Request # 18588844 **NOV 17, 2005**

Ariel To: 64.40.17.85/140.163.217.217

Memorial Sloan-Kettering Cancer Center
Medical Library Nathan Cummings Center (METRO #146)
1275 York Avenue
New York, NY 10021

DOCLINE: Journal Copy Epayment

Title:	Yeast (Chichester, England)
Title Abbrev:	Yeast
Citation:	1995 Jan;11(1):61-78
Article:	Nucleotide sequence and analysis of the centromeri
Author:	Voss H; Tamames J; Teodoru C; Valencia A; Sensen C; Wiemann
NLM Unique ID:	8607637 Verify: PubMed
PubMed UI:	7762303
ISSN:	0749-503X (Print) 1097-0061 (Electronic)
Publisher:	Wiley, Chichester ; New York :
Copyright:	Copyright Compliance Law
Authorization:	A Artale
Need By:	NOV 21, 2005
Maximum Cost:	**$25.00**
Patron Name:	Gangi-Dino, Rita
Referral Reason:	Not owned (title)
Library Groups:	BQSIMB,EFTS,RESOURCE,METRO
Phone:	1.212.639-7441
Fax:	1.646.422-2316
Email:	ill@mskcc.org
Comments:	**Please, we prefer as PDF or Ariel. Thank you! EFTS, METRO, BQSI.**
Routing Reason:	Routed to NYUHHQ in Serial Routing - cell 1
Received:	Nov 17, 2005 (02:58 PM EST)
Lender:	Cold Spring Harbor Laboratory/ Cold Spring Harbor/ NY USA (NYUHHQ)

Bill to: NYUMSK

Memorial Sloan-Kettering Cancer Center
Medical Library Nathan Cummings Center
1275 York Avenue
New York, NY 10021

YEAST VOL. 11: 61–78 (1995)

Yeast Sequencing Reports

Nucleotide Sequence and Analysis of the Centromeric Region of Yeast Chromosome IX

H. VOSS*, J. TAMAMES†, C. TEODORU, A. VALENCIA†, C. SENSEN, S. WIEMANN, C. SCHWAGER, J. ZIMMERMANN, C. SANDER† AND W. ANSORGE

Biochemical Instrumentation and †Biological Structures and Biocomputing Programmes, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany

Received 22 June 1994; accepted 8 September 1994

We have determined the nucleotide sequence of a cosmid (pIX338) containing the centromere region of yeast (*Saccharomyces cerevisiae*) chromosome IX. The complete nucleotide sequence of 33·8 kb was obtained by using an efficient directed sequencing strategy in combination with automated DNA sequencing on the A.L.F. DNA sequencer. Sequence analysis revealed the presence of 17 open reading frames (ORFs), four of them previously known yeast genes (*sly12, pan1, sts1* and *prl1*), a tRNA gene and the centromere motif. Exhaustive database searches detected sequence homologues of known function for as many as 14 of the 17 ORFs. These include a mammalian tyrosine kinase substrate; the *Escherichia coli* cell cycle protein MinD; the human inositol polyphosphate-5-phosphatase (gene OCRL) involved in Lowe's syndrome, a developmental disorder; and helicases, for which the new yeast member defines a distinct DEAD/H-box subfamily. A surprisingly large fraction of the ORFs (at least six out of 17) in the centromeric region are apparently involved in RNA or DNA binding.

The nucleotide sequence reported here has been submitted to the EMBL data library under the accession number X79743.

KEY WORDS — *Saccharomyces cerevisiae*; chromosome IX; centromere; nucleic acid binding proteins.

INTRODUCTION

The Commission of the European Union has organized and is funding a project for sequence determination and analysis of the *Saccharomyces cerevisiae* genome (Oliver *et al.*, 1992; Dujon *et al.*, 1994). In the context of this project, we have determined the nucleotide sequence of a cosmid (pIX338) containing the centromere region of yeast chromosome IX. The complete nucleotide sequence of 33·8 kb was obtained by using an efficient directed sequencing strategy in combination with automated DNA sequencing. The region contains four previously known yeast genes (*sly12, pan1, sts1* and *pri1*), a tRNA gene, the centromere region and 17 apparent open reading frames (ORFs).

After the sequence determination, detailed sequence analysis was carried out. Close contact between the sequencing and computing groups increased confidence in the results of sequence analysis. For example, sequence discrepancies between YIB6c (sequence determined in this work) and *pan1* (sequence reported previously; Sachs and Deardorff, 1992) were resolved in part by making reference to the conservation pattern in the family of sequence-related proteins. In another example, an unexpected amino acid sequence feature in a new member (YIB2c) of the DEAD/H-box family of helicases was confirmed by making reference to the raw sequence data. Convenient access to the

*Corresponding author.

CCC 0749–503X/95/010061–18

raw data in the course of sequence analysis was essential in such cases.

In addition to presenting the overall results of the sequencing of this cosmid, we illustrate the results of sequence analysis by discussing in more detail six examples of particular biological interest. In one case (YIB9w, an snRNP U1/U2 protein), we illustrate the information flow from a new DNA sequence, via database searches, family assignment and phylogenetic characterization, all the way to the construction of an explicit three-dimensional model by homology and mapping of conserved regions on the model structure.

MATERIALS AND METHODS

DNA sequencing

The complete insert of cosmid pIX338 was sequenced by a directed strategy using automated DNA sequencing. Cosmid DNA (a pWE15 construct) was digested with *Eco*RI and separated by agarose electrophoresis. All fragments were purified and subcloned into *Eco*RI-digested pUC18. The resulting subclones were purified in either midi (fragments <3 kb) or maxi (fragments >3 kb) scale DNA preparations (Nucleobond AX kit, Macherey-Nagel). With the exception of subclone pIX338E9, which was sequenced by nested deletion on one strand, all *Eco*RI plasmid subclones were sequenced on both strands by an efficient primer walking strategy using automated DNA sequencing with T7 DNA polymerase and internal labelling by fluorescein-15*-dATP (Voss *et al.*, 93). Inexpensive oligonucleotides for primer walking were synthesized in sets of ten on the EMBL segmental multiple DNA synthesizer (Ansorge *et al.*, 1992). Double-stranded plasmid and cosmid sequencing reactions were performed as described previously (Voss *et al.*, 1993). After sequencing all *Eco*RI plasmid subclones, the fragments were linked by direct cosmid sequencing using either cycle sequencing with fluorescently labelled primers (Zimmermann *et al.*, 1994) or using T7 DNA polymerase and internal labelling with fluorescein-15*-dATP. All DNA sequencing reactions were performed on a Biomek Robotic workstation (Beckman) (Zimmermann *et al.*, 1992) and analysed on a standard A.L.F. DNA sequencer (Pharmacia). Raw data collection and evaluation were performed using the A.L.F. manager software. Walking primers were designed using the OLIGO programme (MedProbe). Sequence assembly, data evaluation and presentation were performed using the EMBL GeneSkipper software (available from EMBL). The pIX338 consensus sequence of 33852 bases was submitted to the MIPS data library (Munich) on 15 February 1994.

ORF definition

All ORFs larger than 70 amino acids were translated with the MAP program (GCG package) (Devereux *et al.*, 1984) and, for each one, independent database searches were carried out. There was no need for excluding any ORF with significant similarity in databases because of its size or because of its overlapping with other ORFs. The presence of 15 ORFs could be confirmed by the presence of homologous sequences. For two other proposed ORFs, no homologies could be detected (Figure 1 and Table 1).

Names of the ORFs are preliminary and may change with the sequencing of the complete chromosome IX. In the absence of a physical map for this region, we name the ORFs A or B instead of the usual left/right nomenclature.

Data analysis

The database searches for homologous sequences have been done in the framework of 'genequiz', a project management, browsing and visualization tool developed by the EMBL protein design group (Scharf *et al.*, 1994).

The following databases were searched: *protein sequences*, PDB (Abola *et al.*, 1987), Swissprot (Bairoch and Boeckmann, 1993), PIR-NBRF (fraction not overlapping with Swissprot (Barker *et al.*, 1993), GENPEPT, a direct translation of the DNA sequences in GENBANK (Benson *et al.*, 1993); *DNA sequences*, EMBL-GenBank (Rice *et al.*, 1993), expressed sequence tags in dbEST (Boguski *et al.*, 1993). Updated versions of the databases were used as of 1 June 1994.

Prior to the database scanning, sequences were masked using the program 'SEG' (Wootton and Federhen, 1993) to avoid spurious hits in regions of obvious composition bias. The scan of the database was done using the Blast (Altschul *et al.*, 1990) and Fasta (Pearson and Lipman, 1988) programs (parameters: BLOSUM62 matrix for Blast; and Ktup=2 for Fasta).

Multiple-sequence alignments were obtained using the programs MAXHOM (Sander and Schneider, 1993), CLUSTALW (Higgins *et al.*,

Figure 1. Restriction and ORF map of the insert of cosmid pIX338. Top: position of *Eco*RI restriction sites. Second row, black boxes: previously sequenced areas, including the region of *cen9*. Second row, open box: a region homologous to tRNA/Glu. Rows a, b, c, d, e, f: the six possible reading frames, with labelled arrow boxes for open reading frames. Open boxes: no significant sequence similarity to database entries. Shaded boxes: one or more homologues detected in databases. Putative functions based on sequence similarities are indicated as in Table 1.

