Protein Science (1992), *1*, 1677–1690. Cambridge University Press. Printed in the USA.
Copyright © 1992 The Protein Society

Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III

PEER BORK, CHRISTOS OUZOUNIS, CHRIS SANDER, MICHAEL SCHARF, REINHARD SCHNEIDER, AND ERIK SONNHAMMER

European Molecular Biology Laboratory, Heidelberg, Germany

(RECEIVED May 29, 1992; REVISED MANUSCRIPT RECEIVED August 4, 1992)

Abstract

With the completion of the first phase of the European yeast genome sequencing project, the complete DNA sequence of chromosome III of *Saccharomyces cerevisiae* has become available (Oliver, S.G., et al., 1992, *Nature 357*, 38–46). We have tested the predictive power of computer sequence analysis on the 176 probable protein products of this chromosome, after exclusion of six problem cases. When the results of database similarity searches are pooled with prior knowledge, a likely function can be assigned to 42% of the proteins, and a predicted three-dimensional structure to a third of these (14% of the total). The function of the remaining 58% remains to be determined. Of these, about one-third have one or more probable transmembrane segments. Among the most interesting proteins with predicted functions are a new member of the type X polymerase family, a transcription factor with an N-terminal DNA-binding domain related to GAL4, a "fork head" DNA-binding domain previously known only in *Drosophila* and in mammals, and a putative methyltransferase. Our analysis increased the number of known significant sequence similarities on chromosome III by 13, to now 67. Although the near 40% success rate of identifying unknown protein function by sequence analysis is surprisingly high, the information gap between known protein sequences and unknown function is expected to widen and become a major bottleneck of genome projects in the near future. Based on the experience gained in this test study, we suggest that the development of an automated computer workbench for protein sequence analysis must be an important item in genome projects.

Keywords: computer methods; genome projects; prediction of protein function; prediction of protein structure; protein sequence analysis

The yeast genome consists of 16 chromosomes containing approximately 14 megabases (Oliver et al., 1992). A collaborative network of 35 research groups in the European yeast sequencing project has now sequenced the entire chromosome III, covering 315 kilobases, 2.3% of the entire yeast genome (Oliver et al., 1992). As this is the first report of a complete eukaryotic chromosome and the longest continuous stretch of DNA ever sequenced, it represents not only a unique rich source of information, but also a challenge for computer sequence analysis. We have therefore taken the 182 open reading frames (ORFs) reported by Oliver et al. and asked a number of technical and biological questions. To what extent can sequence analysis help in the identification of protein function? Is there a single preferred method of database similarity searches for this purpose? What is the total effort involved in performing, analyzing, and reporting comprehensive database searches? How many of the proteins have a known or predicted function? How many are homologous to a protein of known three-dimensional structure? And, extrapolating from the current level of data bases and sequence analysis methods, what can we expect from the entire sequence of the yeast genome?

Reprint requests to: Protein Design Group, EMBL, Meyerhofstrasse 1, D-6900 Heidelberg, Germany.

Results

Sequence analysis predicts the biological function of 67 of the 176 probable proteins (Fig. 1) with high or reasonably high reliability, and the three-dimensional structure of 25 of these. Adding the seven proteins for which the biological function had been determined previously by genetic or biochemical experiment, a biological function can now be assigned to 42% of the proteins of yeast chromosome III. The predictions are, in part, the result of obvious hits in database searches using standard software,

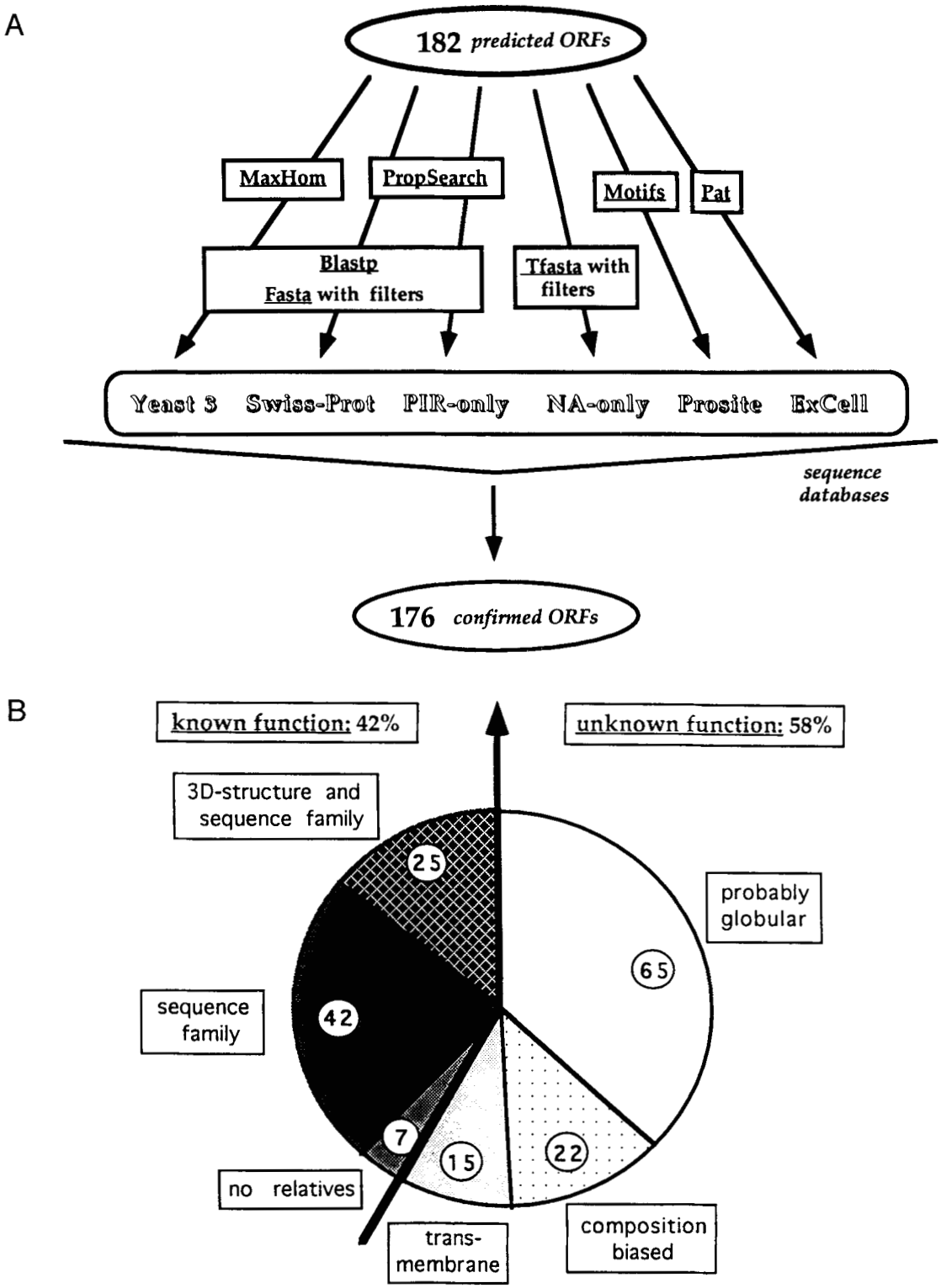

Fig. 1. A: Databases and search tools. The following databases were scanned: (1) Swiss-Prot (Bairoch & Boeckmann, 1991; release 21, 23,742 proteins). (2) PIR-only, containing all sequences in the PIR (Barker et al., 1991; Protein Identification Resource, release 31) database that are not identical in Swiss-Prot (an additional 10,501 proteins), as produced by Peter Rice at EMBL (pers. comm.). (3) The set of 182 yeast chromosome III open reading frames (ORFs) as received by electronic mail from the Martinsried Institute for Protein Sequences on behalf of the European yeast chromosome III collaboration (Oliver et al., 1992). (4) NA-only, a subset of the EMBL Data Library's nucleotide sequence collection, translated in all six reading frames. The subset contains all entries for which there is no explicit EMBL-to-SwissProt or SwissProt-to-EMBL pointer (DR line in either database). The intention is to include all nucleic acid sequence material that possibly has not yet been processed for amino-acid sequences (an additional 48,434 entries, i.e., 290,604 translated amino acid sequences, most of which contain no substan-

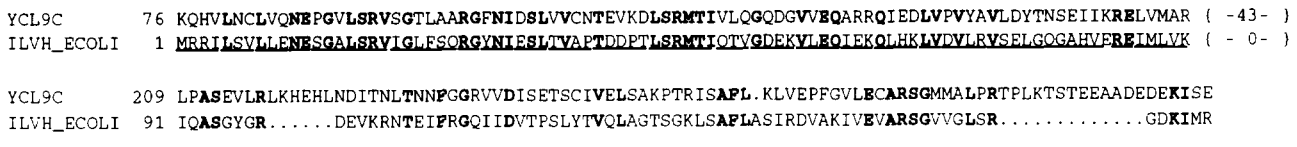

Fig. 2. Optimal alignment of YCL9c with acetolactate synthase. The alignment between the yeast ORF YCL9c and the small subunit of acetolactate synthase from *Escherichia coli* given by Fasta (underlined; identities/length = 40%/90) can be extended considerably using Bestfit (36%/208). Most of the conserved residues (bold) are also present in two other prokaryotic small subunits of acetolactate synthases (data not shown).

and, in part, the result of more detailed analysis near the "twilight zone" of sequence similarity, using a combination of multiple sequence alignment, pattern searches, and other methods (Table 1; see Materials and methods).

