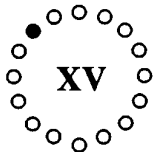
YEAST VOL. 13: 655–672 (1997)



Yeast Sequencing Reports

DNA Sequencing and Analysis of 130 kb from Yeast Chromosome XV

HARTMUT VOSS[1]*, VLADIMIR BENES[1], MIGUEL A. ANDRADE[2,3], ALFONSO VALENCIA[3], STEFANIE RECHMANN[1], CRISTINA TEODORU[1], CHRISTIAN SCHWAGER[1], VACLAV PACES[4], CHRIS SANDER[2] AND WILHELM ANSORGE[1]

[1]*Biochemical Instrumentation Programme, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany*
[2]*Biological Structures and Biocomputing Programme, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany*
[3]*Protein Design Group, CNB-CSIC, Cantoblanco, Madrid 28049, Spain*
[4]*Institute 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 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:

*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

(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 SwissProt; George *et al.*, 1996), GENPEPT (a direct translation of the DNA sequences in GenBank; Benson *et al.*, 1993), TREMBL (Bairoch and Apweiler, 1996); DNA 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

© 1997 by John Wiley & Sons, Ltd

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 2·2 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 human-machine 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.

© 1997 by John Wiley & Sons, Ltd

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

Name	From	To	aa	Identity	Protein/DNA	Description	aa	Score	Features of the ORFs
YOR2964c	465	2714	749	Similar	YK69_YEAST	Hypothetical protein	910	2·7e-247	
YOR3116w	3059	3946	295	Similar	YK71_YEAST	Hypothetical protein	152	1·4e-4	
YOR3120w	4113	5276	387	Similar	Mm0361_1	Lipase-esterase operon product	264	1·1e-1	LIPASE_SER PROSITE
YOR3124w	5559	6611	350	Identical	OSTG_YEAST	Oligosaccharyl transferase γ precursor			
ARS-cons	6679	6689							
ARS-cons	6704	6714							
YOR3141c	6745	10305	1186	Similar	SYT1_CAEEL	Synaptotagmin I	441	3·5e-7	Three C2 domains
tRNA	10965	11038				tRNA-Asn			
YOR3151w	11811	13259	482	Similar	TRP_DROME	Transient receptor potential protein	1275	2·3e-7	Transmemb+coiled-coil
ARS-cons	12048	12058							
YOR3154c	13721	14353	210	Identical	YP51_YEAST	Gtp-Binding Protein YPT_51			
YOR3157c	14648	16366	572	Similar	PDP_BOVIN	Pyr DH (lipoamide)-phosphatase precursor	538	7·0e-19	Protein phosphatase 2C signature
YOR3160w	16789	17994	401	No homologue					contains ORF YOR3162c
YOR3162c	16944	17924	326	Similar	dbest-gnl-73646	*A. thaliana* gene product		n2·5e-7	Leucine_Zipper PROSITE
YOR3165w	18651	20114	487	Similar	Sctrnaorf_2	*S. cerevisiae* ORF	642	n1·7e-177	ATPase α-β
YOR3170c	21029	25975	1648	Similar	dbest-gnl-4055	Human EST		n1·2e-54	
YOR3172w	26318	26869	183	Similar	ARF6_CHICK	ADP-ribosylation factor 6	175	7·2e-66	ADP-ribosylation factors signature
YOR3174c	27075	27851	258	Similar	RPIA_ECOLI	Ribose 5-phosphate isomerase A	219	2·9e-12	
YOR3177w	29317	30290	190	Similar	Sctrnaorf_1	Similar to ribosomal S7	190	6·9e-113	Ribosomal S7e blocks+one intron
YOR3180c	30501	31028	175	No homologue					
YOR3182c	31471	34701	1076	Identical	NUP1_YEAST	Nucleoporin NUP1			
ARS-cons	34926	34936							
YOR3189w	35348	36529	393	Identical	KTR1_YEAST	Probable mannosyltransferase Ktr1			
ARS-cons	36632	36642							
YOR3193c	36818	37801	327	Similar	YMC1_YEAST	Mitochondrial carrier protein Ymc1	307	4·4e-13	Mitochondrial energy carrier signature
YOR3205w	38767	39696	309	Identical	RAS1_YEAST	Ras-like protein 1			
YOR3211c	39972	40373	130	Identical	OSTE_YEAST	Oligosaccharyltransferase 16 kDa subunit			
ARS-cons	40896	40906							
YOR3214w	41165	42013	282	No homologue					
YOR3220w	42644	43495	283	Similar	PE12_YEAST	Vacuolar proteases sorting	288	4·4e-9	PROSITE of epimorphines
YOR3224w	44876	45805	309	Similar	YKO7_CAEEL	Hypothetical protein	221	9·4e-2	
YOR3227w	46550	48238	562	Similar	LEU1_YEAST	2-Isopropylmalate synthase	619	0·0	Aipm_Homocit_Synth 1 & 2 patterns
YOR3231w	48799	52122	1107	Similar	RSD1_YEAST	Recessive suppressor of secretory defect	623	1·1e-50	
YOR3234w	52462	53769	435	No homologue					
YOR3237w	53950	54648	232	Similar	MAF_BACSU	Hypothetical protein	189	7·0e-11	
YOR3240w	55029	57314	761	Similar	Cew07a12_5	*C. elegans* product	1183	3·7e-18	AA transfer class PROSITE
YOR3244w	57596	60340	914	Identical	Z26253	*S. cerevisiae AZF1* gene for zinc finger protein			
YOR3248w	61091	61975	294	No homologue					
YOR3251c	62180	62986	268	Similar	T38532	*S. cerevisiae* EST		n1·3e-48	
YOR3254c	63284	67666	1460	Identical	RPC1_YEAST	DNA-directed *RNA* polymerase III			
YOR3258w	68550	69854	434	Identical	TBP1_YEAST	Tat-binding homolog 1			