Table 1. Assignment of ORFs in the centromeric region of yeast chromosome IX.

ORF	Gene	aa	Function	Closest hit	%id/aa‡	dbest	3D	Motif/pattern
YIA1a		514	Ankyrin repeats	ANK1_HUMAN	19%, 578	21,451 *A. thaliana*		ANK repeats
YIA2c		947	Inositol phosphatase	OCRL_HUMAN	26%, 956	20,970 *A. thaliana*		
YIA3w		294	ATPase	MRP_ECOLI	38%, 246	23,348 *O. sativa*		ATP/GTP binding
YIA4c	*SLY12*	138	Complement YPT1 def.					
YIA5w		154						
YIB1c		251	hnRNP	RO28_SPIOL*	19%, 198	21,181 *A. thaliana*		RNP1,2
YIB2c		994	DEAD/H helicase	DEAD_ECOLI	21%, 960	4062†		DEAD/H
YIB3w		680	Composition bias			12,960†		
YIB4w		433	DnaJ	DNAJ_CLOAB	50%, 70	44,368 *N. tabacum*		DNAJ
YIB5w		149	hnRNP	CST2_HUMAN	31%, 173	15,984 *C. elegans*		RNP1,2
YIB6c	*PAN1*	1481	Poly(A) processing	GP:L21768	22%, 1153	28,976 *P. falciparum*		
YIB7w		765	Endoglucanase	GUN_BURSO	14%, 539			Glycosyl hydrolase
YIB8c	*PRI1*	410	Primase	PRI1_MOUSE	36%, 368	25,700 *H. sapiens*		
YIB9w		112	U2 RNP B	RU2B_HUMAN	31%, 95	21,158 *A. thaliana*	1NRC-A	RNP1,2
YIB10w		577						
YIB11c	*STS1*	320	Toxicity suppressor					
YIB12w		432	β-Transducin repeats	PRO4_YEAST	21%, 431	27,154 *H. sapiens*		gβ-repeats

*YIB1c is closer to CST2_HUMAN and YIB5w (phylogenetic analysis not shown).
†100% identical, most probably a yeast contamination in human library. Indicates that these genes are expressed in yeast.
‡Percentage of identity (%id) over a given sequence length (aa).

1992) or PILEUP (GCG package) and represented using PRETTYPLOT (Peter Rice, EMBL).

Protein secondary structure was predicted from multiple sequence alignments using the PHD neural network method (Rost and Sander, 1993), as implemented on the PredictProtein network server (Rost *et al.*, 1994). Phylogenetic trees based on the neighbour-joining method (Saitou and Nei, 1987) were calculated using the CLUSTALW package (Higgins *et al.*, 1992). Corrections for multiple replacements were applied (Kimura, 1983). The stability of the trees with respect to different choices of subsets of residue positions was checked by bootstrapping experiments (Felsenstein, 1985). Three-dimensional modelling by homology was done with the protein engineering software package WHATIF (Vriend, 1990). The quality of the final model was assessed using directional atomic contract analysis (Vriend and Sander, 1993).

The databases were first searched with the translated protein sequences and later with the full DNA sequence. In some cases it was possible to expand the limits of the families and discover new members by further searches with profiles derived from initial multiple sequence alignments. Profile searches were carried out using PROFILESEARCH (GCG) or MAXHOM (unpublished).

Full results of the database searches, multiple sequence alignments and three-dimensional modelling are available on Internet from ftp.embl-heidelberg.de in the directory/pub/databases/yeast/chr9 as well as via World Wide Web, using the resource locator http://www.embl-heidelberg.de.

RESULTS

Sequence determination

In total, 244 overlapping fragments were sequenced to determine completely the 33,852 bp of pIX338 on both strands. By using an ordered sequencing strategy (90% primer walking, 10% nested deletion), the average number of readings per base was 3·1, significantly lower than figures obtainable with random shotgun sequencing strategies. The sequencing statistics for cosmid pIX338 are summarized in Table 2.

Nucleotide sequence discrepancies with previously known yeast genes were found to be around 0·3%. Most of the deviations are in the 3′ and 5′ extremes, while 100% identical sequences are usually found in the coding regions. In all cases of deviations, the correct consensus sequence of our data was verified on both strands of our raw data. The centromere motif (138 bases) was found to be 100% identical to the previously published sequence (Wustinger and Spevak, 1993).

Table 2. Sequencing statistics for cosmid pIX338.

Final sequence of cosmid pIX338	33,852 bp
Total number of bases sequenced	108,992 bases
Total number of fragments sequenced	244
Number of walking primers synthesized	174
Average coverage	3·1
Average reading length (on 30 cm gels)	446 bases

Sequence analysis

The complete fragment sequenced contains a number of DNA regions previously deposited in databases: (a) from base 614 to base 1027, containing the *sly12* (*bet1*) gene (Newman *et al.*, 1992); (b) from 18,130 to 22,572, containing the *pan1* gene (Sachs and Deardorff, 1992); (c) from 25,741 to 26,970, containing the *pri1* gene (Plevani *et al.*, 1987); (d) from 29,095 to 30,910, containing the *sts1* gene (Bissinger and Kuchler, 1994). Also a tRNA gene was found between 23,080 and 23,183; the encoded tRNA is most similar to Glu tRNAs. The centromeric motif is localized between 8303 to 8440 (Wustinger and Spevak, 1993) (Figure 1). EST homologues were found for 13 of the proposed ORFs. In two of these cases (YIB2c and YIB3w), the human ESTs found are 100% identical to the yeast sequences, which probably is due to a yeast contamination of the human cDNA libraries, but still implies that these ORFs can be expressed in yeast.

Among the identified 17 ORFs in pIX338 we found four yeast genes that had already been identified (see Table 1 and Figure 1). For two of them, *sts1* and *sly12*, we found no homologues in the databases. In both cases, functional information is known about the corresponding protein. Bet1 protein, the product of the *sly12* gene, is needed for transport from the endoplasmic reticulum to the Golgi complex (Newman *et al.*, 1992). The Sts1 protein acts as a suppressor of the toxicity of sporidesmin. For the products of two other known yeast genes, *pan1* and *pri1*, homologous sequences defining sequence families were found. Eps15 (human and mouse) is homologous

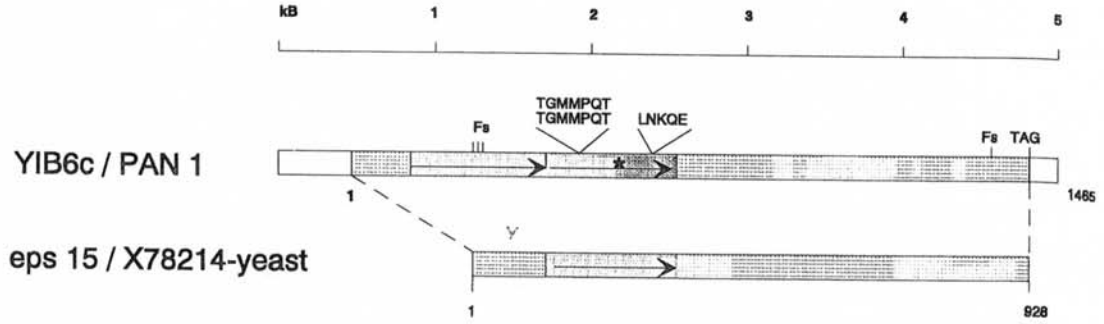


Figure 2. Resolution of sequence discrepancies between YIB6c and the *pan 1* gene. Top bar: organization of the *pan1*/YIB6c gene (Sachs and Deardorff, 1992), with the deduced open reading frame from base 505 to 4914 in light shading and the two duplicated regions in darker shading (Sachs and Deardorff, 1992). The second of these duplicated regions (*) corresponds to the region known to be essential for cell viability. Discrepancies between *pan1* and YIB6c in the coding region: 'Fs' are single deletions at nucleotide positions 1300, 1325, 1326 and 4636; the deletion at 4636 leads to a stop codon at position 4716, correcting the C-terminal end of the protein product; TGMMPQT and LNKQE are insertions at nucleotide positions 1943, 2423. Bottom bar: sequence organization of the *eps15*, a mammalian homologue of *pan1*/YIB6c and of yeast chromosome II homologue X78214. Both of these are shorter than *pan1*/YIB6c and most of the first duplication in the N-terminal half of *pan1*/YIB6c is not present. The C-terminal region of Eps is more similar to YIB6c and does not match the extension originally reported for Pan1p. The homology detected is strong in the second repeat of the N-terminal region (heavy shading); the sequence alignment is in Figure 3.