The most interesting "twilight zone" similarities and their functional interpretation

A number of the detected similarities warrant further discussion, because the similarity with one or more database proteins is not obvious, has not yet been described or is particularly interesting. Exhaustive discussion of all similarities is not feasible. We have selected for discussion some putative enzymes and DNA associated or regulatory proteins.

A eukaryotic protein related to the regulatory subunit of prokaryotic acetolactate synthases

Protein YCL9c has remote similarity to the small regulatory subunits of prokaryotic acetolactate synthases, the first of this type in eukaryotes (Fig. 2). For the large subunits of prokaryotic acetolactate synthases eukaryotic counterparts are known (Wiersma et al., 1989). The large and small subunits appear to be nonhomologous (Wek et al., 1985) . The small subunits of the prokaryotic enzymes are considerably shorter, about 90–170 amino acids in length, compared to the 309 amino acids of protein YCL9c . The similarity is distributed in three "boxes," with large intervening insertions in the eukaryotic enzyme subunit. The strongest contiguous stretch has 40% identical residues, out of 90 (Fig. 2).

A type X DNA polymerase

At least four distinct families of DNA polymerases have been identified so far (Ito & Braithwaite, 1991) . The smallest group, called type X, includes DNA polymerase beta and DNA nucleotidyl exotransferase. Both proteins are relatively small compared to other types of DNA polymerases. So far, only mammals are known to have this type of polymerase. Protein YCR14c appears to be related

Fig. 1. (*continued*)
tial ORFs). (5) Prosite, a database of amino acid sequence patterns characteristic of structural and functional classes (release 8.1, with 530 patterns) (Bairoch, 1991). (6) ExCell, a library of selected sequence patterns characteristic of extracellular domains of mosaic proteins (67 patterns, as collected by one of us; Bork, 1991). For complete searches in each of the databases, the appropriate search software was used. These were, following the numbering of databases given above: (1-3) Blastp (Altschul et al., 1990) and Fasta (Pearson & Lipman, 1988) with Fasta_Filter (R.S. & C.S., unpubl.); (4) Tfasta (Pearson & Lipman, 1988); (5) Motifs in the GCG software package (Devereux et al., 1984); (6) Pat (Bork & Grunwald, 1990). Two additional special purpose tools were used: (7) The program MaxHom (R.S. & C.S., unpubl.), an implementation of the Smith and Waterman (1981) algorithm with linear gap weights (Gotoh, 1982), was used to find internal repeats in each of the proteins of the yeast chromosome III data set and to refine pair alignments. The program reports the K best alignments for each pair comparison according to Waterman and Eggert (1987). (8) The program PropSearch (Sültemeyer, 1988) was used to search the Swiss-Prot database for proteins similar in compositional bias. The fact that no internal repeats were detected in the search may reflect a too stringent significance cutoff—an example of a possibly significant fivefold internal repeat is in the ORF YCR84c (Duronio et al., 1992; van der Voorn & Ploegh, 1992). **B:** Information clock of yeast chromosome III proteins. Information accumulated to date by all methods, experimental and theoretical. Some of the information is direct, e.g., the result of gene disruption experiments, and some of it is indirect, e.g., the result of sequence database searches by computer programs. Information content increases counterclockwise. The principal division is between known and unknown biological function. The categories are approximate, but give an impression of the current state of the art. The information does not necessarily cover the entire protein, as in a number of cases the sequence similarity is restricted to a single domain (Table 1). For 6 out of the 182 predicted ORFs (YCL21w, YCL41c, YCL65w, YCR41w, YCR13c, YCR70w) not included in this figure, there is some doubt as to whether they correspond to protein products. Their protein sequences either match translations of known regulatory elements; or they overlap with other ORFs on chromosome III for which a clear sequence similarity with other proteins in the database was shown—the latter are therefore accepted. For example, YCR13c is completely included in the YCR12w (coding for phosphoglycerate kinase) on the complementary strand (Skala et al., 1992). In addition, three ORFs (YCL75w, YCR69w, YCR9) appear to contain minor sequencing errors (frameshifts, incorrectly predicted start codon).

Table 1. *Sequence analysis of the proteins of yeast chromosome III*[a]

ORF	Length	[Code] Family	New	Closest	%Id/len	Dis	Opt	*p*	3D	Domains/sites
Significant hits (according to BLASTP: $p < 1.0e-10$)										
Enzymes										
YCL18w	364	[LEU3_YEAST] Isopropylmalate dehydrogenases		LEU3_KLULA	86%/360	60.8	1,524	2.0*e*−228	9ICD	
YCL24w	816	Protein kinases		SNF1_YEAST	30%/504	5.6	650	4.7*e*−46	1CPK	[N–550] only
				NIM1_SCHPO	40%/349	14.7	660	3.7*e*−51		[550–C] ins
YCL30c	799	[HIS2_YEAST] HIS−cyclohydrolases		HIS2_NEUCR	41%/688	20.9	1,312	4.2*e*−88		
YCL40w	500	[HXKG_YEAST] Glucokinases/hexokinases		S15885(P)	36%/268	11.0	345	2.4*e*−40	2YHX	
				HXKA_YEAST	34%/384	9.3	589	6.1*e*−30		
YCL43c	522	[JX0182 (P)] Protein disulfide isomerases (PDI)		ER72_MOUSE	58%/403	7.2	613	4.6*e*−59	2TRX	S:ER_target:519
YCL50c	321	[APA1_YEAST] 5′,5‴-P-1,P-4 tetra-P phosphorylases		APA2_YEAST	59%/321	34.1	1,056	1.0*e*−76		
YCL57w	712	Soluble metalloprotease		A36165(P)	36%/561	10.7	1,012	2.9*e*−58		S:zinc_protease:498
YCL64c	360	[SDCSERTHR(E)] L-Serine dehydratases		SDHL_YEAST	53%/330	28.2	973	2.7*e*−120	Yes	
				SDHL_HUMAN	34%/263	9.4	360	8.3*e*−21		
YCL9c	309	Prokaryotic acetolactate synthases, small subunit	New	ILVH_ECOLI	40%/90	15.2	192	1.3*e*−16		
YCR105w	361	Alcohol dehydrogenases		ADH_SCHPO	30%/342	4.7	335	2.6*e*−14	7ADH	
YCR12w	416	[PGK_YEAST] Phosphoglycerate kinases		PGK_KLULA	82%/416	57.4	1,643	7.7*e*−246		
YCR24c	492	Class II tRNA synthases (probably ASP)		SYN_ECOLI	32%/295	6.7	399	3.5*e*−50	Yes	
				SYK2_ECOLI	18%/471	−7.2	123	>1*e*−5		
YCR36w	333	Ribokinase (other prokaryotic sugar kinases)	New	RBSK_ECOLI	38%/96	12.7	150	1.5*e*−10		
YCR45c	491	Subtilisin family		PRTB_YEAST	37%/348	12.3	430	7.8*e*−32	1SBC	
YCR53w	514	[THRC_YEAST] Threonine synthases		THRC_CORGL	39%/465	14.6	760	8.2*e*−27	Yes	
YCR5c	460	[CISZ_YEAST] Citrate synthases		CYSY_PIG	59%/437	33.8	1,432	1.1*e*−153	4CTS	
YCR69w (YCR70w)[b]	170	Peptidyl-prolyl-cis-trans isomerases	New	CYPH_CANAL	37%/122	12.1	176	2.4*e*−13		
YCR8w	603	[SC4(E)] Protein kinases		KCCG_RAT	31%/283	6.3	277	4.1*e*−12	1CPK	[311–592] only
YCR73c	1,314	Protein kinases		JQ1118(P)	34%/285	8.9	321	1.4*e*−19	1CPK	[1,034–1,293] only
YCR91w	726	Protein kinases		KS6_HUMAN	37%/312	12.4	355	7.8*e*−31	1CPK	[319–625] only
				B30311(P)	42%/184	17.0	451	5.2*e*−88		
DNA-associated or regulatory proteins										
YCL11c	426	Poly(A) binding protein		PABP_XENLA	28%/195	8.4	225	2.0*e*−11		
YCL17c	497	Bacterial NIFS genes		NIFS_ANASP	44%/383	18.8	778	4.5*e*−41		[N–100] ins
YCL74w	308	Retroelement, COPIA-like protein		S05465(P)	31%/308	5.6	405	3.4*e*−14		
				POLX_TOBAC	34%/308	9.1	517	7.3*e*−26		
YCR63w	157	G-cycle-specific protein		G10_XENLA	52%/155	26.8	480	6.7*e*−44	Yes	Zinc finger [105–122]
YCR65w	532	HNF3/fork head transcription factors		FKH_DROME	24%/295	−0.7	289	1.9*e*−20		DNA-binding domain
		HNF3 rat gamma		A39533(P)	51%/88	12.1	286	5.1*e*−24		[110–230] only
YCR92c	1,047	DNA repair proteins		DUG_HUMAN	31%/947	5.9	1,120	2.7*e*−45		S:ATP-binding:820

