The EMBO Journal vol.13 no.3 pp.493-503, 1994

Yeast chromosome III: new gene functions

Eugene V.Koonin[1], Peer Bork[2,3] and Chris Sander[2]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [2]European Molecular Biology Laboratory, D-69012 Heidelberg, Germany and [3]Max Delbrück Center for Molecular Medicine, D-13125 Berlin, Germany

Communicated by I.W.Mattaj

One year after the release of the sequence of yeast chromosome III, we have re-examined its open reading frames (ORFs) by computer methods. More than 61% of the 171 probable gene products have significant sequence similarities in the current databases; as many as 54% have already known functions or are related to functionally characterized proteins, allowing partial prediction of protein function, 11 percentage points more than reported a year ago; 19% are similar to proteins of known three-dimensional structure, allowing model building by homology. The most interesting new identifications include a sugar kinase distantly related to ribokinases, a phosphatidyl serine synthetase, a putative transcription regulator, a flavodoxin-like protein, and a zinc finger protein belonging to a distinct subfamily. Several ORFs have similarities to uncharacterized proteins, resulting in new families 'in search of a function'. About 54% of ORFs match sequences from other phyla, including numerous fragments in the database of expressed sequence tags (ESTs). Most significant similarities to ESTs are with proteins in conserved families widely represented in the databases. About 30% of ORFs contain one or more predicted transmembrane segments. The increase in the power of functional and structural prediction comes from improvements in sequence analysis and from richer databases and is expected to facilitate substantially the experimental effort in characterizing the function of new gene products.

Key words: computer methods/genome analysis/prediction of protein function/prediction of protein structure/protein sequence analysis

Introduction

The 315 kb sequence of yeast chromosome III (Oliver *et al.*, 1992) is the first almost completely sequenced eukaryotic chromosome. It is a rich source of information and an ideal test case for computer-assisted sequence analysis. A key goal of this analysis is the prediction of protein function via similarity searches in sequence and structure databases.

In the original sequence report (Oliver *et al.*, 1992), protein sequence analysis was limited to a first step with stringent criteria [182 ORFs longer than 100 amino acid residues; FASTA (Pearson and Lipman, 1988) scores of ≥ 200]. As a result, significant similarities were reported for 47 ORFs, or 26% of the total. A more extensive search for functional similarities with more permissive cut-offs, in combination with methods for multiple alignment and motif search (Bork *et al.*, 1992a,b), raised this fraction to 42% and revealed 16% similarities to proteins of known three-dimensional structure.

Now, a year after the initial analyses, we have attempted to raise the level of functional assignment, in preparation for the expected sequences of other yeast chromosomes (Goffeau *et al.*, 1993a). The progress reported here relies on improvements in similarity matrices for sequence comparison (Henikoff and Henikoff, 1992, 1993), and on more detailed analysis of apparently weak similarities, and to some extent on richer information in DNA and protein databases.

Results and discussion

From ORFs to protein function: similarity search in databases

Currently, 171 ORFs can be considered probable expressed genes (Table I). Of these (Table II, Figure 1), 61% have significant amino acid sequence similarities to proteins in the databases: 11% are similar to functionally uncharacterized proteins, while, remarkably, as many as 50% are similar to proteins of known function (61% = 11% + 50%). Thus 54% of ORFs now have at least some aspect of their function identified: 50% via sequence similarities and 4% by direct experiment.

The power of functional prediction from sequence similarity is varied. In cases of strong similarity to proteins about which much is known, the precise biological function, the intracellular location and/or the role in the respective pathway can be carried over by analogy. Unfortunately, complete functional prediction is currently limited to a small fraction of well characterized enzymes. Nevertheless, almost all of the sequence similarities detected on chromosome III are at least indicative of the general functional class, e.g. 'DNA binding via a zinc finger' or 'G-domain molecular switch'. This dramatically narrows the range of possible functions and provides the starting point for further experimental characterization. Even the prediction of transmembrane regions without further specification of function gives a first idea of the likely type of function (Table I). Below, we give details about a few of the new functional assignments based on sequence similarities. A complete list is given in Table I.

From ORFs to genes: re-evaluation of gene assignments

The reduction in the number of predicted genes from 182 to 171 is due to the fact that nine ORFs (YCL2c, YCL21w, YCL22c, YCL23c, YCL26c, YCL41c, YCL42w,

Table I. Revised gene product assignments in the yeast chromosome III

ORF/gene[1]	Size (aa)	Known or predicted protein function/activity[2]	Closest similarity in the databases[3]	3D or tm	Prosite or other motif(s)	Phylum hits
YCL75w[a,b,4]	146	aspartic protease	PIR: S15611 (*Physarum* copia-like retrotransposon) −2.8e-5	3D	Asp protease	EY
YCL74w[c,4]	308	reverse transcriptase	PIR: S05465 (*Arabidopsis* copia-like retrotransposon) −4.6e-30	3D		EY
YCL70c[b,d,5]	373	membrane transporter	PIR: B40046: 1.6e-2; 1.3e-5	6 tms		BEY
YCL69w[a,b] (555)[6]	458	membrane transporter	GB: YSCSGE1A_1 −1.3e-36 EMRB_ECOLI −3.8e-9	10 tms		BEY
YCL68c[c]	190	**bud site selection**	BUD5_YEAST (YCR38c)- 98%/190			N
YCL67c[c]/MAT-α2	210	**mating type protein**	PIR: S19010 (homeotic protein) −2.5e-5	1HDD	PS00027 (homeobox)	EY
YCL66w[c]/MAT-α1	175	**mating type protein**	none			Y
YCL64c[c]/CHA1	360	**serine/threonine dehydratase**	STDL_YEAST −52%/320	3D	PS00165 pyridoxal-phosphate bind	BEY
YCL63w[b]	128	unknown	GB: YSCGCDA_1 (translation regulator) −2.9e-4; 2.7e-4			Y
YCL61c[b]	329	unknown	YURK_YEAST −90%/170			Y
YCL59c[b]	316	unknown	GB: CEL12C5 −9.6e-23			E
YCL57w[c]	712	metalloproteinase	MEPD_RAT −43%/102		PS00142 (Zn bind)	BE
YCL54w[a,b]	724	methyltransferase?	FTSJ_ECOLI (cell division protein)[7]- 9.1e-18		SAM bind	B
YCL50c[c]/APA1 (DTP)	321	**Ap_4A phosphorylase**	APA2_YEAST −75%/134			Y
YCL48w[c]	463	unknown	SPS2_YEASY (sporulation-specific protein) −56%/239	1 tms		Y
YCL45c[b]	760	unknown	GB T00254 (EST)-4.1e-3; 4.7e-6	2 tms		E
YCL43c[c]/PDI1 (MFP1)	522	protein disulfide isomerase	GB:YSCEUG_1 −48%/231	2TRX	PS00194 (thioredoxin)	BEY
YCL40w[c]/GLK1	500	**glucokinase**	HXKB_YEAST −4.6e-38	2YHX	hexokinase (PS00378)	BEY
YCL39w[b]	759	**regulatory protein of the β-transducin family**	PIR: S11169 (TUP1)-1.6e-7		PS00678 (TPR repeat)	EY
YCL38c[b]	528	membrane transporter	EMRB_ECOLI −6.4e-2; 1.2e-3	6 tms		BEY
YCL37c[b]	466	unknown	EMB D15392 (EST related to LA antigen)[8] −1.5e-6			E
YCL35c[c]	110	**glutaredoxin (thiol-transferase)**	GLRX_YEAST −64%/106	1EGO	PS00195	BEY
YCL33c[a]	168	transcription regulator	PILB_NEIGO −7.4e-13			B
YCL32w[b]/STE50	346	**essential for conjugation**	PIR: S22600 (leucine zipper protein) −7.7e-7			Y
YCL30c[c]/HIS4	799	**histidinol dehydrogenase**	PIR: S21517 −78%/279		PS00611	BEY
YCL29c[b,c]/BIK1	440	**nuclear fusion**; microtubule binding(?)	GB: HUMCLIP1_1 −1.8e-8			E
YCL27w[c]/FUS1	512	**cell fusion**	none			N
YCL27c-a	193[9]	unknown	GB:SCCHRIII_59 (YCLX8c) −8.6e-30 GB:HSA36F062 (EST) −8.1e-5			EY
YCLX8c	192	unknown	YCL27c-a −2.7e-29 GB:HSA36F062 (EST)[10] −5.6e-2; 1.6e-5			EY
YCL25c[c]	633	amino acid permease	GAP1_YEAST −50%/319	10 tms		BEY
YCL24w[c,11]	816	protein kinase	GB: ATHAKIN100A_1 −2.3e-60	2CPK	PS00107; 00108	EY
YCL20w[a]	437	GAG polyprotein	PIR: A22999 −58%/158			EY
YCL19w[a]	1347	POL polyprotein	PIR: B28097 −85%/669			
YCL18w[c]/LEU2	364	β-isopropyl-malate dehydrogenase	LEU3_KLULA −85%/357	9ICD	PS00470	BEY
YCL17c[c,a]/NFS1	(SPL1)497	aminotransferase(?)[12]	PIR: S02507 (bacterial nitrogen fixation protein NifS) −1.0e-42	3D	pyridoxal-phosphate bind	B
YCL11c[c]	427	poly(A) binding protein	PIR: A39720 −4.2e-13		RNA bind	EY
YCL9c[a]	309	regulatory protein	ILVH_ECOLI (acetolactate synthase, small regulatory subunit) −7.9e-17			B
YCL8c[b]	119	ubiquitin-conjugating enzyme? (inactive ?)	GB: ATTS0266_1 −2.6e-2; 5.6e-3	1AAK		EY
YCL4w (+YCL3w)[b,13]/PEL1?	153	phosphatidyl serine synthase	GB: HUMXT01443(EST) −6.4e-4; 3.0e-8			BE