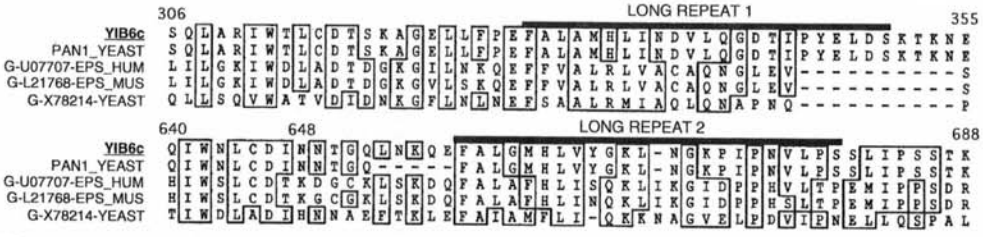


Figure 3. Multiple sequence alignment of the most conserved regions between yeast *Pan1*/YIB6c, mammalian Eps15 and yeast G-X78214. The regions aligned correspond to the most similar regions in the right end of the two duplications (duplications are marked with arrows and heavy shading in Figure 2, eps15/G-X78214-yeast line). All sequences are more similar in the region called 'LONG REPEAT 2', which includes a characteristic 'FALxxHLxxxKxxxG' pattern. Proteins Eps15, G-X78214-yeast and YIB6c differ from the originally reported Pan1p sequence; in particular they do not have the deletion found in Pan1p (positions 652 to 658, LNKQE in YIB6c). The result is a proposed correction of the Pan1p sequence based on that of YIB6c.

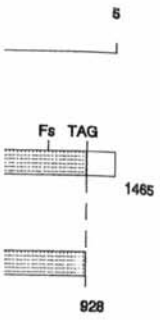
to Pan1p, and at least two homologues for Pri1p were found (mouse and *C. elegans*).

For ten ORFs, the sequence was not previously known in yeast, but homologues in other species were found. Of the remaining three ORFs, two of them do not have any homologue in databases (YIA5w and YIB10w). The third, YIB3w, a protein with composition bias (rich in Gln), matches an EST from a human cDNA library, but this most probably is a result of yeast contamination (NCBI, May 1994) in human cDNA libraries.

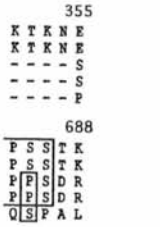
For each ORF, the family classification and the most similar sequence found in the database search are given in Figure 1 and Table 1. Here we describe in detail six examples of particular biological relevance.

Deviations of YIB6c from the previously published pan 1 gene

We identified ORF YIB6c as the previously sequenced *pan1* gene (Sachs and Deardorff, 1992). *pan1* is a single-copy gene in yeast, which is organized in distinct domains containing several repeated sequence elements (Figure 2). The gene encodes the poly(A)-binding protein-dependent poly(A) ribonuclease. Deletion analysis of the gene revealed that only a domain in the N-terminal half of the protein is needed to maintain cell viability



anization of the
to 4914 in light
these duplicated
n1 and YIB6c in
ion at 4636 leads
and LNKQE are
alian homologue
B6c and most of
s more similar to
ng in the second



n1/YIB6c,
ons in the
Figure 2,
' 2', which
B6c differ
in Pan1p
nce based

cular biological

iously published

the previously
Deardorff, 1992).
yeast, which is
ntaining several
re 2). The gene
otein-dependent
ılysis of the gene
N-terminal half
ain cell viability

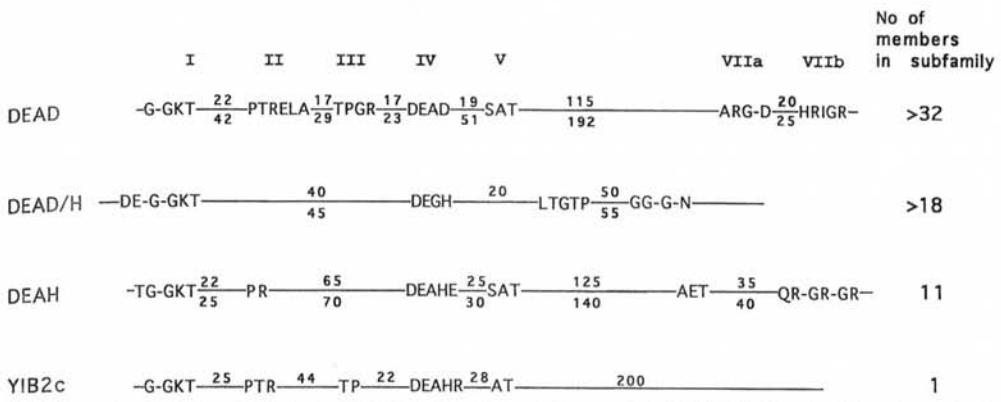


Figure 4. Conserved regions in the DEAD/H family of helicases. The DEAD/H-box family of proteins is characterized by conserved boxes (roman numerals I–VII) with characteristic sequence patterns (upper-case letters) and distances between boxes (smallest/largest distances as integers above/below the line). The known subfamilies are called DEAD, DEAH, DEAD/H. YIB2c matches the common features of the family in patterns I and IV, matches closely the DEAD characteristics in patterns II, III and V, but has clearly different characteristics in patterns IV, VIIa and VIIb. Therefore we propose that this sequence is the first representative of a new DEAD/H-box subfamily.

(Sachs and Deardorff, 1992; labelled with an asterisk in Figures 2 and 3). Comparing the nucleotide sequence of ORF YIB6c to the published *pan1* gene, we noticed a number of discrepancies (Figure 2). A deletion of nucleotide 1300 of the *pan1* sequence results in a frame-shift over 10 amino acids until the deletion of nucleotides 1325 and 1326 re-establishes the original frame. At nucleotide position 1943 of the *pan1* sequence, a 42 bp insertion results in a repeat of the amino acid motif TGMMPQT (amino acids 474 to 480), which is already part of a repetitive motif. At nucleotide position 2423, a 15-bp insertion adds the pentapeptide LNKQE to the sequence, directly adjacent to another repeat. Besides these deviations which maintain the original ORF, a base insertion at position 4636 results in a frame shift in the C-terminal end. This is particularly interesting, as the frame shift results in a stop codon after amino acid 1404 at nucleotide position 4716. The deduced protein is now shortened and lacks the described C-terminal tripeptide CFL, originally thought to be a site for covalent attachment of a lipid moiety.

Database search revealed two homologous sequences: the eps15 gene product corresponding to a receptor Tyr kinase substrate found in human and mouse (Wong *et al.*, 1994; Fazioli *et al.*, 1993) (90% sequence identity between the human and mouse sequences). The protein product of eps15 is shorter than that of YIB6c, lacking most of the N-terminal repeat. The homology detected is only strong in a limited region (dark shading in Figure 2, sequence alignment in Figure 3) that coincides with the only region shown to be essential for cell viability. The homology with eps15 supports our sequence data in the C-terminus where eps15 also does not have the C-terminal extension originally reported for Pan1 protein. Also, sequence similarity with eps15 confirms the sequence of YIB6c in the region of 2423 where YIB6c and eps15 are quite similar and the originally reported sequence of *pan1* has a deletion (Figure 3). In addition, the family can be extended by another protein, GP: x78214-yeast, a yeast sequence from chromosome II without known function, that is more similar in sequence and in length to eps15 than it is to Pan1p/YIB6c. Similarity is again stronger in domain 2, in which eps15 is most similar to Pan1p (Figures 2 and 3). It will be very interesting to see if the homology found here leads to the discovery of a biological connection between control of cell growth mediated by Tyr phosphorylation (eps15) and poly (A)-ribonuclease function (YIB6c/Pan1p), i.e., whether the mammalian eps15 has poly (A)-ribonuclease function and whether yeast YIB6c/Pan1 protein is phosphorylated on a conserved Tyr residue. Sequences from other organisms would also contribute to the understanding of this new protein family.

YIB2c defines a new subfamily in the DEAD/H family of helicases

DEAD/H helicases form an extended family with members in all kingdoms (Gorbalenya *et al.*,

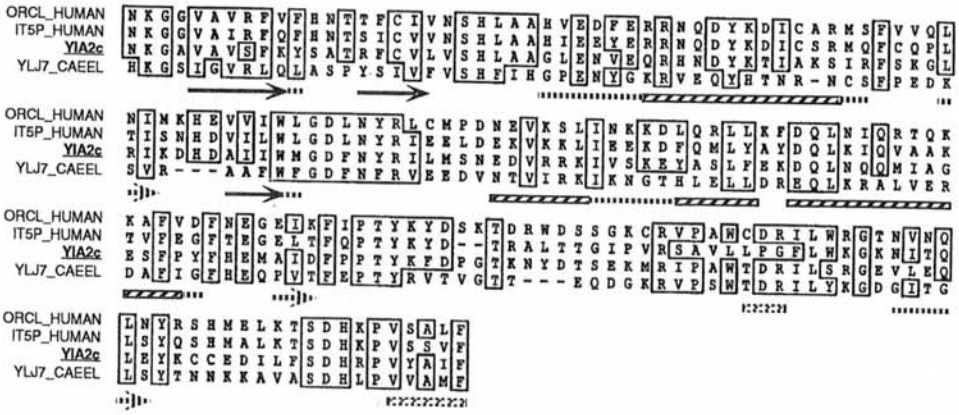


Figure 5. Multiple sequence alignment of the central region of the inositol polyphosphate-5-phosphatase/OCRL family. The most conserved region of the alignment corresponds to the pattern 'hWhGDhNYRh', where h is a hydrophobic residue. In this region the yeast sequence has 50% identical residues relative to the human sequences and 51% to the *C. elegans* sequence. Secondary structure predicted by the PHD program (Rost and Sander, 1993) is shown at the bottom as beta strands (arrows) and alpha helices (boxes). Safe segments (higher reliability) are in bold lines, less safely predicted regions in broken lines.