Not classified										
YCL19w	1,347	[B23496(P)] Transposon B gene family, related to pol genes	New	COPI_DROME POLX_TOBAC	18%/505 20%/393	−7.9 −4.7	248 304	3.2*e*−13 3.0*e*−16		[170-780] only
YCL20w	438	[A23496(P)] Transposon A gene family	New	YTY1_YEAST	49%/438	24.4	1,080	3.0*e*−102		
YCL25c	633	Amino acid permeases, transporter family		GAP1_YEAST	47%/630	22.3	1527	9.3*e*−125		
YCL35c	110	Glutaredoxins		GLRX_YEAST	64%/106	39.4	396	3.2*e*−52		
YCL48w	463	Sporulation-specific protein		SPS2_YEAST	44%/434	19.2	941	4.3*e*−89		
YCR11c	1,049	Multidrug transporter superfamily		WHIT_DROME	27%/445	2.2	512	5.6*e*−17		
YCR18c	225	[YH26_YEAST] Hypothetical protein		YH16_YEAST	94%/126	69.6	577	7.4*e*−87	Yes	GATA-type zinc finger [168-193]
YCR27c	209	[SCTRNGLN(E)] RAS protein family		RAPB_HUMAN	34%/192	9.1	296	4.5*e*−24	5P21	S:farnelysation:206
YCR2c	322	GTP/ATP binding protein		JU0319(P) MMH5RNA(E)	40%/304 38%/322	15.0 13.9	610 660	3.7*e*−36 nd	Yes	
YCR31c	137	[RS14_YEAST] 40S ribosomal proteins		RS14_DROME	82%/133	57.0	530	1.8*e*−78		
YCR52w	483	Mitochondrial targeting protein		SCPCPEIIT(E)	30%/353	4.9	435	nd		[150-C] only
YCR57c	439	Domains, G-proteins, beta subunit		GBB2_HUMAN	32%/160	7.1	266	1.7*e*−17		[170-330] only
YCR67c	1,065	Membrane glycoprotein (ER/Golgi protein transport)		SC12_YEAST	45%/388	20.6	889	2.2*e*−66		S:ER_target:1,062
YCR72c	514	G-protein beta subunits	New	PRO4_YEAST TUP1_YEAST	23%/278 32%/110	−1.8 7.0	191 183	5.7*e*−10 3.0*e*−28		[200-500] only
YCR83w	127	Thioredoxins (but not I, 1, 2 family)		THI2_YEAST	47%/98	22.1	272	1.4*e*−27	3TRX	
YCR84c	713	[TUP1_YEAST] Regulator of glucose repression, G-proteins beta subunit		PRO4_YEAST GBB2_HUMAN	19%/487 34%/163	−6.5 9.2	206 201	2.3*e*−10 1.9*e*−14		[450-660] only
YCR89w	1,609	Agglutinin core protein		S17031(P)	25%/675	0.1	278	4.1*e*−105		[540-1,200] only
Significant hits in the "twilight zone"										
Enzymes										
YCR47c	275	ApoMet-methyltransferases	New	GLMT_RAT	26%/301	8.8	102	>1*e*−5		
YCR64c	136	Carboxypeptidases dipep-peptidase IV	New	CBP8_HUMAN DPP_LACLA	27%/88 25%/105	2.5 0.0	73 71	>1*e*−5 >1*e*−5		
DNA-associated or regulatory proteins										
YCL33c	168	Repressor of pilin promotor	New	PILB_NEIGO	33%/110	7.9	163	2.5*e*−9		[N-398] del
YCL75w	146	Pol-like protein	New	S00954(P)	40%/74	14.6	124	>1*e*−5		[N-70] ins, [90-C] del
YCR104w	124	Glucose repressor/cold shock inducible	New	SRP1_YEAST SCTIPI(E)	27%/115 27%/106	2.2 2.5	123 143	>1*e*−5		[114-C] del (S-rich)
YCR106w	832	Gal4-like DNA/Zn-binding domain	New	CYP1_YEAST GAL4_YEAST	45%/47 19%/168	11.3 −5.8	135 111	>1*e*−5 >1*e*−5	Yes	GAL4-like Zn/DNA-binding domain
YCR14c	582	Type X DNA polymerases	New	DPOB_RAT	26%/393	−0.4	189	>1*e*−5		[N-65] only
YCR66w	487	DNA repair protein RADZ (RAD18)		RPT1_MOUSE	27%/81	2.4	140	>1*e*−5	Yes	S:zinc_finger:28
Not classified										
YCL66w (YCR40w)[c]	175	[MAT1_YEAST] Mating hormone alpha		MATA_NEUCR	35%/58	4.8	126	>1*e*−5		
YCL67c (YCR39c)[c]	210	[MTA2_YEAST] Mating factor, P polypeptide		MTPI_SCAPO	34%/62	5.3	107	>1*e*−5	1HDD	S:homeobox:129

(*continued*)

Table 1. *Continued*

ORF	Length	[Code] Family	New	Closest	%Id/len	Dis	Opt	*p*	3D	Domains/sites
Significant hits in the "twilight zone" (*continued*)										
Not classified										
YCL68c (YCR38c)[c]	190	[A39933(P)] Exchange factor BUD5, related to C-term of CC25 fam						>1*e*−5		
YCL69w	458	Multidrug resistance proteins		TCR2_BACSU	20%/313	−5.4	185	>1*e*−5		
YCR23c	611	Multidrug resistance proteins	New	TCR1_ECOLI	28%/150	3.2	186	>1*e*−5		
YCR26c	743	Mammalian PC1 plasma cell membrane protein phosphodiesterase family	New	PC1_HUMAN	38%/129	13.2	175	1.1*e*−9		
				PPD1_BOVIN	38%/66	10.3	117	5.6*e*−7		[150–300] only
YCR32w	2,167	[S15052(P)] hypothetical protein related to C-terminal "CDC4"-like human fragment	New	HSCDC4A(E)	49%/316	24.3	852	nd		[N–336] ins
YCR107w	363	Auxin-regulated protein from tobacco		NTAUX115(E)	22%/327	−3.3	225	nd		
YCR88w	592	[ABP1_YEAST] Actin-binding protein		HS1_HUMAN	26%/286	1.1	224	3.5*e*−9	Yes	SH3 dom.[537–589] only
YCR96c	119	[MTA2_YEAST] Mating type alpha-2, short form		HME2_HUMAN	33%/71	5.9	81	>1*e*−5	1HDD	S:homeobox:63
YCR97w	126	[MTA1_YEAST] Mating type A1, MTA0,		HM43_CHICK	27%/92	2.4	106	>1*e*−5	1HDD	S:homeobox:70
YCR98c	518	Sugar transporter/symporter	New	A40260(P)	25%/179	0.3	137	1.1*e*−7		