Table I. Continued

ORF/gene[1]	Size (aa)	Known or predicted protein function/activity[2]	Closest similarity in the databases[3]	3D or tm	Prosite or other motif(s)	Phylum hits
YCL3w (+YCL4w)[b,13]/PEL1?	176	phosphatidyl serine synthase	GB: HUMXT01443 (EST) −6.1e-6[14]			BE
YCR2c[a,b,c,15]/CDC10	322	GTPase[16]	BH5_MOUSE −1.8e-60	5P21	GTPase (PS00152)	BEY
YCR4c[b]	247	FMN binding protein; flavodoxin(?)	GB: ECOWRBA_1 (Trp repressor binding protein) −1.2e-26; FLAV_CLOAB −1.4e-4	3FXN; 4FXN	FMN bind	B
YCR5c[c]/CIT2	460	**peroxisomal citrate synthase**	CISY_YEAST −83%/404	4CTS	citrate synthase	BEY
YCR8w[c]	603	protein kinase	NPR1_YEAST −1.8e-26	2CPK	PS00107; 00108	EY
YCR9c[b,c]/RVS161	265	**involved in starvation response**	PIR: S22700 −1.4e-11		SH3 domain	EY
YCR10c[b]	283	permease YAAH_ECOLI −1.2e-15	SCRPC34_2 −78%/267;	4 tms		BEY
YCR11c[c]/ADP1	1049	active transport ATPase	BROW_DROME −9.0e-27	8 tms	PS00152; PS00211	BEY
YCR12w[c]/PGK1	416	**phosphoglycerate kinase**	PGK_KLULA −82%/416	3PGK	PS00111	BE
YCR14c[a]	582	type X DNA polymerase	DPOB_RAT −1.2e-8		PS00522	E
YCR18c[a]	225	transcription regulator(?)	GAT1_HUMAN −1.0; 4.3e-4	1ZNF	PS00344	E
YCR19w[b]/MAK32	363	**maintenance of killer phenotype**; sugar kinase	RBSK_YEAST (YCR36W) −2.9e-1; 1.5e-2		ribokinase family[9]	BEY
YCR20c[b]		transcription regulator	TENA_BSUB −1.0; 5.0e-3; OPTAL −7.2 SD			B
YCR20w-a[b]/**MAK31**	78	**maintenance of killer phenotype**; putative component of snRNP	RUX9_MEDSA - 1.4e-3; 8.0e-9			E
YCR23c[a]	611	membrane transporter	NORA_STAAU −1.2e-9	6 tms		BEY
YCR24c[c]	492	Asn-tRNA synthetase	SYN_ECOLI −3.5e-48	3D	PS00179; 00339	BEY
YCR24c-a/PMP1[17]	40	**plasma membrane proteolipid, H^+-ATPase component**	none	1 tms		
YCR26c[a]	743	membrane phosphodiesterase	PC1_MOUSE −1.6e-16	1 tms		E
YCR27c[c]	209	GTPase	RAS_DICDI −3.3e-27	5P21	GTPase (PS00152)	BEY
YCR28c[c]	512	amino acid permease	DAL5_YEAST −3.2e-7	10 tms		BY
YCR29c-a[18]/**RIM1**	118	**ssDNA binding protein; mitochondrial DNA replication**	BRUSSB_2 (*Brucella abortus*) −1.3e-3; 1.7e-8		PS00735; 00736	BE
YCR31c[c,19]/CRY1	137	**ribosomal protein**	RS14_KLULA −94%/137		PS00054	BE
YCR32w[a]	2167	unknown	GB:HUMCDC4a_1 ('CDC4-related' protein)[20] −1.2e-58	2 tms		E
YCR34w[b]	347	membrane transporter(?), receptor(?)	PIR: S28299 −2.2e-10	4 tms		E
YCR36w[a]/RBK1	333	**Ribokinase**	RBSK_ECOLI −2.2e-16		PS00583; 00584	BEY
YCR37c[b]	923	membrane transporter(?), receptor(?)	YUR2_YEAST −58%/105	GB: RATNASI-_1 −1.5e-6 tms		EY
YCR38c[b,c]/BUD5	538	**bud site selection**; GDP−GTP exchange factor	PIR: S19399 (YCL68c) −98%/190; GNRP_MOUSE −5.2e-6	PS00720	EY	
YCR39c[c]/MAT-α2	210	**mating type protein**	PIR: S19010 (homeotic protein) −2.5e-5	1HDD	Homeobox; PS00027	EY
YCR40w[c]/MAT-α1	175	**mating type protein**	none			Y
YCR42c[c]/TSM1	1407	**lethal temperature-sensitive phenotype**	none	1 tms		N
YCR45c[c]	491	protease	YSP3_YEAST −1.9e-41	1SBC; 1 tms	PS 00136; 00137; 00138	BEY
YCR47c[a]	275	protein carboxyl methylase	PIMT_ECOLI −3.4e-2; 3.2e-7		SAM bind	BEY
YCR51w[b]	222	unknown	LATA_LATMA −5.4e-6		5 ankyrin-like repeats[21]	BE
YCR52w[b]	483	unknown	PIR: S19063 (gene complementing petite mutation) −1.0e-28; PIR: A30222 −1.0e-4			EY
YCR53w[c]/THR4	514	**threonine synthase**	THRC_CORGL −48%/115	3D	threonine synthase	BE
YCR54c[b]/CTR86	563	unknown	GB:MNEEP −5.6e-9			E
YCR57c[c]	439	regulatory protein of the β-transducin family	GBB2_BOVIN −8.3e-23		PS00678 (TPR repeat)	EY
YCR59c[b]	258	unknown	EMBL: TFPOLDNA −2.5e-2; 9.9e-7			B
YCR60w[b]	111	regulatory protein, regulation of mitosis(?), chromosomal segregation(?)	GB: HUMIEF_1 −3.2e-17		TPR repeat	