1989; Schmid and Linder, 1992). The family is part of a superfamily of NTPases involved in RNA and DNA metabolism (Koonin, 1991). The family includes protein functioning in RNA splicing, ribosome assembly, initiation of translation, etc. The best characterized activity for one of these proteins is that of an ATP-dependent RNA helicase (Ray *et al.*, 1985; Schmid and Linder, 1992).

The family has been extensively studied from the sequence point of view. At least three different subfamilies have been identified (DEAD, DEAH and DEAD/H; Gorbalenya *et al.*, 1989; Inoue *et al.*, 1992; Schmid and Linder, 1992; Bork and Koonin, 1993). They differ in a number of conserved regions (see Figure 4). A clear function for these is not known, although some of them are likely candidates for ATP binding (motifs I and IV), or RNA binding (motif VII) (Koonin, 1991).

The new sequence YIB2c clearly belongs to this family (overall sequence identity with DEAD_ECOLI is 47%). The analysis of the characteristic sequence motifs shows a closer relation of YIB2c with the DEAD rather than with the DEAH or DEAD/H subfamilies. The similarity is based on the presence of the specific boxes II, III and V, but there are also clearly distinctive features, i.e. boxes VIIa and VIIb are missing. One characteristic of the new sequence is the presence of a His residue in the DEA pattern instead of the DEAD common to most members of the family; given its importance this position of the sequence was carefully checked in the raw data on both strands. Too little is known about the structure of helicases and about the specific role of the DEAD/H box to assert the functional importance of the H by D substitution found in YIB2c. The functions of YIB2c *in vivo* remain to be proven experimentally, but the present classification as a member of the DEAD/H helicase family could be useful in designing an experimental strategy and points to the existence of a new subfamily.

YIA2c belongs to the inositol polyphosphate-5-phosphatase family responsible for Lowe's syndrome in humans

YIA2c is 52% identical to inositol polyphosphate-5-phosphatase (Figure 5). Inositol polyphosphate-5-phosphatase (75 kDa) is one of the two forms in humans that catalyse the conversion from inositol trisphosphate to inositol bisphosphate.

The inositol polyphosphate-5-phosphatase (75 kDa) protein has extended sequence identity (26%) with a 100 kDa protein that was cloned (Attree *et al.*, 1992) after having been mapped by linkage analysis (Silver *et al.*, 1987; Reilly *et al.*, 1988, 1990) to the Xq25-q26 region in humans. This 100 kDa protein is responsible for the human Lowe's syndrome (Attree *et al.*, 1992; Indo and Matsuda, 1992; Leahey *et al.*, 1993). The syndrome is characterized by developmental disorders

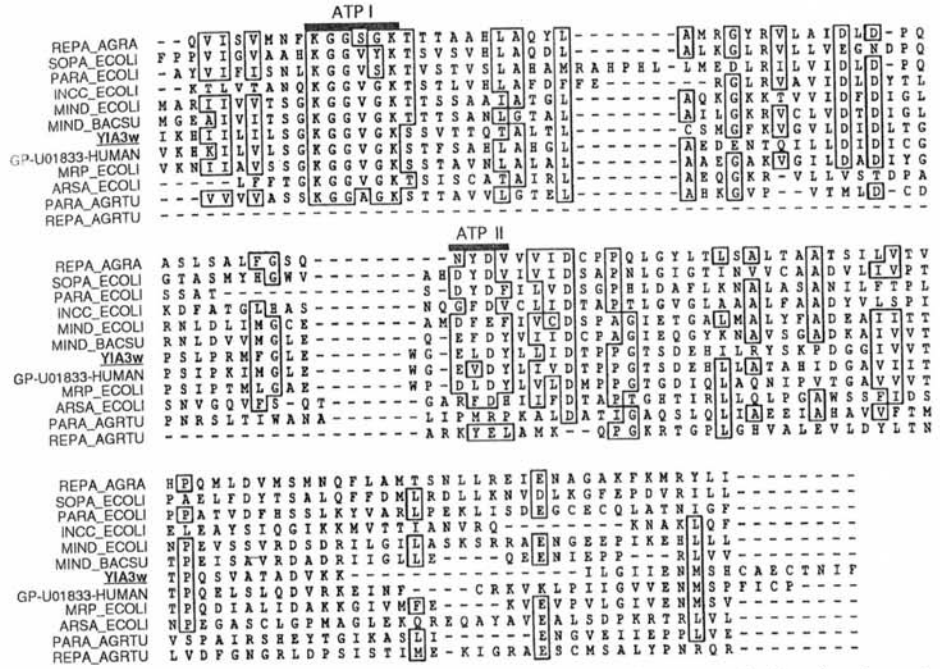


Figure 6. Conserved motifs in the MinD/Mrp family. YIA3w belongs to a family of prokaryotic ATPases, that now have a yeast (YIA3w) and a mammalian (GP-U01833-human) member. Regions common to other ATP-binding proteins are labelled (ATP1 and ATP2). YIA3w and the newly discovered human sequence fit perfectly the conserved sequence pattern, YIA3w has 61% sequence similarity with Mrp and 49% with MinD.

with effects in eye lens, brain and kidneys. Various mutations in this gene (*OCRL* gene) have been found in patients (Indo and Matsuda, 1992; Leahey *et al.*, 1993). The yeast YIA2c sequence adds new information about this important protein, e.g., better definition of conserved residues, and may provide an opportunity for studying the protein in a model experimental system.

Further searches in sequence databases revealed another protein in this family, an ORF of *C. elegans* (Favello, 1993). In Figure 5 the most conserved region is represented. A search with the pattern derived from the central region (WLGDFNY/FR) finds no other related proteins and can be proposed as a characteristic signature of the family. This patch of residues and other residues conserved in all sequences of the family are possible targets for mutagenesis experiments.

YIA3w is homologous to MinD, a prokaryotic cell cycle ATPase

Figure 6 shows the close similarity between YIA3w and a number of prokaryotic ATPases.

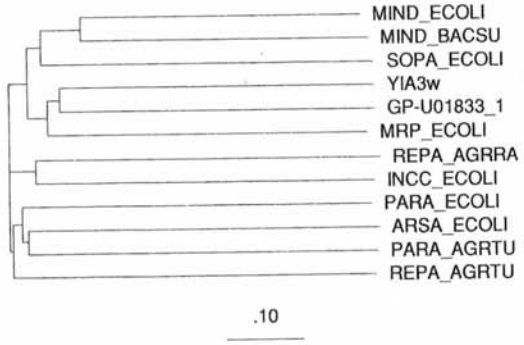


Figure 7. Phylogenetic tree of the MinD/Mrp family. The family tree was generated using the program CLUSTALW (Higgins *et al.*, 1992). All branches are stable at the 90% reliability level (Felsenstein, 1985). The branch including the human ORF, YIA3w, sopA, MinD and Mrp therefore represents a differentiated subfamily of bacterial ATPases. The scale at the bottom applied to the length of the branches roughly estimates the distance between sequences, i.e. MIND_ECOLI and MIND_BACSU are around 30% sequence similar.

The closer homologue is *MinD*, which forms part of the *Min* gene cluster responsible for correct placement of the septum in *B. subtilis* and *E. coli*.

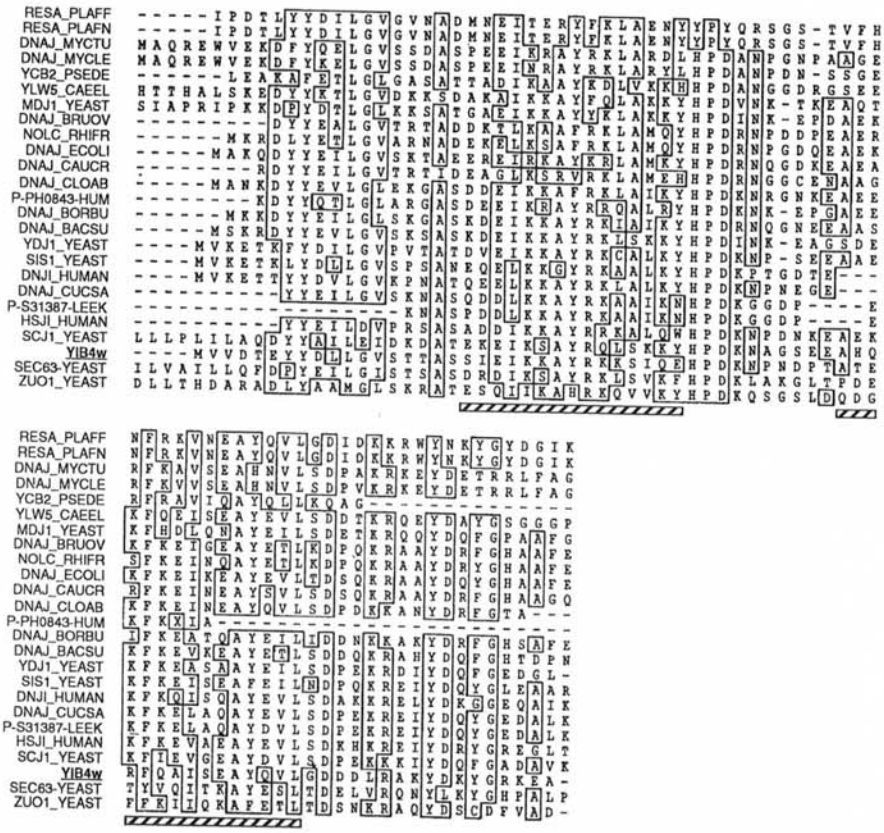


Figure 8. Alignment of the common domain of the DnaJ family. The aligned sequence regions correspond to the 'conserved DnaJ motif' (Figure 9). In this region, the sequences are suprisingly similar, complicating phylogenetic analysis (Figure 9). The motif is predicted to be all-helical (Rost *et al.*, 1994), with two long strongly predicted alpha-helices (expected three-state accuracy 88%).