[a] ORF: Name of the predicted open reading frame (Oliver et al., 1992). [Code]: Sequence database identifier, if available from Swiss-Prot (default), PIR (P), or EMBL (E). Family: Functional protein family. New: Similarity not reported by Oliver et al. (1992). For four of the ORFs (Yc120w, Ycr36w, Ycr69w/Ycr70w, Ycr98c), similarities to other proteins have already been noted (Warmington et al., 1985; Thierry et al., 1990; Franco et al., 1991; Sor et al., 1992). Closest: Database code of the closest relatives, based on Blastp or Fasta searches. %Id/len: Percent amino acid identity/length of the alignment. Dis: Distance (excess) of %id above the length-dependent threshold for structural homology given in percentage points (Sander & Schneider, 1991). Opt: Optimized Fasta score (Pearson & Lipman,1988). *p*: Probability of random occurrence from Blastp (Altschul et al., 1990). Database size was 35,268,075 amino acid residues (Swiss-Prot plus PIR-only). The PAM120 matrix of substitution probabilities was used. If the Blastp probability is below 1.0*e*−10 and the protein is not composition biased, the respective hits are immediately treated as significant; nd, not determined. 3D: Protein Data Bank (PDB) (Bernstein et al., 1977) identifier of the closest homologue of known three-dimensional structure, if available; Yes, 3D structure of a homologue known, but the coordinates of the 3D structure have not been deposited in the Protein Data Bank. Domains/sites: Residue range, in brackets, of the part of the yeast protein with sequence similarity to a known domain (N, N-terminus; C, C-terminus); only, similarity is restricted to the domain; ins (del), the yeast sequence has a substantial extension (deletion) relative to the proteins containing the similar domain. S:, sequence contains an exact match with a motif from the Prosite library (Bairoch, 1991); name of the Prosite pattern and sequence position.

[b] When the two ORFs are corrected (cut and fused), they represent a member of the PPI family.

[c] The three genes occur twice on chromosome III. In addition, two other mating factors are present on the chromosome.

to this group. In addition to the overall similarity, some essential residues involved in primer binding (Date et al., 1991) are conserved (Fig. 3). With the discovery of this type of polymerase in yeast one may expect that it will also be found in a variety of other eukaryotic species.

The methyltransferase superfamily

Protein YCR47c is a putative methyltransferase. A sequence pattern defined on the basis of a few database hits defines a subfamily of methyltransferases that have a common functional site. They all belong to a large superfamily of methyltransferases for various substrates with *S*-adenosylmethionine as donor of a methyl group. The sequence segment corresponding to the pattern therefore is thought to be part of the *S*-adenosylmethionine binding site (Ingrosso et al., 1989; Klimasauskas et al., 1989), with slightly different conservation patterns in functionally distinct subfamilies (R. Roberts, pers. comm.). From the multiple sequence alignment, we have defined a 23-residue-long pattern: `tttxhh[NDE]hGtGxGhhxxxhhxxh`, where t = polar or turn forming, h = hydrophobic, and x = any amino acid, with alternatives at one position in brackets. Searching the protein database with this pattern and allowing a small number of deviations (see Materials and methods), we also detect several epimerases, adenosyl homocysteinases, and various proteins of undetermined function (Fig. 4). The implication is that these proteins may bind *S*-adenosylmethionine as well.

A GAL4-type transcriptional activator

Protein YCR106w is homologous to a 60-residue Zn-containing a (binuclear cluster) DNA-binding domain of known three-dimensional structure (Kraulis et al., 1992; Marmorstein et al., 1992). The domain is common to fungal transcription factors such as yeast transcriptional activator *GAL*4, *CYP*, *MAL*, and *PPR* (Fig. 5). All conserved cysteines and conserved positively charged residues are present. The sequence differs from the consensus motif of the GAL4-like domain in only one position (Prosite entry: Zn2_CY6_FUNGAL; Bairoch, 1991).

A "fork head" DNA-binding domain now also in yeast

Protein YCR65w contains the 110-residue DNA-binding domain known from the region-specific *Drosophila* transcription regulator *fkh* (fork head), the hepatocyte-specific rat transcription factor HNF-3A (Weigel & Jäckle, 1990), and interleukin binding factor (Li et al., 1991). The occurrence of this DNA-binding domain in yeast suggests interesting analogies between transcriptional control in yeast and developmental control in higher eukaryotes.

A new family of stress-induced proteins

The 110-residue protein YCR104w is similar to the N-terminal domains of the yeast gene products *srp1*, induced by glucose, and of *tip1*, inducible by cold shock. This domain precedes a longer serine-rich region in the much larger *srp1* and *tip1* proteins. All three protein share a putative signal sequence. Hybridization experiments had already indicated the presence of homologues of *srp1* and *tip1* in yeast (Marguet et al., 1988; Kondo & Inouye, 1991). Protein YCR104w now appears to be one of these homologues, although it lacks the repetitive serine-rich region.

```
YCR14C      194 ALKRLTK.KYEIEGEKFRARSYRLAKQSMENCDFNVRSGEEAHTKLRNIGPSIAKKIQVILDTGVLPGLNDSVGLE..DKLKYFKNCYGIGSEIAKRWNL
DPOB_HUMAN    1 MLTEIANFEKNVSQAIHKYNAYRKAASVIAKYPHKIKSGAEAK.KLPGVGTKIAEKIDEFLATGKLRKLEKIRQDDTSSSINFLTRVSGIGPSAARKFVD
DPOB_RAT     17 MLVELANFEKNVSQAIHKYNAYRKAASVIAKYPHKIKSGAEAK.KLPGVGTKIAEKIDEFLATGKLRKLEKIRQDDTSSSINFLTRVTGIGPSAARKLVD
TDT_BOVIN   182 AFEILAE.NSEFKENEVSYVTFMRAASVLKSLPFTIISMKDTE.GIPCLGDKVKCIIEEIIEDGESSEVKAVLNDERYQSFKLFTSVFGVGLKTSEKWFR
TDT_HUMAN   171 AFDILAE.NCEFRENEDSCVTFMRAASVLKSLPFTIISMKDTE.GIPCLGSKVKGIIEEIIEDGESSEVKAVLNDERYQSFKLFTSVFGVGLKTSEKWFR
TDT_MOUSE   171 ALDILAE.NDELRENEGSCLAFMGASSVLKSLPFPITSMKDTE.GIPCLGDKVKSIIEGIIEDGESSEAKAVLNDERYKSFKLFTSVFGVGLKTAEKWFR

YCR14C      291 LNFESFCVAAKKDPEEFVSDWTILFGWSYYDDWLCKMSRNECFTHLKKVQKALRGIDPECQVELQGSYNRGYSKCGDIDLLFFKP.FCNDTTELAKIMET
DPOB_HUMAN  100 EGIKTLEDLRKNED.KLNHHQRI..GLKYFGDFEKRIPREEMLQMQDIVLNEVKKVDSEYIATVCGSFRRGAESSGDMDVLLTHPSFTSESTKQPKLLHQ
DPOB_RAT    116 EGIKTLEDLRKNED.KLNHHQRI..GLKYFEDFEKRIPREEMLQMQDIVLNEVKKLDPEYIATVCGSFRRGAESSGDMDVLLTHPNFTSESSKQPKLLHR
TDT_BOVIN   280 MGFRSLSKIMSDKTLKFTKMQKA..GFLYYEDLVSCVTRAEAEAVGVLVKEAVWAFLPDAFVTMTGGFRRGKKIGHDVDFLITSPGSAEDEE.Q..LLPK
TDT_HUMAN   269 MGFRTLSKVRSDKSLKFTRMQKA..GFLYYEDLVSCVTRAEAEAVSVLVKEAVWAFLPDAFVTMTGGFRRGKKMGHDVDFLITSPGSTEDEE.Q..LLQK
TDT_MOUSE   269 MGFRTLSKIQSDKSLRFTKMQKA..GFLYYEDLVSCVNRPEAQAVSMLVKEAVVTFLPDALVTMTGGFRRGKMTGHDVDFLITSPEATEDEEQQ..LLHK
                                                                                  *       * *

YCR14C      390 LCIKLYKDGYI ( -99- ) RLDFFCCKWDELGAGRIHYTGSKEYNRWIRILAA.QKGFKLTQHGL ( - 6- ) LESFNERRIFELLNLKYAEPEHR
DPOB_HUMAN  197 VVEQLQKVHFI ( -29- ) RIDIRLYPKDQYYCGVLYFTGSDIFNKNMRAHAK.EKGFTINEYTI ( -12- ) LPVDSEKDIFDYIQWKYREPKDR
DPOB_RAT    213 VVEQLQKVRFI ( -29- ) RIDIRLIPKDQYYCGVLYFTGSDIFNKNMRAHAL.EKGFTINEYTI ( -12- ) LPVDSEQDIFDYIQWRYREPKDR
TDT_BOVIN   375 VINLWEKKGLL ( -56- ) RVDLVMCPYENRAFALLGWTGSRQFERDIRRYATHERKMMLDNHAL ( - 8- ) LKAESEEEIFAHLGLDYIEPWER
TDT_HUMAN   364 VMNLWEKKGLL ( -56- ) RVDLVLCPYERRAFALLGWTGSR.FERDLRRYATHERKMILDNHAL ( - 8- ) LKAESEEEIFAHLGLDYIEPWER
TDT_MOUSE   365 VTHFWKQQGLL ( -56- ) RVDLVMCPYE.CACALLGWTGSRQFERDLRRYATHERKMMLDNHAL ( -28- ) LEAESEEEIFAHLGLDYIEPWER
```