Table I. Continued

ORF/gene[1]	Size (aa)	Known or predicted protein function/activity[2]	Closest similarity in the databases[3]	3D or tm	Prosite or other motif(s)	Phylum hits
YCR62w (+YCR61w?)[b,22]	120 + 583	transmembrane protein	GB: YSCTSF3G −2.4e-3; 6.0e-7	2 tms + 7 tms		Y
YCR63w[c]	157	nucleic acid binding protein ?	G10_XENLA−4.7e-72		Zn finger type II	E
YCR65w[c]	532	transcription factor; **supressor of calmodulin mutation (Hcm1p)**[23]	GB: DROFD3BPA_1 (transcription factor containing *fork head* domain) −1.7e-22		PS00657; 00658 (*fork head* domain)	E
YCR66w[c]/RAD18	487	**DNA repair** (transcription regulator?)	PIR: S28290−5.4e-9		PS00518 (C3HC4 Zn finger)	EY
YCR67c[c]	1065	Intracellular protein transport	SC12_YEAST (membrane glycoprotein)−47%/226		PS00014 (EPR retention)	EY
YCR69w+	318	peptidyl-prolyl isomerase (N-terminal portion)	CYPC_YEAST (cyclophilin) −7.5e-14	3D		EY
YCR70w[a,24]/CYP4(SC-C3)						
YCR72c[a]	514	regulatory protein of the β-transducin family	PRO4_YEAST −3.0e-20		PS00678 (TPR repeat)	EY
YCRX13w[b,25]	315	NAD-dependent oxidoreductase ?	EMB:SCPAMIBEN (yeast Chr XIV) −73%/134; GB: T01569 (EST)−1.6e-2		NAD bind	EY
YCR73c[c]	1314	protein kinase	ST11_yeast−6.4 e-24	2CPK	PS00107; 00108	EY
YCR75c[c]/ERS1	260	**ER defect supressor;** intracellular protein transport	none	6 tms		N
YCR77c[b]	509	unknown	GB: XELP100A_1 −4.9e-4; 3.3e-10			E
YCR83w[c]	127	Thioredoxin	THI2_yeast−4.6e-23	3TRX	PS00194	BEY
YCR84w[c]/TUP1 (AER2; SFL2; CYC9)	713	Regulatory protein of the β-transducin family	CC4_YEAST−1.5e-27		PS00678 (TPR repeat)	EY
YCRX16c[b,26]	153	nucleic acid binding protein	GB: HUMZFP431 −8.4e-4; 2.0e-6	1ZNF	Zn finger	
YCR88w[b,c]/ABP1,	592	**actin binding protein**	GB:CELF42H10_4−9.7e-5		SH3 domain	E
YCR89w[c]	1609	cell adhesion	PIR: A41258−1.2e-23			EY
YCR91w[c]/KIN82	726	Ser/Thr protein kinase	PIR: B30311−48%/154	2CPK	PS00107; 00108	EY
YCR92c[c]/MSH3	1047	ATPase, mismatch repair	GB: MUSREP3B_1−56%/111		PS00152; 00486	BEY
YCR93w/CDC39[27]	2108	**negative regulator of transcription**	none			
YCR94w[b]	391	unknown	EMB D15884 (EST)−3.7e-8			E
YCR96c[c]/MATα2	119	mating type protein	PIR: S19010 (homeotic protein)−7.9e-6	1HHD	PS00027 (homeobox)	EY
YCR97w[c]/MATa1	126	mating type protein	MTA1_yeast−8.6e-22			N
YCR98c[a,28]	518	carbohydrate transporter	PH84_YEAST−3.0e-14	8 tms	PS00216	BEY
YCR99c[b,29]	155	membrane protein; sialidase (pseudogene?)	GB: SC114_1−60%/151			Y
YCR100c[b,29]	316	membrane protein; sialidase (pseudogene?)	GB: SC114_1−52%/261		4 sialidase repeats	Y
YCR101c[b,29]	182	membrane protein; sialidase (pseudogene ?)	GB: SC114_1−7.6e-17	1 tms		Y
YCR102c[b]	368	alcohol dehydrogenase	GB: ENHADH1A_1−9.6e-4; 8.2e-5			E
YCR104w[a]	124	unknown (cold shock)	SRP1_YEAST−2.6e-4; 9.9e-9		PS00724 (SRP1/TIP1)	Y
YCR105w[c]	361	alcohol dehydrogenase	PIR: S24261−1.4e-25	7ADH	Zn-containing ADH	BEY
YCR106w[b,30]	832	transcription regulator	CYP1_YEAST−2e-4; 1.6e-8	1D66 2 tms	PS00463 (Zn2−Cys6 binuclear cluster)	EY
YCR107w[b,c]	363	aldoketoreductase	GB: PHAAD_1−61%/143			BE

[a] described by Bork *et al.* (1992a,b).
[b] described in this paper.
[c] described by Oliver *et al.* (1992).
[d] described by Goffeau *et al.* (1993a,b).
In the 3D column, Protein Data Bank (PDB) identifier of the most closely related protein with known three-dimensional structure is included, if available; 3D, tertiary structure of a homologue is known, but is not in PDB.

tms, transmembrane segment (s).
In the Motifs column the Prosite identifier is indicated wherever available; when additional motifs that were not in Prosite were found in the given sequence and its homologues, the identifier is shown in parentheses.
In the final column, B indicates a hit with at least one bacterial protein; E, a hit with at least one eukaryotic protein from another phylum (non-fungi); Y, a hit with at least one protein from yeast or other fungi; and N, no similarity in the current sequence databases.
Experimental data are cited only if unavailable at the time of the previous analyses (Bork *et al.*, 1992a,b; Oliver *et al.*, 1992).
[1]The first name starting with 'Y' is the ORF designation from Oliver *et al.* (1992) and the gene name is indicated after a slash whenever available.
[2]Bold type indicates proteins for which the function has been determined experimentally.
[3]Identity of each sequence to itself was disregarded and the next closest similarity was included; where the self-identity was the only significant BLAST hit, 'none' is indicated. In cases of high similarity (>40% identity in an ungapped alignment of >100 amino acid residues in length, generated by BLAST), percentage identity/length is indicated. In all other cases, the first number shown is the lowest Poisson probability of matching by chance given by the BLAST search, and the second number is the lowest probability calculated using MACAW. The MACAW probabilities are shown only when the lowest BLAST probability was higher than 10^{-4}.
[4]The putative transposon has been described by Voytas and Boeke (1992). The sequence downstream of YCL74w encodes an RNase H-related product.
[5]Very recently, a single ORF coding for a putative membrane protein that is strongly similar to YCL70c, YCL71c and YCL73c has been identified in yeast chromosome XI (H.-W.Mewes, personal communication).
[6]The coding region upstream of the 5′-terminal ATG of this ORF showed highly significant similarity to yeast protein SGE1 (indicated in the table; Amakasu *et al.*, 1993), suggesting that the YCL69w protein may be longer by 97 amino acids residues at the N-end than previously believed (see text).
[7]The similarity with FtsJ, but not the SAM-binding motif, has been reported by Tomoyasu *et al.* (1993).
[8]The conserved domain did not include the RNA-binding motif of LA antigen and may have a different, uncharacterized function.
[9]This ORF has been tentatively reconstructed by correcting two probable frameshifts in its 5′-terminal portion.
[10]This particular EST library has been shown to be heavily contaminated with bacterial sequences (Savakis and Doelz, 1993); thus the human origin of HSA36F062 should be considered tentative.
[11]The protein kinase comprises the N-terminal domain; the function of the C-terminal domain is unknown.
[12]The similarity with aminotransferases has been described by Ouzounis and Sander (1993) and by Mehta and Christen (1993).
[13]YCL4W and YCL3W may comprise two portions of a single gene separated by a frameshift.
[14]The BLAST probability is given for YCL3w together with the previously unidentified N-terminal region (see text); the original sequence of YCL3w did not show any significant similarities.
[15]The 5′-terminal portion of this ORF is an autonomous replicating sequence (ARS); the penultimate AUG may be used for translation initiation.
[16]YCR2C belongs to a distinct subfamily of putative GTPases.
[17]This small gene has not been recognized in Oliver *et al.* (1992) and Bork *et al.* (1992a,b) but has been described subsequently by Navarre *et al.* (1992).
[18]Spliced gene that has not been described initially as both exons are <100 codons; expression of the gene, DNA binding properties of the product (RIM1) and its participation in mitochondrial DNA replication have been subsequently studied experimentally (E.Van Dyck *et al.*, 1992).
[19]Spliced gene (Oliver *et al.*, 1992).
[20]In our analysis the human protein related to YCR32W did not show any similarity to yeast CDC4.
[21]Bork (1993).
[22]YCR62W may constitute the C-terminal domain of a larger protein having YCR61 as the N-terminal portion.
[23]Zhu *et al.* (1993).
[24]YCR69W and YCR70W are actually the N-and C-terminal portions of a single protein, respectively, as shown in the original sequencing study (Franco *et al.*, 1991).
[25]This ORF is likely to be expressed instead of YCR74C.
[26]This ORF is likely to be expressed instead of YCR87W.
[27]Collart and Struhl (1993).
[28]The sequence similarity reported by Sor *et al.* (1992).
[29]YCR99C, YCR100C and YCR101C correspond to different portions of PEP1 and may comprise portions of a single (pseudo)gene.
[30]The similarity to CYP1 is in the N-terminal, Zn cluster-containing domain whereas the transmembrane segments are in the C-terminal domain, unique for this new family of uncharacterized proteins.