The sequence corresponding to the *MinD* gene is 42% identical between *E. coli* and *B. subtilis* (Lee and Price, 1993). MinD is an ATPase (de Boer *et al.*, 1991) that interacts with FtsZ (de Boer *et al.*, 1990). This protein was proposed to belong to a more extended family (Motallebi-Vershareh *et al.*, 1990; de Boer *et al.*, 1991), including cell division proteins such as SopA and proteins involved in resistance to toxicity such as ArsA from *E. coli* (Mori *et al.*, 1986; Rosen, 1990). Our sequence search also shows extended similarity with an *E. coli* ORF close to the *metG* gene (mrp_ecoli) (Dardel *et al.*, 1990) of unidentified function, as well as similarity to a human ORF (GP:HO1833), which is the first higher eukaryotic sequence of the family (Figure 6).

All these sequences have in common a well-characterized ATP binding motif common to many other proteins (Rossmann *et al.*, 1974; Walker *et al.*, 1982; Motallebi-Veshareh *et al.*, 1990; de Boer *et al.*, 1991). In addition, there are at least two other well-conserved motifs (see Figure 6). Proteins in the family have different N-terminal and C-terminal extensions.

A phylogenetic analysis of this family of sequences (Figure 7) shows that SopA, Mrp and U01833_1 (a human sequence submitted to EMBL/GenBank by R. D. Howells, Sept. 93, see Figure 6) and YIA3w belong to the same group of sequences, in a stable tree (Figure 7). In particular, the similarity of YIA3w with *E. coli* Mrp is suggestive of a role in replication. The presence of human and yeast sequences in this family raises questions about their possible role in eukaryotic cell cycle control. It would be particularly interesting to know whether it is possible to find a

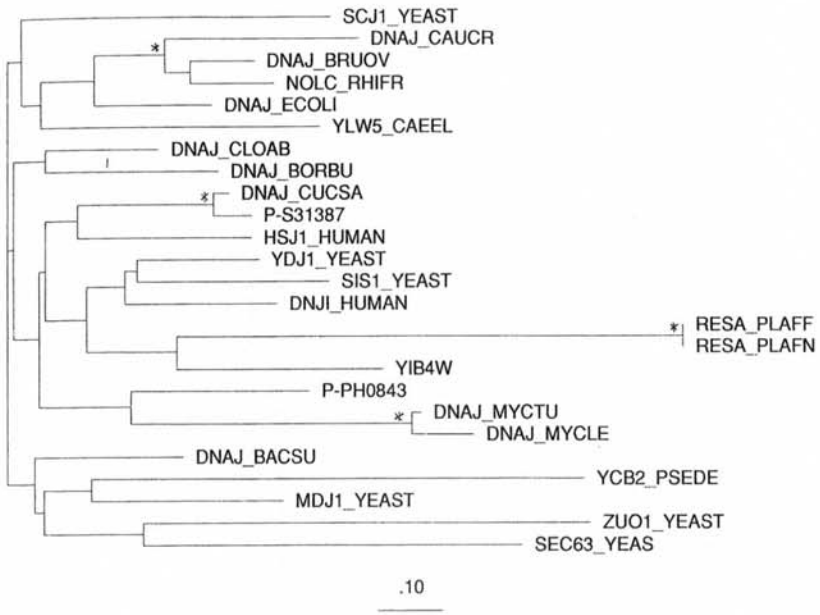


Figure 9. Phylogenetic tree of the DnaJ family. Tree based on the alignment in Figure 8. The yeast sequences are dispersed in different branches of the tree, but no clear specialization of function can be seen for the different groups. Interestingly, YIB4w is closest to the DnaJ from *Plasmodia*, ResA_plaff and ResA_plafn, which are outliers among the currently known family members. However, any functional deduction from this is tentative, as the order of branches in the tree is not particularly stable (see Methods) (Felsenstein, 1985). The scale at the bottom indicates sequence distance as in Figure 7.

molecular partner of YIA3w in yeast, sharing a similar relationship as the one that exists between FtsZ and MinD in *E. coli*.

YIB4w expands the DnaJ family

YIB4w has extensive sequence similarity with DnaJ proteins: in the conserved DnaJ domain YIB4w has 25% identical residues with *E. coli* and 22% with human DnaJ (Figures 8 and 9) (Ohki *et al.*, 1986; Liberek *et al.*, 1991; Oh *et al.*, 1993). DnaJ is implicated in heat-shock response (Rowley *et al.*, 1994) and aids the chaperone function of DnaK (Georgopoulos *et al.*, 1990; Cyr *et al.*, 1994). DnaJ stimulates ATP hydrolysis of DnaK (Liberek *et al.*, 1991) and also interacts directly with specific substrates (Georgopoulos *et al.*, 1990; Wickner, 1990; Cyr *et al.*, 1994).

The DnaJ family is present in many species, from *E. coli* to human. There are six related sequences in yeast (Sadler *et al.*, 1989; Blumberg and Silver, 1991; Caplan and Douglas, 1991; Luke *et al.*, 1991; Atencio and Yaffe, 1992; Zhang *et al.*, 1992; Rowley *et al.*, 1994; Schwarz *et al.*, 1994). The YIB4w protein appears to be most closely related to two ResA proteins from *Plasmodium* (tree in Figure 9), and together they constitute a distinct subfamily of DnaJ proteins. The distinct character of this subfamily is also apparent in the placement of the conserved DnaJ motif relative to other domains (Figure 10). The YIB4w sequence does not contain Gly-rich, Cys-rich, transmembrane or signal sequences. The organization of regions is similar to the one found in the *zuo1* yeast sequence.

The different yeast sequences appear to fulfil different roles in protein transport and heat-shock response (see Rowley *et al.*, 1994). The existence of new yeast sequences such as YIB4w opens new possibilities for the study of the function of DnaJ-like proteins in yeast. The mapping of the conserved residues on the predicted secondary structure elements (Figure 8) could be useful in planning site-directed mutagenesis experiments.

YIB9w is a single-domain U2 snRNP B

YIB9w corresponds to a fragment of DNA previously sequenced as part of an ORF upstream

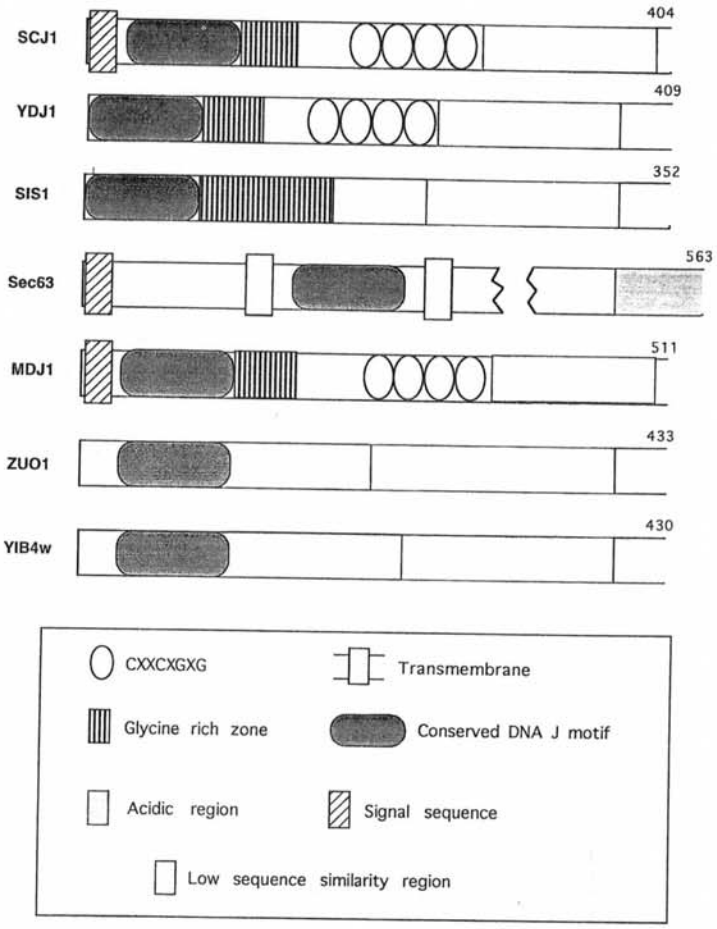


Figure 10. Domain distribution in different yeast DnaJ proteins. Variation in domain distribution of all known yeast DnaJ sequences: Scj1, Ydj1, Sis1, Mdj1, Zuo1, Sec63 and YIB4w. The new YIB4w sequence is most similar in organization to yeast Zuo1 and it does not have transmembrane, Cys-rich, Gly-rich, or acidic regions.

of the *pri1* gene (Plevani *et al.*, 1987). YIB9w is clearly homologous to one of the family of small nuclear ribonucleoproteins (snRNPs) (Figure 11); this family includes the U1 type A and U2 type B snRNPs. Proteins in this family are part of the complex that excises introns, binding to the 5′ intron–exon junction (Kramer *et al.*, 1984; Maniatis and Reed, 1987). U1 snRNP A binds to stem-loop I (Query *et al.*, 1989; Scherly *et al.*, 1989; Jessen *et al.*, 1991) and U2 snRNP B binds to stem-loop IV (Sillekens *et al.*, 1987).