© 1997 by John Wiley & Sons, Ltd

Table 1. *Continued*

Name	From	To	aa	Identity	Protein/DNA	Description	aa	Score	Features of the ORFs
YOR3263w	70378	72081	567	No homologue					
YOR3266c	72313	73767	484	Similar	YOT3_CAEEL	Hypothetical protein	510	1·5e-32	
YOR3269w	74635	75573	312	Identical	GCY_YEAST	GCY protein of unknown function			
YOR3275c	75819	76408	126	Identical	PROF_YEAST	Profilin prevents the polymerization of actin			One intron
YOR3278c	76697	78091	464	Identical	LEO1_YEAST	Unknown function			
YOR3281c	78345	82163	1272	Identical	UBP2_YEAST	Ubiquitin carboxyl-terminal hydrolase 2			
YOR3284c	82551	83369	272	Similar	U13642	*C. elegans* gene product		n2·9e-26	
YOR3287c	83482	84198	238	Similar	M94674	*C. albicans* α-glucosidase (maltase) mRNA (non-translated)		n1·2e-13	
YOR3290w	84691	87714	1007	Identical	SC07421	*S. cerevisiae* S288C rho-type GTPase activating protein			
YOR3293c	87997	89712	571	Identical	PUR6_YEAST	P-Ribosylaminoimidazole carboxidase catalytic subunit			
YOR3296c	90398	93079	893	No homologue					
YOR3299c	93450	94328	292	Similar	YMC1_YEAST	Mitochondrial carrier protein YMC1	307	3·2e-12	Mitochondrial energy carrier signature
δ	95011	95328							
ARS-cons	95163	95173							
tRNA	95479	95550				tRNA-Asp			
YOR3311c	95703	96359	218	Similar	YHFE_ECOLI	Hypothetical protein	252	3·2e-8	
YOR3314w	96696	98351	551	Identical	VP17_YEAST	Vacuolar protein sorting-associated protein VPS17			
YOR3317w	98619	101147	842	Identical	EF2_YEAST	Elongation factor 2			
YOR3320w	102085	103314	409	Similar	SAC7_YEAST	SAC7 protein involved in assembly/function of actin	274	8·6e-18	Probable rho/racGAP domain
YOR3326w	103771	104880	369	Identical	IDH2_YEAST	IDH Mitochondrial sub 2 precursor			
YOR3329c	105334	107202	622	Similar	Scl8093_3	Yeast chromosome XII cosmid	578	3·1e-24	Leucine_Zipper PROSITE
YOR3332c	107830	109845	671	No homologue					
YOR3339w	110502	112802	766	Identical	SFL1_YEAST	Flocculation suppression protein			
YOR3348c	113693	116107	804	Similar	ACT2_YEAST	Actin-like	391	3·0e-6	Actin proteins block, introns(?)
YOR3352w	116577	117566	329	Similar	SUCA_RAT	Succinyl CoA ligase	333	2·8e-124	
tRNA	117875	117945				tRNA-Gly-sup			
δ-remnant	118033	118302							
δ	118341	118672							
YOR3367w	118632	123900	1755	Similar	Sc8229_23	Transposon peptide	1755	0·0	Frameshift (by homology)
δ	123923	124254							
YOR3373c	124903	125862	319	Identical	TH80_YEAST	Thiamin pyrophosphokinase			
YOR3510c	126237	128612	791	Identical	SC0612	EST			
YOR3513c	128867	129523	218	Similar	D28195	Translated cDNA (rice)		n8·0e-29	Fragment