Table II. Summary of sequence similarities and function identification for yeast chromosome III ORFs

	Number of ORFs	Percent
Putative proteins encoded in yeast chromosome III	171[a]	100
Significant sequence similarity to proteins in the database	104	61
to other yeast proteins	67	39
to eubacterial proteins	42	25
to proteins from other eukaryotic phyla	81	48
Total 'cross-phylum hits'	92	54
Known sequence motifs	56	33
Homologues with known 3D structure	32	19
Known or predicted function/activity	93	54
Putative membrane proteins	51	30

[a]ORFs that showed no similarity to other proteins in the databases but instead were found to be similar to non-coding sequences (see text) were excluded. ORFs whose products showed similarity to different portions of a single protein were counted as one, even keeping in mind that some of them may be pseudogenes and the frameshifts separating them may be real; exons of spliced genes also were considered as one ORF.

YCL65w, YCR13c and YCR44c) have obvious similarities to regulatory and other functionally characterized untranslated DNA sequences (data not shown). Further reduction comes from sequences that were merged in cases where two ORFs (three cases) or even three ORFs (one case) apparently contained portions of a single gene (Table I and below). On the other hand, several segments between the original proposed ORFs (Oliver *et al.*, 1992) appear to be expressed (see below).

Another type of error in gene assignment can occur when a smaller ORF is completely contained in a larger one or overlaps strongly with it, typically on the reverse strand.

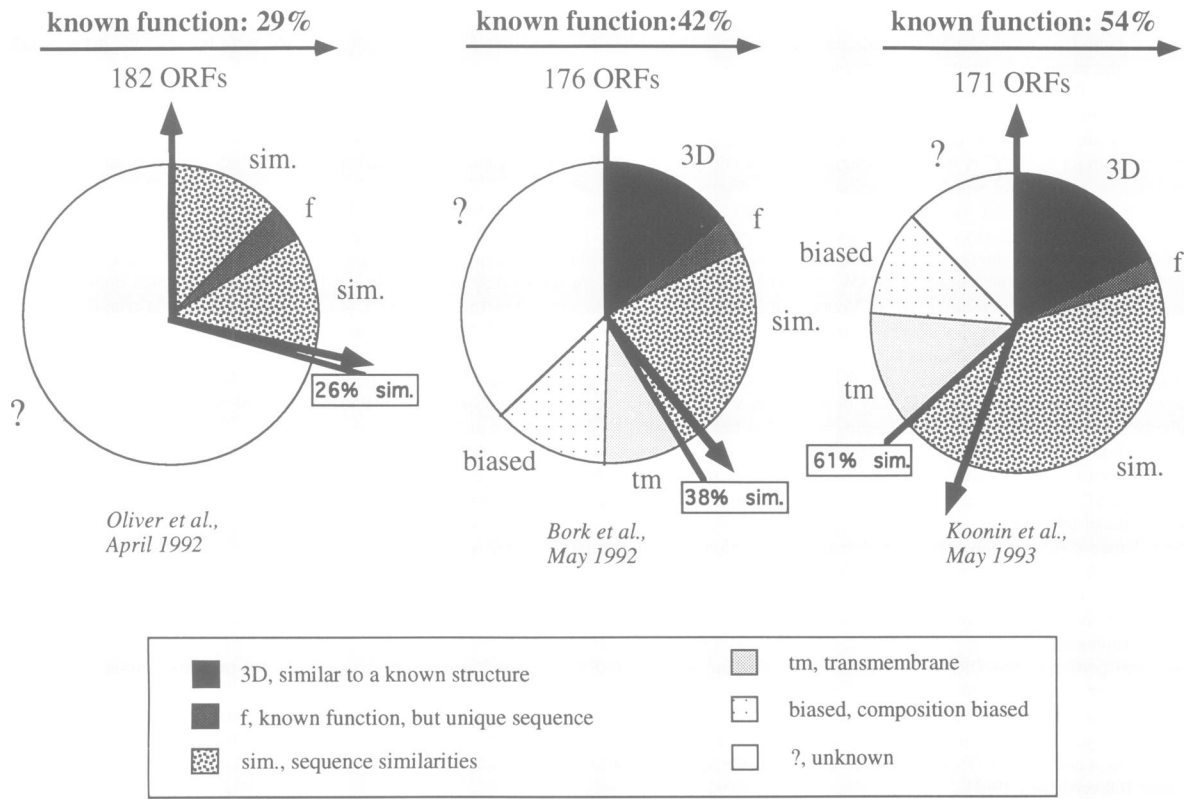

Fig. 1. Growth of information about ORFs in yeast chromosome III. The 'information clock' shows the accumulated knowledge about the known and predicted protein products at three different time points, according to (**left**) Oliver *et al.* (1992), (**center**) Bork *et al.* (1992a) and (**right**) this work. From top clockwise: 3D = ORFs homologous to proteins of known three-dimensional structure; f = ORFs with experimentally known function, but no relatives in the databases; sim. = ORFs with relatives in the sequence databases, but of unknown three-dimensional structure; function = ORFs with known or predicted function; tm = ORFs with one or more predicted transmembrane segments; bias = ORFs with significant bias in amino acid composition untypical of globular proteins; ? = all other ORFs. In the left-hand chart, three-dimensional similarities were not studied; the actual number is indicated by two 'sim.' regions.

Only one of the two is likely to be expressed, but not always the larger one, as originally supposed (Oliver *et al.*, 1992). An example of such a reassignment is YCRX16c (opposite strand of the larger YCR87w) which is related to a group of zinc finger proteins. A unique signature for this subfamily is Cx_2C[GS]$KxFx_5$&xHx_{2-3}[CH].

In addition, as noticed by Tanaka and Isono (1993), in two cases the longest ORF for a given region has been identified erratically. Specifically, ORFs YCLX8c and YCRX13w should be considered candidate genes instead of the smaller ORFs YCL26c and YCR74c, respectively. We found supporting sequence similarities for both these ORFs.

YCLX8c (opposite strand of YCL26c) is closely related to another ORF (perhaps a pseudogene; see below) located in the same region of chromosome III, downstream of YCL27w. Both ORFs are also similar to a human expressed sequence tag (EST) (Table I) and share a conserved motif with a group of uncharacterized bacterial membrane proteins (data not shown). The signature $\&x_2Wx_2$[UA]x_3GLGx_2 LQ[NH] uniquely characterizes the current members of this new protein family (residue types: U, bulky aliphatic; &, bulky hydrophobic; x, any type; alternative residues are in square brackets).

YCRX13W is closely related to an uncharacterized ORF downstream of the paraaminobenzoate (PABA) synthase gene on yeast chromosome XIV (Edman *et al.*, 1993; Table I). The ORF on chromosome XIV is, however, interrupted by a stop codon.

From ORFs to genes: detecting frameshifts via sequence similarity

When two adjacent ORFs have a match in the database to adjacent regions of the same protein, a frameshift, possibly as a result of a sequencing error, is the likely cause. There are at least three such cases in the reported sequence of yeast chromosome III:

(i) YCR69W and YCR70W are two parts of a gene encoding a cyclophilin (Franco *et al.*, 1991; Bork *et al.*, 1992; Table I).

(ii) YCR99c, YCR100c and YCR101c together correspond to a gene coding for a membrane protein related to yeast PEP1 (L.Van Dyck *et al.*, 1992; Table I). The presence of two frameshifts makes it more likely that this is an unexpressed pseudogene. However, we found that PEP1 and YCR100c have four 20 residue repeats conserved in a wide variety of sialidases (neuraminidases) and probably involved in activity (Rothe *et al.*, 1991). So at least PEP1 is likely to be a sialidase.