Fig. 3. Multiple alignment of type X DNA polymerases. YCR14c can be aligned with mammalian DNA polymerases β (DPOB) and DNA nucleotidyl exotransferases (TDT). Only the three most conserved segments are shown. Sequence positions are at left, length of sequence gaps in parentheses. Asterisks mark functionally important residues (Date et al., 1991) that are also present in YCR14c. The Prosite pattern for this class is underlined. Several conserved (bold) charged residues outside the known pattern probably have a functional role.

```
                                                          ttt hh-hGtG Ghh   hh  h  h hh
HIOM_BOVIN  hydroxyindole O-methyltransferase          178 PFPLICDLGGGSGALAKACVSLYPGCRAI
CRTF_RHOCA  hydroxyneurosporen methyltransferase       228 DAKRVMDVGGGTGAFLRVVAKLYPELPLT
CARB_STRTH  RRNA methyltransferase                      74 PGEVVLEVGAGNGAITRELARLCRRVVAY
KSGA_ECOLI  S-adenosylmethionin dimethyltransfer.       37 KGQAMVEIGPGLAALTEPVGERLDQLTVI
MLS1_STAAU  RRNA adenyl-N-6-methyltransferase           30 KQDNVIEIGSGKGHFTKELVKMSRSVTAI
MTPS_PROST  modification methyltranferase PSTI          57 GEHEILDAGAGVGSLTAAFVQNATLNGAK
PIMT_BOVIN  protein-beta-aspart. methyltransferase      77 EGAKALDVGSGSGILTACFARMVGPSGKV
GLMT_RAT    glycine methyltransferase                   56 GCHRVLDVACGTGVDSIMLVEEGFSVTSV

YCR47c      yeast ORF                                   47 PCSFILDIGCGSGLSGEILTQEGDHVWCG

BIOC_ECOLI  protein involved in biotin conversion       42 KYTHVLDAGCGPGWMSRHWRERHAQVTAL
YT37_STRFR  hypoth. protein in transposon TN4556       126 PGESALDLGCGPGTDLGTLAKAVSPSGRV
YAT1_SYNP6  hypoth. protein in the GYRA 5' region       71 GRPRILDAGCGTGVSTDYLAHLNPSAEIT
YFAB_ECOLI  hypoth. 26.6KD protein                      56 FGKKVLDVGCGGGILAESMAREGATVTGL
SAHH_HUMAN  adenosylhomocycteinase                     340 AEGRLVNLGCAMGHPSFVMSNSFTNQVMA
GALE_ECOLI  UDP-glucose-4-epimerase                    254 PGVHIYNLGAGVGNSVLDVVNAFSKACGK
```

Fig. 4. Result of a pattern search with the *S*-adenosylmethionine binding site of methyltransferases. In the preliminary Fasta search, several methyltransferases match a short segment in YCR47c. The segment corresponds to subtype of an *S*-adenosylmethionine binding site found in various methyltransferases (Ingrosso et al., 1989; Klimasauskas et al., 1989). Based on these hits, a property consensus pattern (Bork & Grunwald, 1990) for this region was derived. A sequence database search with the pattern picks up more than 50 different methyltransferases. A few of them are shown in the figure (upper). In addition, *S*-adenosyl homocysteine transferases and UDP-glucose-4-epimerases match the pattern, as well as some hypothetical proteins (lower). Because of the clear separation of the background of nonrelated proteins (data not shown), an *S*-adenosylmethionine binding site is also suggested in these proteins. Each line has the Swiss-Prot database code, protein name, sequence position of the motif, sequence segment. Top line: sequence consensus (capitals, conserved residues; h, hydrophobic; t, turn forming or polar, -, D, N, or E). The most conserved positions are in bold caps.

A regulatory domain common to eukaryotes and prokaryotes

A new regulatory domain common to eukaryotes and prokaryotes is defined by the significant relationship between the yeast protein YCL33c and the C-terminal segment of the PILB protein. PILB is a repressor of pilin promoter activity (Taha et al., 1988). Pilin, in turn, is the major protein of the pili in *Neisseria gonorrhoeae*, which play an important role in virulence by mediating adhesion in the human host (Taha et al., 1988). This similarity of YCL33c to a part of the larger PILB protein may reflect a regulatory mechanism of gene expression common to eukaryotes and prokaryotes.

Functional and structural classes of proteins on chromosome III

The distribution of different protein types on yeast chromosome III is probably not identical to that on other chromosomes. Yet it is of some interest to provide a statistical overview of the different structural and functional classes for this chromosome (Fig. 1).

Structural classes

About 14%, i.e., 25 of the 176 probable proteins of chromosome III, have domains with homologues of known three-dimensional structure (Table 1), as deduced from a

```
                 {      METAL-BINDING DOMAIN      }       { LINKER }

                 ========             ========       +++++++++
YCR106w      9 PRLRLVCLQCKKIKRKCDKLRP...ACSRCQQNSLQ..CEYEERTDLSAN
GAL4_YEAST   5 SSIEQACDICRLKKLKCSKEKP...KCAKCLKNNWE..CRYSPKTKRSPL
ARG2_YEAST  15 AKTFTGCWTCRGRKVKCDLRHP...HCQRCEKSNLP..CGGYDIKLRWSK
LAC9_KLULA  89 EVMHQACDACRKKKWKCSKTVP...TCTNCLKYNLD..CVYSPQVVRTPL
LEUR_YEAST  31 RKRKFACVECRQQKSKCDAHERAPEPCTKCAKKNVP..CILKRDFRRTYK
AMDR_ASPNI  14 GNGSAACVHCHRRKVRCDARLVG.LPCSNCRSAGKTD.CQIHEKKKKLAV
MALR_SACCA   2 GIAKQSCDCCRVRRVKCDRNKP....CNRCIQRNLN..CTYLQPLKKRGP
PDR1_YEAST  40 SKVSKACDNCRKRKIKCNGKFP....CASCEIYSCE..CTFSTRQGGARI
PPR1_YEAST  28 SKSRTACKRCRLKKIKCDQEFP...SCKRCAKLEVP..CVSLDPATGKDV
QA1F_NEUCR  70 QRVSRACDQCRAAREKCDGIQP...ACFPCVSQGRS..CTYQASPKKRGV
QUTA_ASPNI  43 QRVSRACDSCRSKKDKCDGAQP...ICSTCASLSRP..CTYRANPKKRGL
UGA3/YEAST  11 KYSKHGCITCKIRKKRCSEDKP...VCRDCRRLSFP..CIYISESVDKQS
CYP1_YEAST  58 NRIPLSCTICRKRKVKCDKLRP...HCQQCTKTGVAHLCHYMEQTWAEEA
YCO1_YEAST  11 SKAFKTCLFCKRSHVVCDKQRP....CSRCVKRDIAHLCREDDIAVPNEM

                       *  **  * **         *  *        *
```

Fig. 5. Multiple alignment of the N-terminal GAL4-like DNA-binding domains. GAL4-like fungal transcription factors are given by their Swiss-Prot codes. The conserved cysteines as well as positively charged positions are marked by asterisks. Functionally important residues of this domain (bold in GAL4) of known three-dimensional structure (Kraulis et al., 1992; Marmorstein et al., 1992) are mainly conserved in YCR106w. The similarity with the yeast protein YCO1 (bottom) has not been reported before, as far as we know.

sufficiently high level of sequence similarity (see Materials and methods; Sander & Schneider, 1991). This percentage is lower than that in the current databases of protein sequences, where it is about 25% of 24,000 sequences (unpubl.). The difference is most likely due to the fact that the current database is strongly biased toward well-studied protein families. Approximate three-dimensional models of these domains could be built, given sufficient interest (Holm & Sander, 1991). Such models would be quite accurate in the three-dimensional protein core and less accurate in loop regions. Three-dimensional models can be very useful for analyzing the role of conserved residues and for planning point mutation experiments to explore molecular details of function.