The protein/DNA column indicates the closest homologue in the database. Unless stated otherwise, only SwissProt and TREMBL protein sequence identifiers are used throughout the paper. SwissProt identifiers are presented with two words in capital letters joined by an underscore, the second referring to species. TREMBL is a database of protein translation product from the EMBL DNA database. The TREMBL identifiers are composed of the corresponding EMBL identifier followed by an underscore and a number that indicates the order of the translation product (since many consecutive translation products are frequently reported from the same EMBL entry). The EMBL accession number is given for DNA, one or two capital letters followed by a number. The score indicates the degree of homology, BLAST scores are listed when the closest homologue is a protein sequence, BLASTX scores in case of a nucleotide sequence (indicated by an 'n').

© 1997 by John Wiley & Sons, Ltd

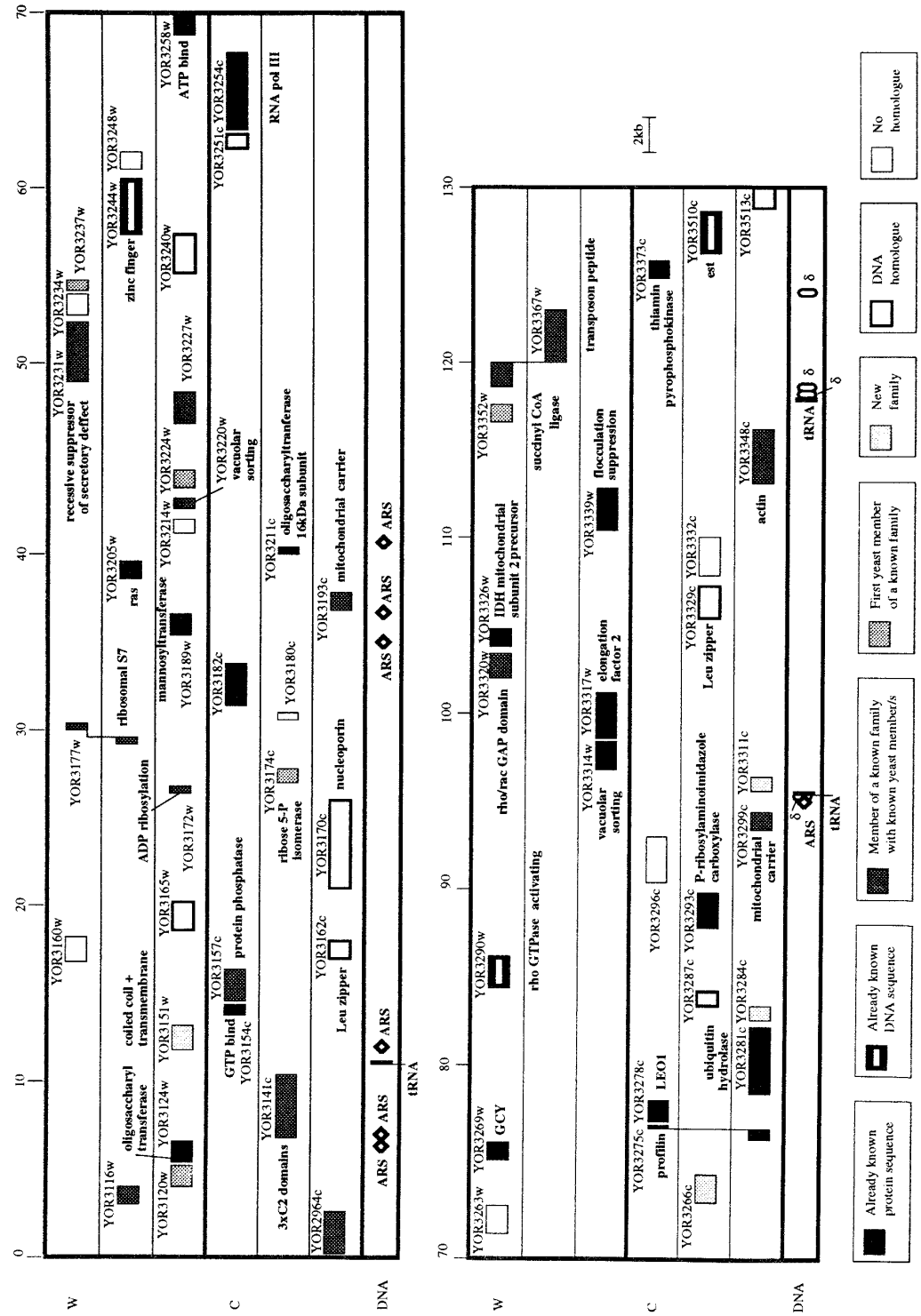


Figure 1. Protein and DNA features of accession no. X94335. Each box represents an ORF. Patterns and line thickness indicate the homology and functional characteristics of the corresponding ORF as stated in the legend below the figure. For each ORF the identifier and a short feature description (if known) are given. Note the striking accumulation of Watson ORFs from position 41 000 to 62 000.

© 1997 by John Wiley & Sons, Ltd