(iii) YCL3w together with YCL4w probably codes for a phosphatidyl serine synthetase (PSS) (Table I; Figure 2). Both ORFs have modest similarity to *Escherichia coli* and human PSS. When the sequence upstream of the presumed initiation codon of YCL4W was translated and compared with PSSs, a much stronger conservation became apparent. The putative new PSS appears be completely unrelated to the known yeast PSS that is in the family of CDP-alcohol phosphatidyltransferases (Nikawa *et al.*, 1987; Hjelmstad

```
consensus          .U.&&..PA.F.E.U......S      ..L.V.&UUD..RG.R.
                         S   D        A                    A
PSS E.coli     25  DVDFFYRPADFRETLLEKIASA 32  PELDVRVLVDWHRAQRG
                   ::*    *::* : *  **  :     * * *  *:*  *: *
YCL4w          33  EIDIIESPSQFYDLLKTKILNS 28  PKLKVSFLLDGLRGTRE
                    : :: **::*::*:*   *:*       **** ***  **:*
PSS human      (3) HVRVLSSPAEFFELMKVDCLES 11  SNLKVSILLDFTRGSRG


consensus          V...U@..P.............E.UGU.H
PSS E.coli     22  VDVPVYGVP----------INTREALGVLH
                   **  :*                *:**: *
YCL4w          21  VDCRLYKTPAYHGWKKVLVPKRFNEGLGLQH C-end
                   *   *: **: **  ::*:* **** :****
PSS human      19  VRVSLFHTPHLRGLLRLLIPERFNETIGLQH


consensus          &K.@.&DN.VU&SGA.L.D.Y&.Q..D..&
                         D          N    N  N
PSS E.coli         FKGFIIDDSVLYSGASLNDVYLHQHDNIAY 283  P23830
                    * :  *: *: *** * : *  : ::  *
YCL3w'             LKIYGFDNEVILSGANLSNDYFTNRQDRYY 204
                   :*:* *** *********: ** :  ***
PSS human          IKVYLFDNSVILSGANLSDSYFHQPSDRYV (4) M77859G
```

Fig. 2. Frameshift merger: YCL4w and YCL3w together are a putative phosphatidyl serine synthase (PSS). When merged into a single sequence, YCL4w and YCL3w show significant similarity to a family of PSSs. The alignment is shown as a set of conserved blocks, with the distances between them as well as the distance from the N-terminus indicated as numbers. YCL3w′ is the amino acid sequence encoded upstream of the original initiation codon of YCL3w; the rest of YCL3w had some additional similarity to the *E.coli* PSS (not shown). Identities between the yeast sequences and each of the two PSS species are shown by asterisks and similar amino acid residues are shown by colons. The consensus line shows residues conserved in the three aligned sequences; U designates a bulky aliphatic residue (I, L, V or M); @ designates an aromatic residue (F, Y or W); & designates a bulky hydrophobic residue (either aliphatic or aromatic); and dot designates any residue. The conserved positions are also highlighted by bold typing. The overlined signature, LxVx&LUDx$_2$R[AG]xR, is unique for this emerging family of PSSs. For the two PSS sequences accession numbers from SwissProt or GenBank (marked by G) are indicated.

and Bell, 1991). Recently, the *PEL1* gene has been assigned to this region of chromosome III indicating its expression and possible role in the mitochondrially mediated control of cell division (Janitor and Subik, 1993).

Other instances of apparent frameshifts (Table I) include (i) YCR61w and YCR62w, which may encode portions of a single transmembrane protein, (ii) the 3′-terminal region of YCL74W (Voytas and Boeke, 1992), (iii) the 5′-terminal region of YCL8C, (iv) two probable frameshifts in the newly discovered ORF YCL27c-a and (v) an apparent pseudogene in the centromeric region related to *DOM34* on chromosome XIV. In addition, the coding sequence upstream of YCL69c, separated by a termination codon in the reported sequence, is highly similar to yeast protein SGE1, suggesting that the actual protein may be 97 amino acids longer at its N-terminus.

It remains to be determined which frameshifts are sequencing errors, with the respective genes actually encoding active proteins, and which correspond to 'fresh' pseudogenes that have not yet accumulated numerous mutations.

Gene duplications in yeast

Sequence similarities provide evidence of two types of apparent gene duplication. Firstly, strong similarities between yeast sequences indicate recent duplication. Secondly, evolutionarily distant duplications, occurring before the divergence of yeast from other species, are likely when similarity is stronger between species than within the yeast genome. On chromosome III, very few cases belong in the first class. These are two identical copies of mating factors and a partially identical copy of the *BUD5* gene (Oliver *et al.*, 1992). In addition, the adjacent ORFs, YCLX8c and YCL27c-a, resemble each other and appear to be a tandem duplication.

Between yeast chromosomes, there are strong similarities between YCRX13w (on chromosome II) and a putative NAD-dependent dehydrogenase (XIV); between centromeric regions of chromosomes III and XIV (Lalo *et al.*, 1993), including the two citrate synthases CIT2 (YCR5c, III) and CIT1 (XIV), putative membrane proteins YCR10c (III) and FUN34 (XIV), putative RNA-binding proteins YCL11c (III) and TOM34 (XIV); and an apparent pseudogene interrupted by multiple frameshifts (III) and DOM34 (XIV) (Lalo *et al.*, 1993). YCL61 and YCR37c represent an additional interesting case of apparent interchromosomal duplication. They are closely related to two uncharacterized ORFs flanking the yeast uridine kinase (*URK*) gene [chromosomal location unknown (Kern, 1990)] from the 5′-end and the 3′-end, respectively (Table I). There is no homologue of *URK1* itself on chromosome III, and YCL61 and YCR37c are separated by a large number of genes, indicative of a gene rearrangement accompanying the duplication.

The second type of duplication, with much weaker intra-yeast similarity, includes at least 11 likely membrane transporters on chromosome III, four probable regulatory proteins related to transducin β-2 subunit, four Ser/Thr protein kinases, and YCR19w and YCR36w, probable sugar kinases of the ribokinase type, identified by characteristic sequence patterns (Table I; Bork *et al.*, 1993).

From these first observations, one may conclude that yeast chromosomes have diversified in part by duplication of chromosomal segments, accompanied in some instances by gene shuffling and/or by frameshift disruption.

Unexpected conserved motifs and very subtle similarities

In some cases, even modest sequence similarity may be sufficient for functional identification, e.g. if a conserved functional motif is detected. Such motifs typically consist of one or more conserved sequence boxes, each ~15–30 residues long. Examples are a sugar kinase motif in YCR19w and the methyltransferase signature in YCR47c (Bork *et al.*, 1992). Based on the conservation of the putative SAM-binding motif (Bork *et al.*, 1992; Koonin, 1993), we now suggest that YCL54w, which is related to *E.coli* cell division protein FtsJ (Tomoyasu *et al.*, 1993), may be another methyltransferase.

Motif conservation also establishes the similarity between YCR4c, the closely related *E.coli* Trp repressor-binding protein WrbA, and distantly related bacterial flavodoxins. We found that all these proteins contain a conserved FMN-binding motif. The intriguing implication is that both YCR4c and WrbA, identified as an accessory regulator of the tryptophan operon transcription (Yang *et al.*, 1993), are FMN-binding proteins.