Inspection of output from the Blastp and Fasta programs (Pearson & Lipman, 1988; Altschul et al., 1990) and use of the program PropSearch (Sültemeyer, 1988) revealed 49 composition-biased proteins. This percentage is much higher than that in mammalian proteins sequenced so far (Brendel et al., 1992). Among these are 26 that are predicted to have at least one transmembrane region or are otherwise rich in hydrophobic amino acids: 7 with extended charge clusters, 7 with amino acid composition typical of coiled coils (Lupas at al., 1991), and 9 that are rich in a particular amino acid, such as serine or proline (e.g., YCR30c). At least for transmembrane proteins, most of which probably are predominantly helical, and for coiled coiled proteins an approximate overall structure can be predicted. Secondary structure prediction can easily be performed for all proteins, but its information content is limited.

Functional classes

What is the balance between different functional types of proteins on this chromosome? Among the predicted function of 74 of the 176 probable proteins, three major functional groups can be discerned: membrane proteins, soluble enzymes, and DNA-associated, regulatory proteins. One characteristic is an apparent accumulation of regulatory or DNA-associated proteins on chromosome III: 19 out of 176 proteins are in this category. Enzymes are a second major group (24 proteins), among them mainstream metabolic enzymes as well as regulatory and extracellular ones. Proteins in a third large group, 26 out of 176, are predicted to have at least one transmembrane region. Of these, 5 proteins (YCL25c, YCL69w, YCR11c, YCR23c, YCR98c) appear to be transporter proteins (Ringe & Petsko, 1990, and references therein). Another one is predicted as a 7-helix membrane protein (YCR75c; Hardwick & Pelham, 1990).

The transposon (Ty-17) region of chromosome III is well characterized (Warmington et al., 1986; Table 1). But there is another region of the chromosome containing retroelements (YCL74w, YCL75w). The first ORF is a copia-like protein and the following ORF has similarity to reverse transcriptases, in a region of 80 residues (Table 1). Other proteins located on chromosome III include mating factors, present in three different regions, a 40S ribosomal protein, several β-subunits of G-proteins, and a mitochondrial targeting protein (Table 1).

The function of 102 proteins (i.e., 58%) remains unclear at this point, although putative transmembrane regions, predicted coiled coil structures, and charge clusters (Fig. 1b) narrow the range of possibilities. It is also at present unclear to what extent "overrepresentation" of certain functional classes on particular chromosomes may have functional or evolutionary meaning.

Discussion

Current methods in protein sequence analysis by computer

Which methods work best?

In our experience, several of the basic database search tools, such as Fasta and Blastp work efficiently for the first scan. They are rapid, and, when a sufficiently strict threshold on the similarity scores or probability estimates is applied in reading the output, the similarities reported are reliable. Scanning against a database of six-frame translations of nucleic acids (e.g., using Tfasta) is essential for quality control of predicted ORFs (e.g., do they match known regulatory regions?); and, for increasing the chances of detecting similarities in the most recent database entries.

Considerable difficulties arise, however, in pursuing low level resemblances. Often a combination of several methods is required to increase confidence, and rarely are two cases alike. On occasion, additional information gleaned from publications or the intuitive grasp of the human expert performing the analysis is essential. After executing database scans on a small number of workstations, the main effort was expended in establishing the validity of some of the more remote similarities using a combination of tools and in generating the overall summary of results (what to report and how to present it). For the more detailed analysis, it is therefore essential to use a combination of methods and different ways of assessing significance–no single method embodies all currently available, proven expertise of sequence analysis. The real "art" lies not so much in the individual tools but rather in the users' expert application of these tools to pursue subtle relationships.

Where are the main limitations?

The principal limiting factors currently are the following: inadequate selectivity (accuracy) of algorithms for the detection of similarities in the "twilight zone"; missing or inadequate methods for the delineation of domains and for the prediction of protein tertiary structure; inadequate software tools for the detection of compositional

bias, in spite of recent progress in the statistical theory (Karlin & Brendel, 1992); inadequate integration of nonstandard methods into the software working environment; inadequate accessibility of published information that is buried in the literature; and, inadequate software for generating reports and feeding the derived data into the public databases.

Future need: development of an automatic scan-select-refine-store procedure

What is the most pressing need?

As genome projects come of age and start producing increasing amounts of data, the procedures used here will be woefully inadequate. They will become a bottleneck unless many more people are trained in the "art." What would be ideal is an integrated procedure that goes from the raw sequence and genetic map data to the final report and into the databases in automated fashion: scan-select-refine-store. This is utopia—but a view of utopia guides research.

In our opinion, development in the area of protein sequence analysis should focus on three major areas. The first is improved algorithms for the detection of real, but difficult to catch, homologies and for the direct prediction of structure and function. This is the theoretician's realm. The second is integration of heterogeneous tools into an overall working environment with facile exchange of information between tasks. This is the software engineer's challenge—the development of a sequence analysis "workbench." The third is the direct, "on-line, on-desk" availability of all relevant information in the biological literature. This is a task for database people and, in part, a problem of commercial politics.

How many more unknown proteins in yeast?

Probability of accurate prediction of function and structure

How much biological knowledge can we expect to gain between now and the time all yeast chromosomes have been sequenced? Given a newly sequenced open reading frame in yeast, what is the probability that the function (p(F)) or structure (p(S)) can be reliably predicted on the basis of similarity searches in databases? These probabilities depend crucially on the content of current databases and, to some extent, on the current level of technology in sequence comparison and prediction methods. Let us make a rough estimate based on the limited experience with chromosome III. After exhaustive sequence analysis, the fraction of known three-dimensional structures is $f(S) = 25/176$ and that of known functions $f(F) = 74/176$. By known we mean that the properties of at least one domain of the protein is already known or was predicted with a high probability of being correct. This fraction may be a slight overestimate in the sense that domains with known and unknown properties can be present in the same protein; and, because there may be a few errors, for whatever reason. Accepting a certain margin of error, we extrapolate from observed fraction to probabilities and estimate $p(F) = 0.42$ and $p(S) = 0.14$. In our opinion, a more than 42% probability of being able to identify the function of a new yeast protein on the basis of sequence analysis and a 14% probability of reliably predicting its three-dimensional fold is surprisingly high (Bork et al., 1992). We are almost halfway toward the goal of a first basic understanding of all protein types of a simple eukaryotic organism.

How many different protein types exist in yeast?

It is plausible to assume that most proteins in yeast have a specialized function, rather than simply being redundant duplicates of another protein. Yet for many purposes, e.g., biochemical classification, two proteins with the same basic function can be considered to be of the same basic type, e.g., two protein kinases with different specificities. In this sense, the redundancy in basic functional types on chromosome III appears to be relatively small: of the functionally identified 74 proteins there are approximately 65 different basic types. The definition of basic type is intuitive, at this point. For example, we assume here that two mating factors are of the same type, whereas a protein kinase and a ribokinase are of a different type. With these caveats, a rough estimate of redundancy is 1.1–1.2, defined as the total number of proteins divided by the number of different basic functional types. Yeast chromosome III has about 180 distinct proteins. Extrapolating to the entire yeast genome, which is 44 times larger than chromosome III, one arrives at about 8,000 distinct proteins and 7,000 basic types of proteins. These numbers are upper limits because (1) chromosome III may be more densely packed with open reading frames than other chromosomes, and (2) the probability of detecting remote similarities through the use of multiple alignments and pattern searches increases as more members of particular families become known.

Extrapolation to the human genome?

It is interesting to speculate about the approximate numbers of different proteins and basic protein types in the human genome. Genomes with more elaborate splicing mechanisms have a lower fraction of coding sequences, say 2–5% of the total genome. Also, the redundancy of basic types is probably higher in higher eukaryotes, e.g., there may be many more types of protein kinases with different regulatory roles. So, assuming an excess of noncoding to coding DNA by a factor of 20–50 and allowing for a factor of 3 in redundancy of basic types, the $3.6 * 10^6$ kilobases of the human genome would correspond to 30,000–100,000 distinct proteins and 10,000–30,000 dif-

ferent basic types. Although these estimates are highly speculative, it is only after the determination of the first sizable continuous fraction of a eukaryotic genome that they have a factual basis. With time, this type of estimate will become increasingly accurate.