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), *LEO1* (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

© 1997 by John Wiley & Sons, Ltd

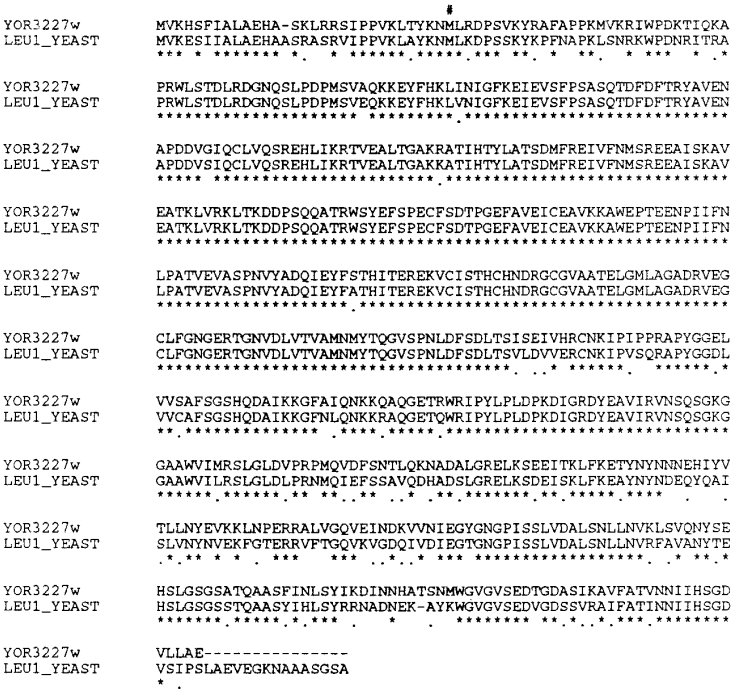


Figure 2. Case 1. Alignment of the sequence YOR3227w with LEU1_YEAST. (Note: there is an annotation conflict; the translation product of the *LEU4* gene is annotated as LEU1_YEAST in SwissProt.) The homology between the two sequences is around 80%. The position of the second Met (labelled with #) is a possible second initiation site conserved in both sequences. The sequence of LEU1_YEAST is longer (619 aa).

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.

© 1997 by John Wiley & Sons, Ltd

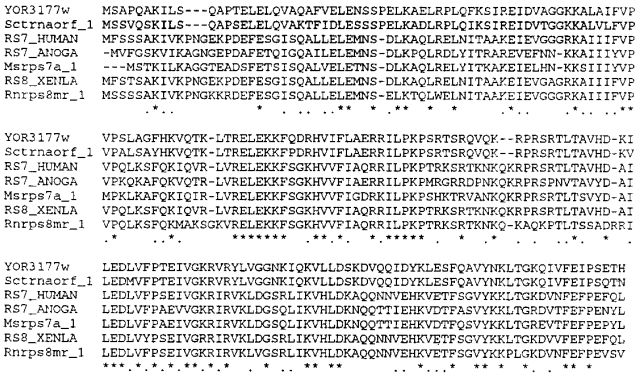


Figure 3. Case 2. Multiple alignment of the S7, 40S ribosomal subunit protein family. The alignment includes YOR3177w, its homologue Sctrnaorf_1, a translation product from yeast (X85811), and examples from human (RS7 Human), *Xenopus laevis* (RS8_XENLA), *A. gambiae* (RS7_ANOGA), a ribosomal S7 protein from an insect *M. sexta* (Msrps7a 1) and rat (Rnrps8mr_1). The sequences labelled 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 contains other yeast sequences Case 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 OCRL-inositol polyphosphate-5-phosphatase family (Voss *et al.*, 1995). Inositol polyphosphate-5-phosphatases catalyse the conversion of inositol-trisphosphate 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