Both types of snRNPs, U1A and U2B, contain two repeats separated by a sequence fragment of variable length. The N-terminal repeat is enough for RNA binding (Scherly *et al.*, 1989, 1990) and is stable enough to be crystallized (Nagai *et al.*, 1990; Jessen *et al.*, 1991). Each one of these repeats is around 90 residues long and contains two very conserved sequence motifs called RNP2 and RNP1 (Figure 11). The RNP motifs are responsible for RNA binding (Brunel *et al.*, 1985; Maniatis and Reed, 1987; Sillekens *et al.*, 1987), and different mutagenesis experiments have identified particular residues in these motifs that are implicated in specificity and affinity of RNA binding (Scherly *et al.*, 1990; Tan, 1982).

The length of the YIB9w sequence (112 amino acids) is equivalent to one of the repeats and indeed YIB9w contains the RNP2 and RNP1 motifs and shares extended sequence similarity

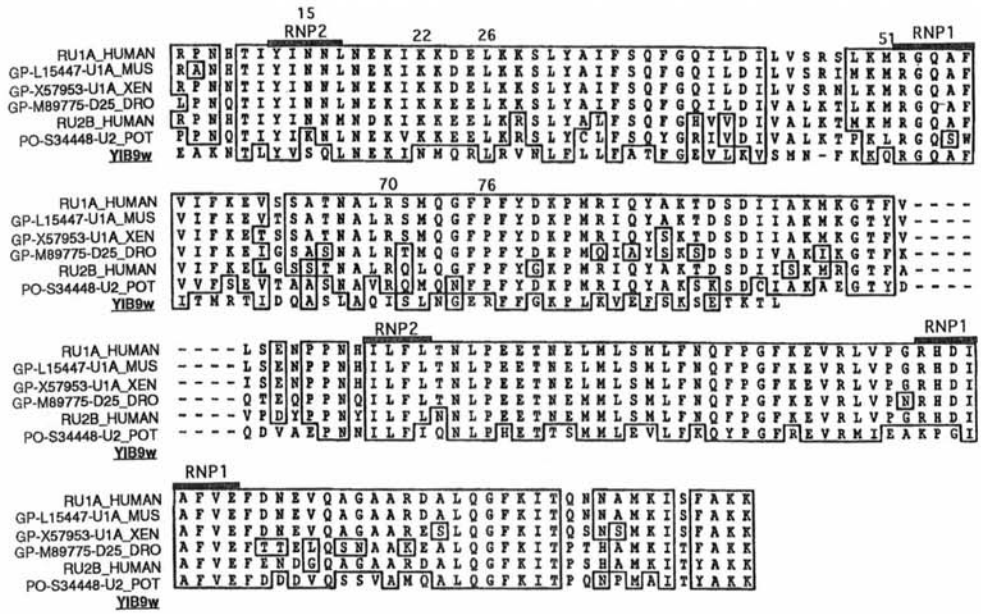


Figure 11. Multiple sequence alignment of the snRNP U1/U2 A/B′ family. The positions of the RNA-binding motifs (Dreyfuss *et al.*, 1988; Bandziulis *et al.*, 1989) are shown as RNP1 and RNP2. These sequences normally have two of these motifs corresponding to a two-domain organization; YIB9w appears to be the first single-domain protein in this family. YIB9w is very similar to the other sequences of the family, especially in the conserved domains; the few positions in which it differs from the others are labelled by their residue number (15, 51, 22–26 and 70–76) and are also shown in Figure 13.

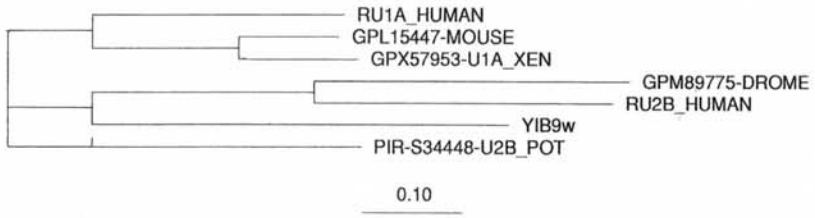


Figure 12. Phylogenetic tree of the snRNP U1/U2 A/B′ family. Tree based on sequences in Figure 11. The branches of the tree are stable at the 90% confidence level. As YIB9w is grouped in the tree with GPM89755_DROME and RU2B_HUMAN, it can be classified as a U2 snRNP type B′ protein.

with the N-terminal repeat of human sequence RU2B (31% identity). The N-terminal repeat is precisely the one that is necessary for specific binding to the cognate RNA molecule (Scherly *et al.*, 1989, 1990). So, YIB9w is the first single-domain member of the family.

Although both U1 snRNP-A and U2 snRNP-B′ are very similar, a phylogenetic analysis shows a reliable separation between them, with YIB9w classified as a U2 type B′ snRNP (Figure 12). The YIB9w is the first lower eukaryotic sequence known to belong to this class. This classification may be important as type B′ snRNPs are related to lupus erythematosus pathology in humans (Tan, 1982).

The three-dimensional structure of snRNP type A in humans (Protein Data Bank code 1NRC-A) is known (Stark *et al.*, 1992). Using the core of this structure as a framework for the family, one can construct a molecular model for YIB9w (Figure 13). The core of the RNA-binding site is located between two central beta strands and corresponds to the RNP sequence motifs. The positions in which YIB9w differs in sequence from the family