A combination of multiple sequence alignment and motif definition using MACAW and OPTAL shows that YCR20C is similar to the *Bacillus subtilis* transcription enhancer TenA (Pang *et al.*, 1991). The two proteins are of almost identical size and the conserved residues are in several regions. When combined with a third sequence, that of an alleged human cDNA clone (Figure 3), the probability of random

```
consensus          H.F...U..GTU..D..&.&YU.QD..&&.....U..........
HSA40F human   28  HPFVQAIAAGTVPNDV-LNYYVEQDEHYLKDYLQVTALTITKTDNV
                   * *  : ***:  *  *  *: **  :   *  :   * :
YCR20c         21  HKFAKELCAGTLK-DRSLYIYLSQDLQFFETSLRLICKTTSLAPTT
                   * *  : ***  **    *: **  :     ::     : *
TenA Bs        20  HPFVQGIGDGTLPIDR-FKYYVLQDSYYLTHFAKVQSFGAAYAKDL

consensus      .....U..........E........EU....U....R
HSA40F         DDINQLLTTAQ-FVQNESRAHQVMLEITGHTIENWRRE  (4) Z13332G
                 :  *     * ::*       **: : ::    *:
YCR20c         HALITLAKKIGFFSNDENSYFHDCLELLAPSLTKEERDNFDNKAIPGVDAYIN
                  :*       : * :      ***   :: ***  *
TenA Bs        YTTGRMASHAQGTYEAEMALHREFAELL--EISEEERKAFKPSPTAYSYTSHM

YCR20c         FLDELRKDASITWPSLVTSLWVAEELYWRWARDTPRAPGLHWKYQKWIDLHDG
               :  *       *    :  * :       **  :           *  *****    *
TenA Bs        YRSVLSGNFAEILAALLPCYW----LYYEVGEKLLHCDPGHPIYQKWIGTYGG

YCR20c         EHFQTWCEFLKAEVDKFPVEEVESI-------FVKVSQFEFEFFESCY  2
               : *   *     *    :  * :          **  * :*::*:   *
TenA Bs        DWFRQQVEEQINRFDELAENSTEEVRAKMKENFVISSYYEYQFWGMAY 23  P25052
```

Fig. 3. Detecting very subtle similarities: YCR20C may encode a novel transcription regulator. The consensus is shown only for the region where the three-way alignment was available. HSA40F is an alleged human cDNA clone (EST); the origin of this EST should be considered tentative (Savakis and Doelz, 1993). Regardless of this, the alignment with HSA40F supports the observed sequence conservation between the sequences of YCR20c and *B.subtilis* (Bs) transcription enhancer TenA. Other notation is as in Figure 2.

occurrence of the triple similarity is very low (estimated as $\sim 10^{-11}$).

These and other examples (Table I) illustrate that detailed analysis, including derivation of multiple alignments and motifs, can reveal significant similarities that are not evident in initial searches. While caution is appropriate in the interpretation of such observations, they may provide very useful clues for further functional analysis of gene products.

New protein families in search of a function

For 17 putative proteins encoded in chromosome III, significant sequence similarities were observed with proteins of unknown function. For example, YCR59c is similar to proteins with no established function from three very different bacterial species (Figure 4). YCLX8c and YCL27c-a, which are similar to each other and to an alleged human cDNA clone and share a conserved motif with three uncharacterized bacterial proteins, provide another example. While less informative than similarities that permit functional prediction, these observations define new protein families. If the database hit is from a very different organism, e.g. bacteria, the alignment typically contains a characteristic motif (e.g. Figure 4) that is very probably involved in function. This is useful not only for extending databases of protein motifs, such as ProSite (Bairoch, 1993), but also for functional analysis using site-directed mutagenesis.

Intergenic regions and small ORFs

Database searches with the 36 long (>700 nt) intergenic regions present between the originally (Oliver *et al.*, 1992) assigned ORFs in chromosome III revealed six significant similarities. These include (i) the apparent continuation of YCL74w across a frameshift, (ii) the 5′-terminal extension of YCL69w across a termination codon, (iii) the reconstruction of YCL27c-a from pieces smaller than 100 codons by correcting two probable frameshifts and (iv) the pseudogene homologue of *DOM34*. Interestingly, there is moderate but statistically significant sequence similarity between *DOM34*, the chromosome III pseudogene and the family that includes yeast omnipotent suppressor of nonsense codons (*SUP1* gene product) and related proteins from plants

```
consensus      S.FU.........E.A..&U...K..      .A.H...A@..
                            Q      R
YCR59c   145   STFMAFAAHVTSEEQAFAMLDLLKTD  4  KANHVMSAWRI  10
               * *:* **   ***:* * *   :       * *   *:**
ORF  Tf  (4)   SRFLAKAAPAASEEEALAFLAAHREP  0  QATHNAYAYRI   6
YigZ Ec  20    SRFITMLAHTDGVEAAKAFVESVRAE  2  DARHHCVAWVA  10
L    Bs  20    SRFICHLSRVSTEQEAQEFIQKIKKQ  2  NATHNCSAYVI   8

consensus      .D.DGE..A.G...UL..U....U....UUV.R@@GG..UG..
                       G A          A
YCR59c         SDDDGET-AAGSR-MLHLITIMDVWNVIVVVARWFGGAHIGPD  20
               ** ***  * *   :** *   :  *:*:* *:*** :*
ORF  Tf        SD-DGEPRAPGRP-ILHAIEAAGLDRVVVLVVRYFGGVKLGAG  91  X66105G
YigZ Ec        SD-DGEPAGTAGKPMLAQLMGSGVGEITAVVVRYYGGILLGTG  93  P27862
L    Bs        ND-DGEPSGTAGVPMLEVLKKRRLKDTCAVVTRYFGGIKLGAG  67  A30191P
```

Fig. 4. Conserved proteins in search of function. YCR59C encodes a protein that helps define a new family with both eukaryotic and prokaryotic members. One of the latter, an uncharacterized, incomplete ORF, was found downstream of the DNA polymerase gene of *Thermus flavus* (Tf); Ec, *E.coli*; Bs, *B.subtilis*. The overlined signature, $ULx_2Ux_4Ux_5UVxR@@GGx_2UG$, was unique for this family of uncharacterized proteins.

and animals. The fifth example involves two exons of YCR29c-a or the *RIM1* gene (E.Van Dyck *et al.*, 1992) which individually are smaller than 100 codons and had escaped detection. Finally, a previously described small protein, YCR20′w (MAK31), turned out to be similar to components of small nuclear ribonucleoproteins. MAK31 is required for the maintenance of the killer phenotype, which is conferred by double-stranded RNA elements (Wickner *et al.*, 1986). The possible presence of such elements in small nuclear RNP complexes is intriguing.

Although examples of small genes expressed in yeast are known, including the *PMP1* (YCR24c-a) gene in chromosome III (Navarre *et al.*, 1993), only one gene identified by homology in fact encodes a protein with <100 residues. In the other five cases, the similarities detected in the 'intergenic' regions resulted in the reconstruction of larger genes that had not previously been detected.

Transmembrane segments and low complexity regions

Transmembrane helices were predicted in 50 proteins (Table II), i.e. in ~30% of all proteins encoded in chromosome III. This is a conservative estimate, counting only proteins for which two methods (Eisenberg *et al.*, 1984; Rao and Argos, 1986) agreed on at least one transmembrane segment. The higher estimate of 35–40% (Goffeau *et al.*, 1993b) may be too high [for example, three proposed transmembrane proteins (YCR12w, YCR5c and YCL18w) are in fact homologues of proteins with known three-dimensional structure that are clearly not membrane-associated; Table I]. The actual fraction of proteins with transmembrane segments is probably ~30–35% of all chromosome III proteins.

Seventeen proteins, or 10% of the total, were predicted to contain more than four transmembrane helices and may be thought to function as membrane transporters (permeases) or as receptors. Eleven of these have sequence similarity to known permeases (Table I). The fact that 34 of the 78 proteins of still unknown function (Figure 1) appear to have transmembrane segments narrows down their possible function.

Analysis of regions of low complexity revealed 10 proteins in which such regions cover >20% of the total length (Figure 1). Only two of these are proteins of known function, namely YCR89w and YCL11c. For all others, the unusual amino acid composition remains to be interpreted.