© 1997 by John Wiley & Sons, Ltd

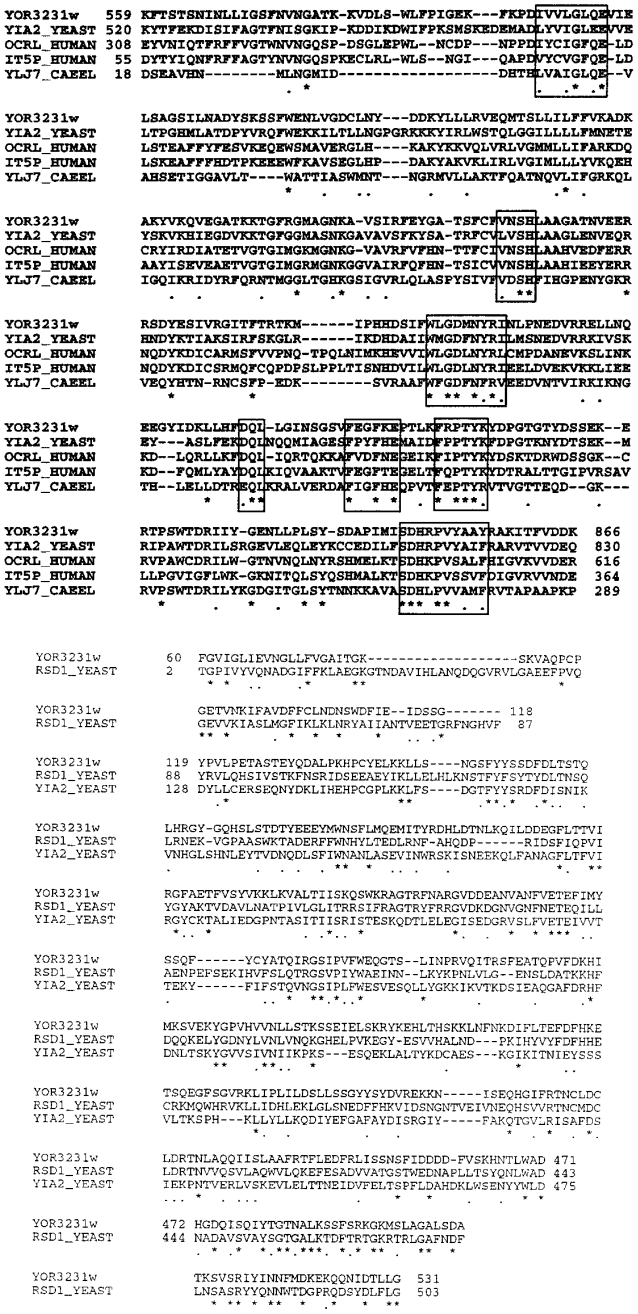


Figure 4. Case 5. YOR3231w. Alignment of sequences: (a) C-terminal against OCRL_HUMAN and its homologues and (b) N-terminal against RSD1_YEAST and YIA2_YEAST.

