). Fully automated Sanger sequencing protocol for double stranded DNA. Meth. Mol. Cell. Biol. 3, 39–42.
- Zimmermann, J., Wiemann, S., Voss, H., Schwager, C. and Ansorge, W. (1994). Improved fluorescent cycle sequencing protocol allows reading nearly to 1000 bases. *BioTechniques* **17**, 302–307.