Table III. Proteins similar to expressed sequence tags in yeast chromosome III

ORF	Known or predicted function/activity	Similar EST(s)
1 YCL59c	unknown	*Caenorhabditis elegans*
2 YCL45C	unknown	*C.elegans*
3 YCL43c	protein disulfide isomerase	*C.elegans*, *Arabidopsis thaliana*, human
4 YCL40w	glucokinase	human
5 YCL37c	unknown	rice
6 YCL33c	transcription regulator	rice
7 YCL27c-a	unknown	human[a]
8 YCL24w	protein kinase	*C.elegans*, rice, human
9 YCL19w	POL polyprotein	human
10 YCL18w	β-isopropylmalate dehydrogenase	*C.elegans*
11 YCL17c	aminotransferase (NifS homologue)	*C.elegans*
12 YCL11c	RNA binding protein	*C.elegans*, rice, human
13 YCL4w+ YCL3w[b]	phosphatidylserine synthase	human
14 YCR2c	GTPase	*C.elegans*, human
15 YCR5c	citrate synthase	*C.elegans*, *A.thaliana*
16 YCR8c	protein kinase	*C.elegans*, rice, human
17 YCR9c	starvation response protein (conserved SH3 domain)	*C.elegans*
18 YCR11c	active transport ATPase	*C.elegans*, rice, human
19 YCR12w	phosphoglycerate kinase	*C.elegans*, rice, human
20 YCR24c	Asn-tRNA-synthetase	*C.elegans*
21 YCR27c	GTPase	*C.elegans*, rice, *A.thaliana*
22 YCR31c	ribosomal protein S14	rice, *A.thaliana*
23 YCR51w	ankyrin repeat protein	*C.elegans*, human
24 YCR57c	regulatory protein	*C.elegans*, *A.thaliana*, human
25 YRC60w	regulatory protein	*Plasmodium falciparum*
26 YCR63w	nucleic acid-binding protein?	rice, human
27 YCR66w	DNA repair protein	*C.elegans*, *A.thaliana*
28 YCR69w	peptidylprolyl isomerase (N-terminal portion)	*C.elegans*, *A.thaliana*, rice, human
29 YCR72c	transcription regulator ?	*C.elegans*, *A.thaliana*, human
30 YCR73c	protein kinase	*C.elegans*, *A.thaliana*, rice, *P.falciparum*, human
31 YCRX13w	NAD-dependent oxidoreductase?	*C.elegans*
32 YCR83w	thioredoxin	*C.elegans*, *A.thaliana*, rice
33 YCR84w	transcription regulator	*C.elegans*, human
34 YCR91w	protein kinase	*C.elegans*, *A.thaliana*, *P.falciparum*, human
35 YCR94w	unknown	rice
36 YCR105w	alcohol dehydrogenase	*A.thaliana*

[a]The same EST showed lower similarity to YCLX8c which is a diverged tandem copy of YCL27c-a (see text and Table I).
[b]Both YCL4w and the N-terminal extension of YCL3w are similar to the same EST (see text and Table I).

Similarities with expressed sequence tags

In addition, 37 proteins on chromosome III were found to have similarities ($P < 0.001$) in the database of partial cDNA fragments or ESTs [dbEST (Boguski *et al.*, 1993), 21 781 entries or ~6.9 megabases as of June 16, 1993]. Remarkably, most of these 36 are well characterized, highly conserved proteins (e.g. protein kinases, GTPases and ATPases). Moreover, for 20 of the 37 gene products, similarity was observed with ESTs from more than one source (Table III). As ESTs are likely to be a representative collection of highly expressed genes from several organisms (Adams *et al.*, 1993; Boguski *et al.*, 1993; Green *et al.*, 1993), similarities with ESTs may be indicative of highly expressed genes. Data on the correlation between yeast chromosome III ORFs and transcripts (Tanaka and Isono, 1993) can be used to test this hypothesis. Only four putative proteins matched isolated ESTs, i.e. those without related proteins in the current databases: YCL27c-a, YCL37c, YCL45c and YCR94w (Table I). Specific conclusions based on similarities with ESTs should be drawn with caution as the accuracy of these sequences may be low and their organismic origin not always secure (Savakis and Doelz, 1993).

Members of ancient protein families

Each round of chromosome III sequence analysis increased the fraction of ORFs for which similarities were found, with the current fraction exceeding 61% of all ORFs (Figure 1). The majority (>87%) of these similarities (Table III) are 'cross-phylum hits', i.e. relationships with proteins from phylogenetically distant organisms. The observed proportion of cross-phylum hits among all ORFs (54% of the total) was remarkably close to recent theoretical estimates of 50–60% for a representative gene collection (Claverie, 1993; Green *et al.*, 1993). About 40% of the cross-phylum hits included similarities to prokaryotic proteins, typical of members of ancestral protein families. Most of the chromosome III/EST similarities fit into well known, evolutionarily conserved families, an observation compatible with the concept of a limited set of 'ancient conserved regions' in protein sequences, a significant fraction of which is already known (Claverie, 1993; Green *et al.*, 1993).

Increasing the efficiency of functional identification

Our results indicate that computer-assisted analysis of genome sequences may be able to identify significant similarities for a considerably larger fraction of putative ORF

products than is currently appreciated (Adams *et al.*, 1993). Some of the new observations are the result of database growth, but most of them resulted from the increase in the sensitivity of similarity searches and from combined use of several different computer methods. Most of the observed sequence relationships have immediate interpretation in terms of function. The practical consequences are that a high level of computer-assisted functional assignments leads to more efficient strategies for the complete functional characterization of sequenced genes.

Within the next few years, with rapid growth of the size of sequence databases and refinement of computer methods, the fraction of detectable similarities in genome projects is likely to increase well beyond the 60% level. However, the level of functional identification cannot grow as fast because of the much slower progress of experimental characterization of proteins. The goal of 100% functional identification can only be reached within a reasonable time by a combination of functional experiments and further improvements in computer-assisted genome analysis.

Materials and methods

ORF selection and search of databases

The sequence of yeast chromosome III is entry SCCHRIII (X59720) in the EMBL/GenBank database. 182 ORFs longer than 100 codons were taken from Oliver *et al.* (1992). In addition, 33 ORFs that were longer than 100 codons, but completely contained in, or strongly overlapping with, some of the original 182 ORFs ('X' in the GenBank notation), were used. Finally, 36 regions longer than 700 nucleotides and located between the 215 ORFs were subjected to database searches.

Daily database updates were taken from the National Center for Biotechnology Information (NIH) and EMBL: a 'non-redundant' nucleotide database, the result of merging non-identical entries from GenBank (Benson *et al.*, 1993) and EMBL (Rice *et al.*, 1993); a 'non-redundant' protein sequence database, generated by merging non-identical sequences from PIR (Barker *et al.*, 1993) and SwissProt (Bairoch and Boeckmann, 1993) and amended by translations of GenBank and EMBL databases.

Initial database screening

Initial searches were performed using programs based on the BLAST algorithm (Altschul *et al.*, 1990). BLASTP compares a protein sequence with a protein sequence database, TBLASTN compares a protein sequence with the translation of a nucleotide sequence database in all six possible reading frames, and BLASTX (Gish and States, 1993) compares the 6-frame translation of a nucleotide sequence with a protein database. The latter program was used for the 36 inter-ORF regions. The BLAST tools are fast and give a statistically robust significance estimate for each local alignment (Karlin and Altschul, 1993 and references therein), but do not take into account insertions or deletions. Hits with the probability of occurrence by chance (P) of $< 10^{-4}$ were considered significant, and those with $10^{-4} < P < 1.0$ were subjected to a (T)FASTA search (Pearson and Lipman, 1988), with optional reordering according to length-dependent significance using the program FASTA-FILTER (C.Sander and R.Schneider, unpublished). (T)FASTA is slower than BLAST, but attempts to join blocks into gapped alignments. The BLOSUM62 matrix (Henikoff and Henikoff, 1993) was used in BLAST searches, and the PAM250 matrix (Dayhoff *et al.*, 1978), considered optimal for the detection of long but weak similarities, in FASTA searches.

A relatively permissive significance cut-off cannot be used productively in database screening unless the sequences are prefiltered to exclude low complexity (compositionally biased) regions that tend to produce spurious hits (i.e. the usual significance estimates and empirical cut-offs do not apply for these regions). Accordingly, the query sequences were routinely searched for low complexity segments using the SEG program (Wootton and Federhen, 1993). These segments were masked and excluded from the subsequent searches.

Motif search

Each sequence was searched for motifs form the ProSite library (Bairoch, 1993). Conserved ProSite motifs were also searched in BLAST outputs using the BLA program (Tatusov and Koonin, 1994). Motifs with low information content (e.g. phosphorylation and glycosylation sites) were omitted from these searches. BLAST and FASTA outputs were also searched for segments of the query sequence that matched more than one sequence, as such segments may comprise new motifs. New motifs were searched for in the databases using the programs DBSITE (Claverie, 1993; J.-M.Claverie, personal communication) and PROPAT (Rohde and Bork, 1993).

Generation of families and multiple alignment

To exploit the possible transitivity of similarity (if A is similar to B and B is similar to C, then A may be similar to C), sequences identified as similar to a query protein from chromosome III were subjected to new BLASTP and FASTA searches, unless they belonged to established sequence families. If new significant hits ($P < 10^{-4}$ for BLAST or opt > 150 for FASTA) were obtained, the procedure was repeated iteratively until the putative new family was completed. For such putative families, multiple alignments were generated and analyzed in detail in order to make functional predictions based on the observed conservation pattern.