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

© 1997 by John Wiley & Sons, Ltd

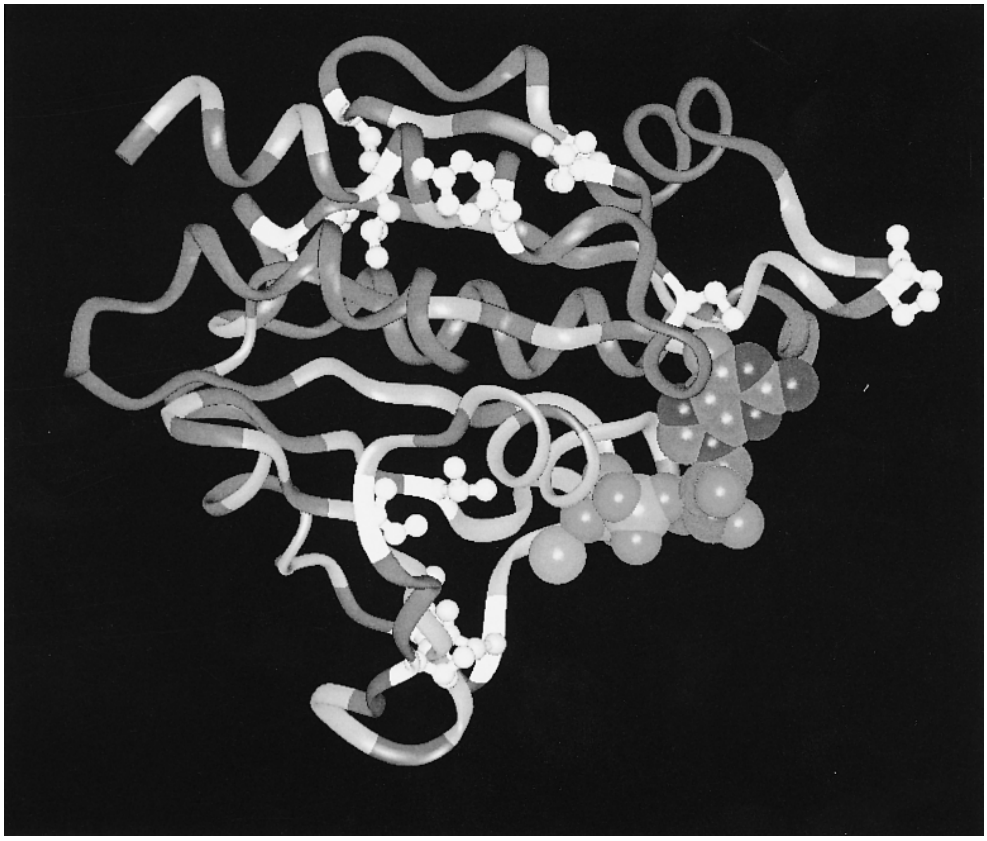


Figure 5. Case 7. The conservation levels in the arf family are mapped in the three-dimensional structure of the human ADP-ribosylation factor 1 complexed with GDP. The backbone is coloured in orange for those residues conserved in the whole arf family. White colour and side chains indicate residues 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 Mg^{2+} ions (Ser-Thr exchange in position 45).

Case 8: YOR3141c (10 305–6745; 1186 aa) contains three C2 domains. C2 domains, probably involved in Ca^{2+} and phospholipid binding, have been described in different protein families such as Ca^{2+}-dependent protein kinases C (Clark *et al.*, 1991); synaptotagmines, which are related to synaptic 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 β2–β8). The phylogenetic analysis of

© 1997 by John Wiley & Sons, Ltd

```
YOR3141C/1  385 IGILEITVKNAKGLKRTSS-ILNESIDPYLSFEFND----ISIAKTRTVRD-TLNPVWDETLYVLL--
YOR3141C/2  657 IGAIRVFIEKANDLRNLE---KFGTIDPYCKVLVNG----LSKGRTDFKSQ-TLNPVWNQVIYVAVT-
YOR3141C/3  991 SGDLTIMSRSAENLIASD---LNGYSDPYLKYYINNEED--CAYKTKVVKK-TLNPKWNDEGTIQIN-
RSP5_YEAST    2 PSSISVKLVAAESLYKRD---VFRSPDPFAVLTIDG----YQTKSTSAAKK-TLNPYWNETFKFD---
SYT1_RAT/1  155 NNQLLVGIIQAAELPALD---MGGTSDPYVKVFLLPDKK--KKFETKVHRK-TLNPVFNEQFTFKVP-
SYT1_RAT/2  285 AGKLTVVILEAKNLKKMD---VGGLSDPYVKIHLMQNGKRLKKKKTTIKKN-TLNPYYNESFSFEVP-
PIPA_DICDI  673 YSRLIVNVISARQLPKYTKSTKGEVIDPYVTLSIVGTHFDQKVEKTKVIDNNGFNPHWGEEFEFPLYN
UN13_CAEEL  736 SAKITLTVLCAQGLIAKD---KTGKSDPYVTAQVGK-----TKRRTRTIHQ-ELNPVWNEKFHFECH-
A42142      275 HGRFVGVTIKVPACVDLAK--KQGTCDPFVVCTAHYSNKHQVTRRTKQRKK-TVDPEFDEAMYFDLHI
KPC2_HUMAN  170 RDVLIVLVRDAKNLVPMD---PNGLSDPYVKLKLIPDPKSESKQKTKTIKC-SLNPEWNETFRFQLK-

YOR3141C/1      --NSFTDP-LTISVYDKR---------AKLK--DKVLGRIQYN-----------LNTLHDKT  481
YOR3141C/2      --SPNQR--ITLQCMDVET---------VNK--DRSLGEFNVNVQDLFKKD--ENDKYEETI  760
YOR3141C/3      --NRLNDVL-RIKVMDWD---------STSA--DDTIGTAEIPLNKVKVEGTTELDVPVEGL 1099
RSP5_YEAST      DINENSI--LTIQVFDQK---------KFKKKDDGFLGVVNVRVG--------DVLGHLDED  101
SYT1_RAT/1      -YSELGGKTLVMAVYDFD---------RFSK--HDIIGEFKVPMNTVDFG--HVTEEWRDLQ  263
SYT1_RAT/2      -FEQIQKVQVVVTVLDYD---------KIGK--NDAIDKVFVGYNSTG----AELRHWSDML  394
PIPA_DICDI      --SQLSM--LLIRVDDKD---------KVGH---NRIGHHCIRVENIRPGYR--ILKLKNNF  784
UN13_CAEEL      --NSTDR--IKVRVWDEDNDLKSKLRQKLTRESDDFLGQTVIEVRTLSG----EMDVWYNLE  847
A42142          DADAGST---NTTGSNKS-------AGSLES--SANKGYSIYPVGGADLVE-IVVSVWHDAH  388
KPC2_HUMAN      --ESDKDRRLSVEIWDWD---------LTSR--NDFMGSLSFGISELQKA---SVDGWFKLL  278
```