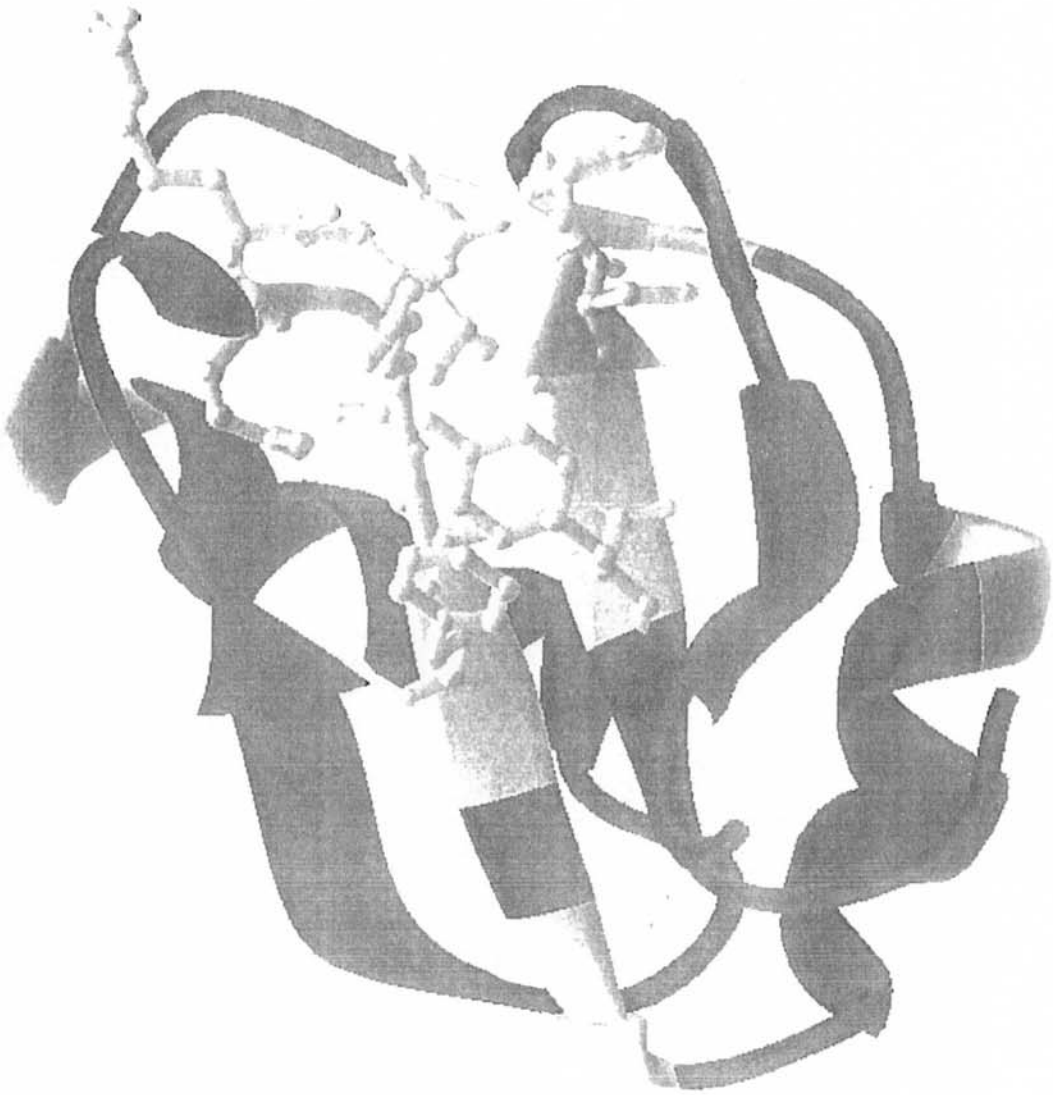


Figure 13. Three-dimensional model of YIB9w based on its homology with the snRNP1A 1NRC-A of human (Stark *et al.*, 1992). The model was generated replacing first all side-chains, trying to conserve as many atoms as possible in position and later optimizing side-chain positions choosing optimal rotamers and avoiding bumps with other residues as implemented in the package WHATIF (Vriend, 1990). The final model was regularized with a short energy minimization run. The figure represents the backbone of the model and the side-chains in the active site or most different between YIB9w and the rest of the family; residues are those given in Figure 11.

consensus are labelled in Figure 11 and highlighted in Figure 13. Most of these correspond to the beginning of two alpha helices (residues 22–25 and 70–76). Both helices are located far away from the active site and do not appear to correspond to different modes of RNA binding. Two sequence differences in YIB9w do map to the active site, residues 15 (S for N) and 51 (Q for M), and may indicate different specificity that could be subject to experimental analysis.

DISCUSSION

We have analysed in detail a fragment of 33·8 kb around the centromere of chromosome IX in

yeast. Maintaining a close collaborative interaction between experimental sequence determination and computational sequence analysis has proven useful in resolving difficult cases and differences to previously published data (e.g., *pan1*, DEAD/H helicase).

Use of frequently updated databases and advanced sequence search tools has enabled us to assign functions to 11 of the 14 ORFs (82%). This is an unusually high percentage compared to the average 55–60% level of functional ORF detection for yeast sequences (Bork *et al.*, 1992; Dujon *et al.*, 1994). The number of composition-biased proteins, i.e., proteins with amino acid compositions rich in particular amino acids, was correspondingly low—only one ORF (6%), compared to the yeast average of about 15%. Undoubtedly more information is needed to judge the significance of these local deviations.

In the computational sequence analysis, special emphasis was placed on family characterization of sequences. The resulting higher quality of functional classification demonstrates the relevance of massive sequencing toward obtaining new biological information. Examples are YIB4w as a new DnaJ protein, YIA3w as a member of the MinD family, the yeast homologue of the human OCRL, and the detection of new members of established families, such as human homologues of MinD or Pan1p. Family analysis is also a powerful tool for detecting conserved and divergent sequence regions and pinpointing residues important for protein function and specificity. This is particularly clear when three-dimensional information is available, as in the case of YIB4w, homologue of a U2 snRNP protein of known three-dimensional structure.

With respect to the types of proteins found in this fragment of chromosome IX, it appears that at least six of the 17 ORFs are RNA or DNA binding proteins of different types (including RNPs, helicase and primase). Indeed, there are three RNPs, and two of them (YIB1c and YIB5w) belong to the same family of heteronuclear RNA binding proteins. Such a concentration of related proteins has so far not been observed with sequences from other yeast chromosomes. However, the analysed region may be too short to give statistically significant data on this subject. The existence of functionally related clusters of sequences is an open possibility that sequencing projects and careful data analysis will be able to clarify in the future.

ACKNOWLEDGEMENTS

We are indebted to our colleagues at EMBL who developed search and analysis software and helped us in its use. We are particularly grateful to Georg Casari, Reinhard Schneider and Michael Scharf for maintaining the software and updating the databases. We also thank Christos Ouzounis and Peter Rice for critical reading of the manuscript. This work was carried out with financial support of the EC Biotech Program.

REFERENCES

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. and Weng, J. (1987). *Crystallographic Databases. Information Content, Software Systems, Scientific Applications*. Data Commission of the International Union of Crystallograpy, Bonn.

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

Ansorge, W., Voss, H., Wiemann, S., Schwager, C., Sproat, B., Zimmermann, J., Erfle, H., Hewitt, N. and Rupp, T. (1992). High-throughput automated DNA sequencing with fluorescent labels at the EMBL. *Electrophoresis* **13,** 616–619.

Atencio, D. P. and Yaffe, M. P. (1992). MAS5, a yeast homolog of DNA J involved in mitochondrial protein import. *Molec. Cell Biol.* **12,** 283–291.

Attree, O., Olivos, I. M., Okabe, I., Bailey, L. C., Nelson, D. L., Lewis, R. A., McIness, R. R. and Nussbaum, R. L. (1992). The Lowe's oculocerebrorenal syndrome gene encodes a protein highly homologous to inositol polyphosphate-5-phosphatase. *Nature* **358,** 239–242.

Bairoch, A. and Boeckmann, B. (1993). The SWISS-PROT protein sequence data bank, recent developments. *Nuc. Acids Res.* **21,** 3093–3096.

Bandziulis, R. J., Swanson, M. J. and Dreyfuss, G. (1989). RNA-binding proteins as developmental regulators. *Genes Dev.* **3,** 431–437.

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

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

Bissinger, P. H. and Kuchler, K. (1994). Molecular cloning and expression of the *Saccharomyces cerevisiae* STS1 gene product. A yeast transporter conferring mycotoxin resistance. *J. Biol. Chem.* **269,** 4180–4186.

Blumberg, H. and Silver, P. A. (1991). A homologue of the bacterial heat-shock gene DNA J that alters protein sorting in yeast. *Nature* 627–629.

Boguski, M. S., Lowe, T. M. J. and Tolstoshev, C. M. (1993). dbEST database for expressed sequence tags. *Nature Genet.* **4,** 332–333.

Bork, P. and Koonin, E. V. (1993). An expanding family of helicases within the DEAD/H superfamily. *Nuc. Acids Res.* **21,** 751–752.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E. (1992). Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Protein Sci.* **1,** 1677–1690.

Brunel, C., Sri-Widada, J. and Jeanteur, P. (1985). *Progress in Molecular and Subcellular Biology.* Springer-Verlag, Heidelberg.

Caplan, A. J. and Douglas, M. G. (1991). Characterization of YDJ1: A yeast homologue of the bacterial DNA J protein. *J. Cell. Biol.* **114,** 609–621.

Cyr, D. M., Langer, T. and Douglas, M. G. (1994). DNA J-like proteins: Molecular chaperones and specific regulators of Hsp70. *Trends. Biochem. Sci.* **19,** 176–181.

Dardel, F., Panvert, M. and Fayat, G. (1990). Transcription and regulation of expression of the *Escherichia coli* methionyl-tRNA synthetase gene. *Mol. Gen. Genet.* **223,** 121–133.

de Boer, P. A. J., Crossley, R. E., Hand, A. R. and Rothfield, L. I. (1991). The minD protein is a membrane ATPase required for the correct placement of the *Escherichia coli* division site. *EMBO J.* **10,** 4371–4380.

de Boer, P. A. J., Crossley, R. E. and Rothfield, L. I. (1990). Central role for *Escherichia coli* minC gene product in two different cell division-inhibition systems. *Proc. Natl Acad. Sci. USA* **87,** 1129–1133.

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

Dreyfuss, G., Swanson, M. J. and Pinol-Ruma, S. (1988). Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem. Sci.* **13,** 86–91.

Dujon, B. *et al.* (1994). The complete DNA sequence of chromosome XI of *Saccharomyces cerevisiae. Nature* **369,** 371–378.

Favello, A. D. (1993). Submitted to Swissprot data bank, P34370.

Fazioli, F., Minichiello, L., Matoskova, B., Wong, W. T. and Difiore, P. (1993). *Molec. Cell. Biol.* **13,** 5814–5828.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39,** 783–791.

Georgopoulos, C., Ang, D., Liberek, K. and Zylic, M. (1990). *Stress Proteins in Biology and Medicine.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. and Blinov, V. M. (1989). Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. *Nuc. Acids Res.* **17,** 4713–4730.

Higgins, D. G., Bleasby, A. J. and Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8,** 189–191.

Indo, Y. and Matsuda, I. (1992). *Nippon-Rhinso.* **50,** 3086–3092.

Inoue, S. B., Sakamoto, H., Sawa, H. and Shimura, Y. (1992). Nucleotide sequence of a fission yeast gene encoding the DEAH-box RNA helicase. *Nuc. Acids Res.* **20,** 5841.

Jessen, T.-H., Oudbridge, C., Teo, C. H., Pritchard, C. and Nagai, K. (1991). Identification of molecular contacts between the U1 A small nuclear ribonucleoprotein and U1 RNA. *EMBO J.* **10,** 3447–3456.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

Koonin, E. V. (1991). Similarities in RNA helicases. *Nature* **352,** 290.