For multiple sequence alignments, the MACAW (Schuler *et al.*, 1991), OPTAL (Gorbalenya *et al.*, 1989), CLUSTALV (Higgins *et al.*, 1992) or MaxHom (C.Sander and R.Schneider, unpublished) programs were used. In order to characterize sequence families further, PROFILES and/or local motifs were generated and used for additional database searches (Gribskov *et al.*, 1987; Rohde and Bork, 1993).

Additional structural analysis

Sequences homologous to proteins of known three-dimensional structure were identified by lookup in the HSSP database of structure–sequence alignments (Sander and Schneider, 1993). Putative transmembrane segments in the ORFs were predicted using the algorithms of Eisenberg *et al.* (1984) and Rao and Argos (1986), as implemented in the PCGENE package (Moore *et al.*, 1988), respectively. Gene products were considered to be probable membrane proteins only if both algorithms predicted at least one transmembrane helix.

Acknowledgements

E.V.K. is grateful to Dr J.C.Wootton for his input at the initial stages of this project, to Dr J.-M.Claverie for providing the DBSITE program, to Dr R.L.Tatusov for programming, and to Dr M.S.Boguski for critical reading of the manuscript. P.B. and C.S. thank their colleagues in the Protein Design Group at EMBL for providing their software tools.

Note

Selected alignments corresponding to sequence similarities reported here are available by anonymous ftp (file transfer protocol) from ftp.embl-heidelberg.de in the directory /pub/databases/protein_extras/yeast and from ncbi.nlm.nih.gov in the directory pub/koonin/yeast.

References

Adams,M.D., Kerlavage,A.R., Fields,C. and Venter,J.C. (1993) *Nature Genet.*, **3**, 266–272.

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

Amakasu,H., Suzuki,Y., Nishizawa,M. and Fukusawa,T. (1993) *Genetics*, **134**, 675–683.

Bairoch,A. (1993) *Nucleic Acids Res.*, **21**, 3097–3103.

Bairoch,A. and Boeckmann,B. (1993) *Nucleic Acids Res.*, **21**, 3093–3096.

Barker,W.C., George,D.G., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1993) *Nucleic Acids Res.*, **21**, 3089–3092.

Benson,D., Lipman,D.J. and Ostell,J. (1993) *Nucleic Acids Res.*, **21**, 2963–2965.

Boguski,M.S., Lowe,T.M.J. and Tolstoshev,C. (1993) *Nature Genet.*, **4**, 332–333.

Bork,P. (1993) *Proteins Struct. Funct. Genet.*, **17**, 363–374.

Bork,P., Ouzounis,C., Sander,C., Scharf,M., Schneider,R. and Sonnhammer,E. (1992a) *Protein Sci.*, **1**, 1677–1690.

Bork,P., Ouzounis,C., Sander,C., Scharf,M., Schneider,R. and Sonnhammer,E. (1992b) *Nature*, **358**, 287.

Bork,P., Sander,C. and Valencia,A. (1993) *Protein Sci.*, **2**, 31–40.

Claverie,J.-M. (1993) *Nature*, **364**, 19–20.

Collart,M.A. and Struhl,K. (1993) *EMBO J.*, **12**, 177–186.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl. 3, pp. 345–352.

Edman,J.C., Goldstein,A.L. and Erbe,J.G. (1993) *Yeast*, **9**, 669–675.
Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) *J. Mol. Biol.*, **179**, 125–142.
Franco,L., Jimenez,A., Demolder,J., Molemans,F., Fiers,W. and Contreras,R. (1991) *Yeast*, **7**, 971–979.
Gish,W. and States,D.J. (1993) *Nature Genet.*, **3**, 266–272.
Goffeau,A., Nakai,K., Slonimski,P. and Risler,J.-L. (1993a) *FEBS Lett.*, **325**, 112–117.
Goffeau,A., Slonimski,P., Nakai,K. and Risler,J.-L. (1993b) *Yeast*, **9**, 691–702.
Gorbalenya,A.E., Blinov,V.M., Donchenko,A.P. and Koonin,E.V. (1989) *J. Mol. Evol.*, **28**, 256–268.
Green,P., Lipman,D.J., Hillier,L., Waterston,R., States,D. and Claverie,J.-M. (1993) *Science*, **259**, 1711–1716.
Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
Henikoff,S. and Henikoff,J. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) *Comput. Appl. Biosci.*, **8**, 189–191.
Hjelmstad,R.H. and Bell,R.M. (1991) *J. Biol. Chem.*, **266**, 5094–5104.
Janitor,M. and Subik,J. (1993) *Curr. Genet.*, **24**, 307–312.
Karlin,S. and Altschul,S.F. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 5873–5877.
Kern,L. (1990) *Nucleic Acids Res.*, **18**, 5279.
Koonin,E.V. (1993) *J. Gen. Virol.*, **74**, 733–740.
Lalo,D., Stettler,S., Mariotte,S. and Slonimski,P.P. (1993) *C. R. Acad. Sci. Paris (Sciences de la vie)*, **316**, 367–373.
Mehta,P.K. and Christen,P. (1993) *Eur. J. Biochem.*, **211**, 373–376.
Moore,J., Engelberg,A. and Bairoch,A. (1988) *BioTechniques*, **6**, 566–572.
Navarre,C., Ghislain,M., Leterme,S., Ferroud,C., Dufour,J.-P. and Goffeau,A. (1992) *J. Biol. Chem.*, **267**, 6425–6428.
Nikawa,J.-I., Kodaki,T. and Yamashita,S. (1987) *J. Biol. Chem.*, **262**, 4876–4881.
Oliver,S.G. *et al.* (1992) *Nature*, **357**, 38–46.
Ouzounis,C. and Sander,C. (1993) *FEBS Lett.*, **322**, 159–164.
Pang,A.S., Nathoo,S. and Wong,S.L. (1991) *J. Bacteriol.*, **173**, 46–54.
Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Rao,M.J.K. and Argos,P. (1986) *Biochim. Biophys. Acta*, **869**, 197–214.
Rice,C.M., Fuchs,R., Higgins,D.G., Stoehr,P.J. and Cameron,G.N. (1993) *Nucleic Acids Res.*, **21**, 2967–2971.
Rohde,K. and Bork,P. (1993) *Comput. Appl. Biosci.*, **9**, 183–189.
Rothe,B., Rothe,B., Roggentin,P. and Schauer,R. (1991) *Mol. Gen. Genet.*, **226**, 190–197.
Sander,C. and Schneider,R. (1993) *Nucleic Acids Res.*, **21**, 3105–3109.
Savakis,C. and Doelz,R. (1993) *Science*, **259**, 1677–1678.
Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) *Proteins Struct. Funct. Genet.*, **9**, 180–190.
Sor,F., Cheret,G., Fabre,F., Faye,G. and Fukuhara,H. (1992) *Yeast*, **8**, 215–222.
Tanaka,S. and Isono,K. (1993) *Nucleic Acids Res.*, **21**, 1149–1153.
Tatusov,R.L. and Koonin,E.V. (1994) *Comput. Appl. Biosci.*, in press.
Tomoyasu,T., Yuki,T., Morimura,S., Mori,H., Yamanaka,K., Niki,H., Hiraga,S. and Ogura,T. (1993) *J. Bacteriol.*, **175**, 1344–1351.
Van Dyck,E., Foury,F., Stillman,B. and Brill,S.J. (1992) *EMBO J.*, **11**, 3421–3430.
Van Dyck,L., Purnelle,B., Skala,J. and Goffeau,A. (1992) *Yeast*, **8**, 769–776.
Voytas,D. and Boeke,J. D. (1992) *Nature*, **358**, 717.
Wickner,R.B., Fujimura,T. and Esteban,R. (1986) *Basic Life Sci.*, **40**, 149–163.
Wootton,J.C. and Federhen,S. (1993) *Comput. Chem.*, **17**, 149–163.
Yang,W., Ni,L. and Sommerville,R. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 5796–5800.
Zhu,G., Muller,E.G.D., Amacher,S.L., Northrop,J.L. and Davis,T.N. (1993) *Mol. Cell. Biol.*, **13**, 1779–1787.

Received on August 12, 1993; revised on October 29, 1993