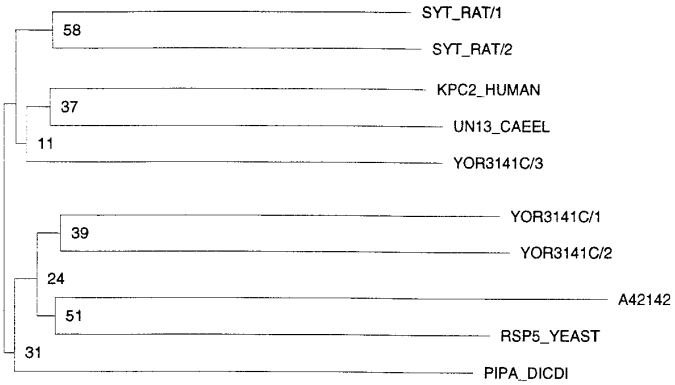


Figure 6. (a) Case 8. Alignment of the three C2 domains from YOR3141c with the C2 domains of other representative proteins. The alignment extends around 120 aa. The proteins are indicated by their SwissProt identifiers: KPC2_HUMAN, protein kinase C; PIPA_DICDI, phosphatidyl inositol phosphodiesterase; 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 β7 is pointed out by the fact that it is mutated into a 'D' in the second C2 domain of synaptotagmin that is not functional. (b) Case 8. Tree of the previous alignment 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 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

© 1997 by John Wiley & Sons, Ltd

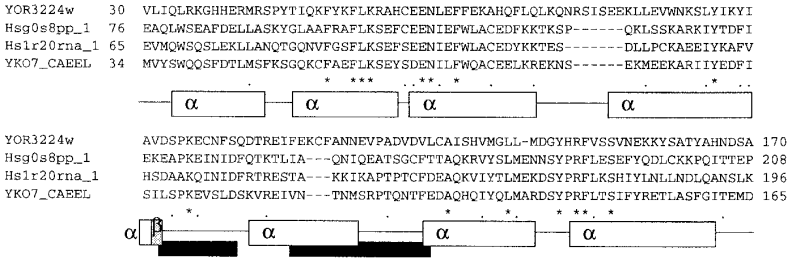


Figure 7. Case 10. YOR3224w is similar to two ORFs from human Hsg0s8pp_1 (Hong *et al.*, 1993) and Hs1r20rna_1 (Siderovski *et al.*, 1994), and to one ORF from *C. elegans* (YK07_CAEEL). Multiple sequence alignment with secondary structure prediction using PHD (Rost, 1994). Completely conserved residues 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-loop-helix (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

© 1997 by John Wiley & Sons, Ltd

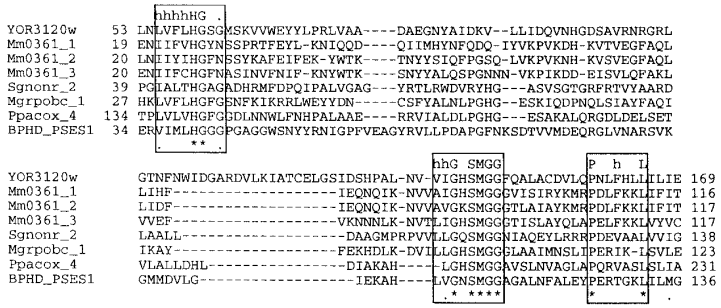


Figure 8. Case 11. YOR3120w. Alignment against products of the lipase-esterase operon from Mycoplasm (Mm0361_1, Mm0361_2 and Mm0361_3), Sgnonr_2 (antibiotic-resistance protein, 279 aa, *Streptomyces griseus*), Mgrpobc_1 (unknown product, Mycoplasm), Ppacox_4 (dihydrolipoamide acetyltransferase, Pseudomonas)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