The protein function information gap

The probability of identifying the function of a new yeast protein as the result of searching databases is, in our view, surprisingly high, at about 40% – a cause for optimism. However, the functions of the remaining 60% will be difficult to determine. As large fractions of the yeast genome and other genomes are sequenced, the sequence database will grow rapidly in content. As a result, given a new protein sequence of unknown function, the probability of finding in the databases at least one other protein related by sequence similarity will steadily, and then rapidly, increase. However, more and more often the similarity will be with a protein that also has unknown function, sequenced blindly. The information gap between the number of proteins sequenced and significant similarities detected, on the one hand, and the number of experimentally determined functions, on the other hand, will increase dramatically. Unless the technology of determination of function makes qualitative jumps, there will be tens of thousands of sequences of proteins with unknown function long after the first round of genome sequencing projects is completed.

Methods

Data

The entire DNA sequence of yeast chromosome III (315,357 base pairs) has been deposited in the EMBL database (European Molecular Biology Laboratory, Heidelberg; code: SCCHRIII, accession number: X59720). From the DNA sequence, 182 ORFs were predicted (ORFs longer than 100 amino acids) and made publicly available in March 1992 by Oliver et al. (1992), via the MIPS database (Martinsried Institute for Protein Sequences). We have used the amino acid sequences of these putative ORFs directly and subjected them to a variety of procedures that screen the databases of sequences, patterns, and three-dimensional structures for significant sequence similarities.

Search

For each of the ORFs, several databases were scanned for sequence similarities using a standard set of search tools. For maximum selectivity, the search was performed at the amino acid sequence level, rather than at the level of nucleotide sequences (Fig. 1). Each of the search methods is optimal for a certain type of question, with some overlap between them.

For example, Blastp (Altschul et al., 1990) performs a very rapid first scan for ungapped alignments with significance assessment based on probability estimates (Karlin & Altschul, 1990). In practice, the assessments work particularly well for fast detection of clear similarities among globular proteins. Fasta and Tfasta (Pearson & Lipman, 1988) perform a rapid screen based on the occurrence of n-tuples and then report optimal alignments with gaps. Our significance assessment using Fasta is based on the distribution of high scores relative to the database background and can be improved by sorting the best alignments by a variety of criteria (Fasta_Filter, R.S. & C.S., unpubl.). To find internal repeats in a protein, a full dynamic programming alignment (with gaps) of the sequence against itself is useful (for details see Fig. 1).

Database sequences similar to a query sequence in terms of amino acid composition and other global properties, rather than sequential alignment, can be detected using PropSearch (Sültemeyer, 1988; Hobohm & Sander, unpubl.). This is particularly useful for proteins that appear to have compositional bias. In such cases, one can verify by inspection that the sequence belongs to a cluster of proteins with similar residue composition, suggestive of a particular functional or structural adaptation.

Libraries of sequence patterns that characterize certain structural or functional features can provide additional information, beyond the sensitivity range of homology search methods. Here, we searched the Prosite motif library (Bairoch, 1991) and a library of patterns for extracellular domains (ExCell; Bork, 1991). Transmembrane segments can be detected by inspection of hydropathy profiles (Kyte & Doolittle, 1982).

Select

As a first criterion, the significance estimate of Blastp was used to detect obvious similarities, with a cut-off of $p = 10^{-10}$ for the probability of random occurrence of an alignment. With the current size of the database (10.6 million residues) and using a PAM120 matrix, an overall cutoff of $p < 10^{-10}$ was chosen based on the observation that only one probable false positive had a lower p-value: YCR44c against GARP_PLAFF had $p = 4.5 * 10^{-12}$ for a single best local alignment, later rejected because of composition bias. A cutoff at larger p-values would have brought in too many false positives. It is important to note that alignments with $p > 10^{-10}$ were not ignored but were treated as "twilight zone" candidates and investigated further.

In addition, high scores from Fasta and Tfasta were taken as an indication of significance, down to an optimized (Pearson & Lipman, 1988) value of about 100, complemented by two other scores. These are the excess percentage identity relative to the length-dependent thresh-

old of structural homology (Sander & Schneider, 1991) and the number of identical residues in an alignment. Note that Oliver et al. (1992) used a conservative cutoff of Fasta scores at the 200 level for their first scan for clear-cut sequence similarities. In the absence of a complete mathematical theory covering all cases, final judgment of significance was partly subjective, based on extensive experience in homology searches. Where possible, multiple sequence alignment of an entire protein family and other methods were used to confirm or reject putative homologies.

Refine

In interesting cases, additional methods were used to verify and extend the results. This includes more sophisticated alignment methods, pattern or profile search programs (e.g., Gribskov et al., 1987), various significance estimates for "twilight zone" candidates, as well as inclusion of any available experimental information. The term "twilight zone" loosely defines the region of sequence similarity in which "it is virtually impossible to distinguish between chance similarity and common ancestry" (Doolittle, 1985)—with pairwise sequence alignment methods. Here, we consider similarities with a Blastp (Altschul et al., 1990) p-value of $p > 10^{-10}$ to be in the "twilight zone," worthy of further analysis (for a database of about 10.6 million residues).

Refinement of pair alignments

The very fast search programs Blastp and Fasta do not necessarily capture the full extent of sequence similarity. A careful and more time-consuming application of full dynamic programming alignment methods, e.g., that of Smith and Waterman (1981) with various parameter sets, can in some cases detect much longer common subsequences. For example, for the putative type X DNA polymerase (YCR14c), the Fasta alignment of 24% identical amino acids out of 213 (notation: 24%/213) can be extended to about double the length (27%/411). Here we used the program Bestfit of the GCG package (Devereux et al., 1984), with default parameters (gap open: 3.0, gap elongation: 0.1). In another case, that of the predicted acetolactate synthase (YCL9c), the alignment was extended to more than double the length, i.e., from 40%/90 to 36%/208 (Fig. 2). A third example is that of the putative methyltransferase YCR47c, for which the improvement went from 36%/66 to 26%/301. In each of these examples, the longer alignments did not have an excessive number of gaps.

Multiple alignment

Another way of verifying results is the use of multiple sequence alignment methods, e.g., Pileup of the GCG package (Devereux et al., 1984), if more than one related protein sequence is known. The presence of certain conserved residues or properties in all sequences increases the confidence in the detected relationships, especially when the conserved residues are known to be functionally important. An example is the putative DNA polymerase YCR14c (Fig. 3).

Similarity search by sequence patterns

In some cases, short sequence motifs, but not longer subsequences, are common to a family of functionally related proteins and can be treated as marker regions. Some of these motifs are well known (e.g., several different motifs for ATP-binding sites) and documented in pattern libraries (Prosite, ExCell). Others can be derived directly from multiple sequence alignments. For example, when database searches with YCR47c detected several methyltransferases, albeit at a low level of similarity, a pattern of conserved residues common to the entire family became apparent. Such patterns are usually indicative of a common function and mechanism involving the conserved residues (Fig. 4). The patterns are usually too sparse to be detected as significant in standard sequence database searches. However, with the additional information derived from a known protein family, a match with a pattern takes on special significance, compared to a match of apparently equal strength that involves only nonconserved or nonfunctional residues.

Matching domain location

Further support for remote relationships comes from knowledge of the domain structure of matching candidates. For example, a segment of the query sequence may match a well-defined and previously characterized domain of a database protein. In addition, if the location of this domain in the query sequence matches that in the database protein, one has more confidence in the putative homology, e.g., YCR106w matches the N-terminal GAL4-like DNA-binding domains of otherwise dissimilar transcription factors (Fig. 5) precisely with its first 53 residues.

Resources and effort

Computer time

Searching the sequence databases with several hundred query sequences requires considerable computing resources. Fortunately, Blastp (Altschul et al., 1990) provides a very rapid and reasonable first scan: searches with all 182 ORFs against Swiss-Prot and PIR-only took less then 3 h CPU (central processor unit) time on one R3000 processor, clocked at 33 MHz, of a Silicon Graphics 4D/480 workstation. Larger databases, slower programs, and slower machines require considerably more time. For example, the (slower) program Tfasta was used to screen all six translations of 48,434 NA-only entries, i.e., an 11 times larger database of about 300,000 translated amino acid sequences, and would require a total of about 23 days

CPU time on the same machine. The actual time used can be—and was—reduced by using several processors or/and several workstations in parallel. The corresponding Fasta runs against 35,000 protein sequences would take about 90 h on a (slower) VAX 6040; on a single i860 processor of an Alliant FX 2800 this was reduced to 15 h. Using a combination of software and hardware, all runs, including setup and testing, were completed in 2 weeks.