Kramer, A., Keller, W., Appel, B. and Luhrmann, R. (1984). The 5′ terminus of the RNA moiety of U1 small nuclear ribonucleoprotein particles is required for the splicing of messenger RNA precursors. *Cell* **42,** 725–736.

Leahey, A. M., Charnas, L. R. and Nussbaum, R. L. (1993). Nonsense mutations in the OCRL-1 gene in patients with the oculocerebrorenal syndrome of Lowe. *Hum. Mol. Genet.* **2,** 461–463.

Lee, S. and Price, C. W. (1993). The minCD locus of *Bacillus subtilis* lacks the minE determinant that provides topological specificity to cell division. *Molec. Microbiol.* **7,** 601–610.

Liberek, K., Marszalek, J., Georgoulos, C. and Zylic, M. (1991). *E. coli* DnaJ and GrpE heat shock proteins jointly stimulate ATPase activity of DnaK. *Proc. Natl Acad. Sci. USA* **88,** 2874–2878.

Luke, M. M., Sutton, A. and Arndt, K. T. (1991). Characterization of SIS1, a *Saccharomyces cerevisiae* homologue of bacterial DNA J proteins. *J. Cell. Biol.* **114,** 623–638.

Maniatis, T. and Reed, R. (1987). The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature* **325,** 673–678.

Mori, H., Kondo, A., Ohshima, A., Ogura, T. and Hiraga, S. (1986). Structure and function of the F plasmid genes essential for partitioning. *J. Mol. Biol.* **192,** 1–15.

Motallebi-Vershareh, M., Rouch, D. A. and Thomas, C. M. (1990). A family of ATPases involved in active partitioning of diverse bacterial plasmids. *Molec. Microbiol.* **4,** 1455–1463.

Nagai, K., Oudbridge, C., Jessen, T. H., Li, J. and Evans, P. R. (1990). Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. *Nature* **348,** 515–520.

Newman, A. P., Groesch, M. E. and Ferro-Novick, S. (1992). Bos1p, a membrane protein required for ER to

Golgi transport in yeast, co-purifies with the carrier vesicles and with Bet1p and the ER membrane. *EMBO J.* 11, 3609–3617.

Oh, S., Iwahori, A. and Kato, S. (1993). Human cDNA encoding DnaJ protein homologue. *Biochim. Biophys. Acta* **1174,** 114–116.

Ohki, M., Tamura, F., Nishimura, S. and Uchida, H. (1986). Nucleotide sequence of the *E. coli dnaJ* gene and purification of the gene product. *J. Biol. Chem.* **261,** 1778–1781.

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

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85,** 2444–2448.

Plevani, P., Francesconi, S. and Lucchini, G. (1987). The nucleotide sequence of the *PRI1* gene related to DNA primase in *Saccharomyces cerevisiae. Nuc. Acids Res.* **15,** 7975–7989.

Query, C. C., Bentley, R. C. and Keene, J. D. (1989). A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNA protein. *Cell* **57,** 89–101.

Ray, B. K., Lawson, T. G., Kramer, J. C., Cladaras, M. H., Grifo, J. A., Abramson, R. D., Merrick, W. C. and Thach, R. E. (1985). ATP-dependent unwinding of mRNA stucture by eucaryotic initiation factors. *J. Biol. Chem.* **260,** 7651–7658.

Reilly, D. S., Lewis, R. A., Ledbetter, D. H. and Nussbaum, R. L. (1988). Tightly linked flanking markers for the Lowe oculocerebrorenal syndrome, with application to carrier assessment. *Am. J. Hum. Genet.* **42,** 748–755.

Reilly, D. S., Lewis, R. A. and Nussbaum, R. L. (1990). Genetic and physical mapping of the Xq24-q26 markers flanking the Lowe oculocerebrorenal syndrome. *Genomics* **8,** 62–70.

Rice, C. M., Fuchs, R., Higgins, D. G., Stoher, P. J. and Cameron, G. N. (1993). The EMBL data library. *Nuc. Acids Res.* **21,** 2967–2971.

Rosen, B. P. (1990). *Res. Microbiol.* **141,** 336–341.

Rossmann, M. A., Moas, D. and Olsen, K. W. (1974). Chemical and biological evaluation of a nucleotide-binding protein. *Nature* **250,** 194–199.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232,** 584–599.

Rost, B., Sander, C. and Schneider, R. (1994). PHD—an automatic mail server for protein secondary structure prediction. *CABIOS* **10,** 53–60.

Rowley, N., Prip-Buus, C., Westermann, B., Brown, C., Schwarz, E., Barrell, B. and Neupert, W. (1994). Mdj1p, a novel chaperone of the DNA J family, is involved in mitochondrial biogenesis and protein folding. *Cell* **77,** 249–259.

Sachs, A. B. and Deardorff, J. A. (1992). Translation initiation requires the PAB-dependent poly(A) ribonuclease in yeast. *Cell* **70,** 961–973.

Sadler, I., Chiang, A., Kurihara, T., Rothblatt, J., Way, J. and Silver, P. (1989). A yeast gene important for protein assembly into the endoplasmatic reticulum and the nucleus has homology to DNA J, an *Escherichia coli* heat shock protein. *J. Cell. Biol.* **109,** 2665–2675.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425.

Sander, C. and Schneider, R. (1993). The HSSP data base of protein structure-sequence alignments. *Nuc. Acids Res.* **21,** 3105–3109.

Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C. and Sander, C. (1994). Genequiz: A workbench for sequence analysis. *2nd International Conference on Intelligent Systems for Molecular Biology,* in press.

Scherly, D., Boelens, W., van Venrooij, W. J., Dathan, N. A., Hamm, J. and Mattay, I. W. (1989). Identification of the RNA binding segment of the human U1A protein and definition of its binding site on U1 snRNA. *EMBO J.* **8,** 4163–4170.

Scherly, D., Dathan, N. A., Boelens, W., Van Venrooij, W. J. and Mattaj, I. W. (1990). The U2B″ RNP motif as a site for protein–protein interactions. *EMBO J.* **9,** 3675–3681.

Schmid, S. R. and Linder, P. (1992). DEAD protein family of putative RNA helicases. *Mol. Microbiol.* **6,** 283–292.

Schwarz, E., Westermann, B., Caplan, A. J., Ludwig, G. and Neupert, W. (1994). XDJ1, a gene encoding a novel nonessential DNA J homolog from *Saccharomyces cerevisiae. Gene,* in press.

Sillekens, P. T. G., Habets, W. J., Beijer, R. P. and Van Venrooij, W. J. (1987). cDNA cloning of the human U1 snRNA-associated A protein: extensive homology between U1 and U2 snRNP-specific proteins. *EMBO J.* **6,** 3841–3848.

Silver, D. N., Lewis, R. A. and Nussbaum, R. L. (1987). *J. Clin. Invest.* **79,** 282–285.

Stark, W., Pauptit, A., Wilson, K. and Jansonius, J. N. (1992). The structure of neural protease from *Bacillus cereus* at 0·2 nm resolution. *Eur. J. Biochem.* **207,** 781.

Tan, E. M. (1982). *Adv. Immunol.* **33,** 167–240.

Voss, H., Wiemann, S., Gothues, D., Sensen, C., Zimmermann, J., Schwager, C., Stegemann, J., Erfle, H., Rupp, T. and Ansorge, W. (1993). Automated low-redundant large-scale DNA sequencing. *BioTechniques* **15,** 714–721.

Vriend, G. (1990). WHAT IF: a molecular modelling and drug design program. *J. Mol. Graphics* **8,** 52–56.

Vriend, G. and Sander, C. (1993). Quality control of protein models: directional atomic contact analysis. *J. App. Cryst.* **26,** 47–60.

Walker, E. J., Saraste, M., Runwick, M. J. and Gay, N. J. (1982). Distantly related sequences in the α- and β- subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1,** 945–951.

Wong, W. T., Kraus, M. H., Carlomagno, F., Zelano, A., Druck, T., Croce, C. M., Huebner, K. and Di Fiore, P. P. (1994). The human *eps15* gene, encoding a tyrosine kinase substrate, is conserved in evolution and maps to 1p31–p32. *Oncogene* **9,** 1591–1597.

Wootton, J. C. and Federhen, S. (1993). *Comput. Chem.* **17,** 149–163.

Wustinger, K. and Spevak, W. (1993). Isolation of CENIX and CENXII from *Saccharomyces cerevisiae. Nuc. Acids Res.* **21,** 3221.

Zhang, S., Lockshin, C., Herbert, A., Winter, E. and Rich, A. (1992). Zuotin, a putative Z-DNA binding protein in *Saccharomyces cerevisiae. EMBO J.* **11,** 3787–3796.

Zimmermann, J., Dietrich, T., Voss, H., Erfle, H., Schwager, C., Stegemann, J., Hewitt, N. and Ansorge, W. (1992). Fully automated Sanger sequencing protocol for double stranded DNA. *Methods in Molec. Cell Biol.* **3,** 39–42.

Zimmermann, J., Wiemann, S., Voss, H., Schwager, C. and Ansorge, W. (1994). Improved fluorescent cycle sequencing protocol allows reading up to 1000 bases. *BioTechniques* **16,** in press.