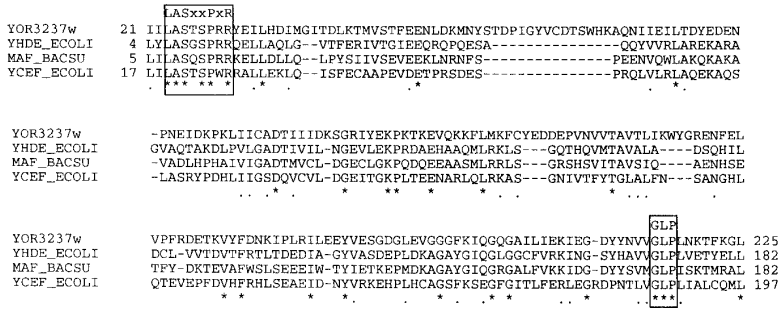


Figure 9. Case 12. YOR3237w is similar to *B. subtilis* and *E. coli* proteins with possible roles in the cell cycle. Multiple sequence alignment with main motifs boxed. Asterisks indicate residues conserved throughout the whole family; dots mark similar residues.

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

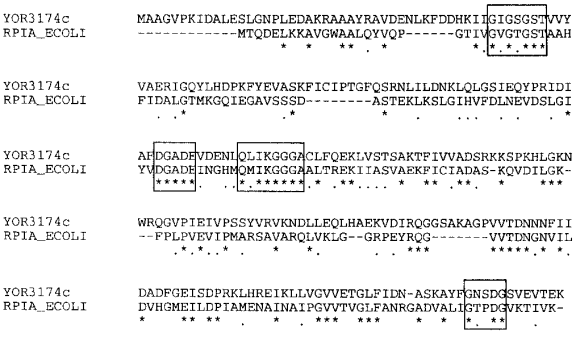


Figure 10. Case 13. Full sequence alignment of YOR3174c to RPIA_ECOLI. Highly conserved motifs are boxed, conserved residues are marked by an (*), similar residues are indicated as a (.).

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

© 1997 by John Wiley & Sons, Ltd

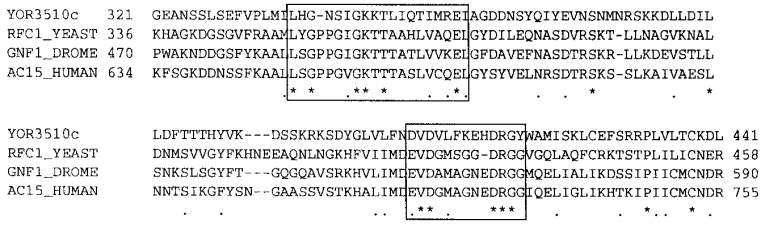


Figure 11. Case 14. Multiple alignment of YOR3510c with DNA binding proteins GNF1_DROME, AC15_MOUSE, AC15_HUMAN and RFC1_YEAST. Homologous profiles are boxed, conserved residues are marked by a (*), similar residues are indicated as a (.).

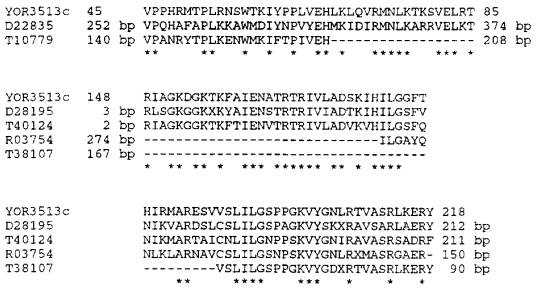


Figure 12. Case 15. Alignment of ORF YOR3513c against several ESTs (see text for details). Asterisks indicate positions where all the fragments show identical residues. No gaps were allowed for the alignment.

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 LxGxxxhGKxTxxxhxxEh or aVDxhxxxxDRG, where × is any amino acid, h 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 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.

© 1997 by John Wiley & Sons, Ltd

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 non-mitochondrial 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.

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 PLA_2 contains a Ca^{2+}-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.

© 1997 by John Wiley & Sons, Ltd

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 G-domain 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.

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 amino-terminal 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 trans-splicing 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 Ca^{2+}/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.

© 1997 by John Wiley & Sons, Ltd

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). 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.

© 1997 by John Wiley & Sons, Ltd