Human time

An additional 6 weeks of two persons working full time was needed for verification and documentation of the results, and for preparation of the final report. The fact that the amount of human time spent on data analysis was the major limiting factor points to the relatively low level of automation of low stringency, state-of-the-art computer sequence analysis.

Acknowledgments

We thank Rich Roberts for detailed information on methyltransferases, Amos Bairoch for useful hints, Mark Boguski for helpful comments, the BRIDGE programme of the European Community and the Human Frontiers Science Programme (HFSP) for financial support, the EMBL computer group for systems support, SUN Microsystems for a Scientific Equipment Grant, and the Gesellschaft für Mathematik und Datenverarbeitung (GMD) for access to a 16-processor parallel computer.

References

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

Bairoch, A. (1991). PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res. 16*, 2241–2245.

Bairoch, A. & Boeckmann, B. (1991). The SWISS-PROT protein sequence databank. *Nucleic Acids Res. 19*, 2247–2249.

Barker, W.C., George, D.G., Hunt, L.T., & Garavelli, J.S. (1991). The PIR protein sequence database. *Nucleic Acids Res. 19*, 2231–2236.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Lett. 286*, 47–54.

Bork, P. & Grunwald, C. (1990). Recognition of different nucleotide binding sites using a property pattern approach. *Eur. J. Biochem. 191*, 347–358.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., & Sonnhammer, E. (1992). What's in a genome? *Nature 358*, 287.

Brendel, V., Bucher, P., Nourbakhsh, I.R., Blaisdell, E., & Karlin, S. (1992). Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci. USA 89*, 2002–2006.

Date, T., Yamamoto, S., Tanikara, K., Nishimoto, Y., & Matsukage, A. (1991). Aspartic acid residues at positions 190 and 192 of rat DNA polymerase β are involved in primer binding. *Biochemistry 30*, 5286–5292.

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

Doolittle, R.F. (1985). Proteins. *Sci. Am. 253*, 88–99.

Duronio, R.J., Gordon, J.I., & Boguski, M.S. (1992). Comparative analysis of the b transducin family with identification of several new members including PWP1 a non-essential gene of *Saccharomyces cerevisiae* that is divergently transcribed from NMT1. *Proteins 13*, 41–56.

Franco, L., Jimenez, A., Demolder, J., Molemans, F., Fiers, W., & Contreras, R. (1991). The nucleotide sequence of a third cyclophilin-homologous gene from *Saccharomyces cerevisiae*. *Yeast 7*, 971–979.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol. 162*, 705–708.

Gribskov, M., McLachlan, A.D., & Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA 84*, 4355–4358.

Hardwick, K. & Pelham, H. (1990). ERS1, a seven transmembrane domain protein from *Saccharomyces cerevisiae*. *Nucleic Acids Res. 18*, 2177.

Holm, L. & Sander, C. (1991). Database algorithm for generating protein backbone and side chain co-ordinates from a C(alpha) trace. Application to model building and detection of co-ordinate errors. *J. Mol. Biol. 218*, 183–194.

Ingrosso, D., Fowler, V., Bleibaum, J., & Clarke, S. (1989). Sequence of D-aspartyl/L-isoaspartyl protein methyltransferase from human erythrocytes. *J. Biol. Chem. 264*, 20131–20139.

Ito, J. & Braithwaite, D.K. (1991). Compilation and alignment of DNA polymerase sequences. *Nucleic Acids Res. 19*, 4045–4057.

Karlin, S. & Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA 87*, 2264–2268.

Karlin, S. & Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science 257*, 39–49.

Klimasauskas, S., Timinskas, A., Menkevicius, S., Butkiene, D., Butkus, V., & Janulaitis, A. (1989). Sequence motifs characteristic of DNA [cytosine-N4]methyltransferases: Similarity to adenine and cytosine-C5 DNA-methylases. *Nucleic Acids Res. 17*, 9823–9832.

Kondo, K. & Inouye, M. (1991). TIP1, a cold shock-inducible gene of *Saccharomyces cerevisiae*. *J. Biol. Chem. 266*, 17537–17544.

Kraulis, P.J., Raine, A.R.C., Gadhavi, P.L., & Laue, E.D. (1992). Structure of the DNA-binding domain of zinc GAL4. *Nature 356*, 448–450.

Kyte, J. & Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 157*, 105–132.

Li, C., Lai, C., Sigman, D.S., & Gaynor, R.B. (1991). Cloning of a cellular factor, interleukin binding factor, that binds to NFAT-like motifs in the human immunodeficiency virus long terminal repeat. *Proc. Natl. Acad. Sci. USA 88*, 7739–7743.

Lupas, A., Van Dyke, M., & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science 252*, 1162–1164.

Marguet, D., Guo, X.J., & Languin G.J.-M. (1988). Yeast gene SRP1 (serine-rich) protein. *J. Mol. Biol. 202*, 455–470.

Marmorstein, R., Carey, M., Ptashne, M., & Harrison, S.C. (1992). DNA recognition by GAL4: Structure of a protein–DNA complex. *Nature 356*, 408–414.

Oliver, S.G., van der Aart, Q.J.M., Agostoni-Carbone, M.L., Aigle, M., et al. (1992). The complete DNA sequence of yeast chromosome III. *Nature 357*, 38–46.

Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA 85*, 3338–3342.

Ringe, D. & Petsko, G.A. (1990). A transport problem? *Nature 346*, 312–313.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins 9*, 56–68.

Skala, J., Purnelle, B., & Goffeau, A. (1992). The complete sequence of a 10.8kb segment distal of SUF2 on the right arm of chromosome III from *Saccharomyces cerevisiae* reveals seven open reading frames including the RVS161, ADP1 and PGK genes. *Yeast 8*, 409–417.

Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol. 147*, 195–197.

Sor, F., Cheret, G., Fabre, F., Faye, G., & Fukuhara, H. (1992). Sequence of the HMR region on chromosome III of *Saccharomyces cerevisiae*. *Yeast 8*, 215–222.

Sültemeyer, R. (1988). Vergleich von Proteinen anhand charakteristischer Sequenzkenngrößen. Diploma thesis, Medizinische Informatik, Fachhochschule Heilbronn, Germany.

Taha, M.K., So, M., Seifert, H.S., Billyard, E., & Marchal, C. (1988). Pilin expression in *Neisseria gonorrhoeae* is under both positive and negative transcriptional control. *EMBO J. 7*, 4367–4378.

Thierry, A., Fairhead, C., & Dujon, B. (1990). The complete sequence of the 8.2 kb segment left of MAT on chromosome III reveals five ORFs, including a gene for a yeast ribokinase. *Yeast 6*, 521–534.

van der Voorn, L. & Ploegh, H.L. (1992). The WD-40 repeat. *FEBS Lett. 307*, 131–134.

Warmington, J.R., Anwar, R., Newlon, C.S., Waring, R.B., Davies, R.W., Inolge, K.J., & Oliver, S.G. (1986). A 'hot-spot' for Ty transposition on the left arm of yeast chromosome III. *Nucleic Acids Res. 14*, 3475–3485.

Warmington, J.R., Waring, R.B., Newlon, C.S. Indge, K.J., & Oliver, S.G. (1985). Nucleotide sequence characterization of Ty 1–17, a class II transposon from yeast. *Nucleic Acids Res. 13*, 6679–6693.

Waterman, M.S. & Eggert, M. (1987). A new algorithm for best subsequence alignments with application to tRNA–tRNA comparison. *J. Mol. Biol. 197*, 723–728.

Weigel, D. & Jäckle, H. (1990). The *fork head* domain: A novel DNA binding motif of eukaryotic transcription factors? *Cell 63*, 455–456.

Wek, R.C., Hauser, C.A., & Hatfield, G.W. (1985). The nucleotide sequence of the *ilvBN* operon of *E. coli*: Sequence homologies of the acetohydroxy acid synthase isozymes. *Nucleic Acids Res. 13*, 3995–4011.

Wiersma, P.A., Schmiemann, M.G., Condie, J.A., Crosby, W.L., & Moloney, M.M. (1989). Isolation, expression and phylogenetic inheritance of an acetolactate synthase gene from *Brassica napus*. *Mol. Gen. Genet. 219*, 413–420.