YEAST VOL. **13:** 1363–1374 (1997)

Yeast Functional Analysis Reports

# Characterization of New Proteins Found by Analysis of Short Open Reading Frames from the Full Yeast Genome

MIGUEL A. ANDRADE[1]*, ANTOINE DARUVAR[1], GEORG CASARI[2], REINHARD SCHNEIDER[2], MICHEL TERMIER[3] AND CHRIS SANDER[1]

[1]*European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Cambridge CB10 1SD, U.K.*
[2]*European Molecular Biology Laboratory, D-69012 Heidelberg, Germany*
[3]*Institute de Génétique et Microbiologie, Bât. 400, Centre Universitaire d'Orsay, 91405 Orsay Cedex, France*

Received 14 February 1997; accepted 1 May 1997

We have analysed short open reading frames (between 150 and 300 base pairs long) of the yeast genome (*Saccharomyces cerevisiae*) with a two-step strategy. The first step selects a candidate set of open reading frames from the DNA sequence based on statistical evaluation of DNA and protein sequence properties. The second step filters the candidate set by selecting open reading frames with high similarity to other known sequences (from any organism). As a result, we report ten new predicted proteins not present in the current sequence databases. These include a new alcohol dehydrogenase, a protein probably related to the cell cycle, as well as a homolog of the prokaryotic ribosomal protein L36 likely to be a mitochondrial ribosomal protein coded in the nuclear genome. We conclude that the analysis of short open reading frames leads to biologically interesting discoveries, even though the quantitative yield of new proteins is relatively low. © 1997 John Wiley & Sons, Ltd.

Yeast 13: 1363–1374, 1997.
No. of Figures: 5. No. of Tables: 2. No. of References: 20.

KEY WORDS — *Saccharomyces cerevisiae*; short ORFs; computational ORF verification; ORF properties; sequence similarity

## INTRODUCTION

With the recent sequencing of whole genomes we can for the first time look at the entire set of instructions for construction and maintenance of organisms. One of the first steps towards understanding how this information is converted into a living cell is extracting the full set of proteins encoded by the genome.

The computational identification of DNA regions translated to proteins is typically based on heuristic rules. The simplest of such rules is taking DNA stretches between a start codon and the next stop codon, the open reading frames (ORFs). The translations of those ORFs are initially just hypothetical proteins. The biochemical characterization of the translation product definitely proves that an ORF actually codes for a protein (although significant sequence similarity to other proteins of known function provides strong circumstantial evidence). However, such experiments are often difficult, expensive, and not always feasible. Many ORFs resulting from genome projects accumulate in the databases as hypothetical proteins waiting to be confirmed or withdrawn.

*Correspondence to: Miguel A. Andrade, European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Cambridge CB10 1SD, U.K.

CCC 0749–503X/97/141363–12 $17.50

Table 1. Small-protein families with yeast members.

| Family | Subfamily | Length (aa) |
|---|---|---|
| GroES | | 90–100 |
| 40S subunit ribosomal proteins | S14 | 50–120 |
| | S27e | 82–86 |
| | S21e | 82–87 |
| | RS28 | 64–69 |
| 60S subunit ribosomal proteins | L37e | 50–100 |
| | L46e | 50 |
| | L40e | 52 |
| Mitochondrial ATPase complex, non-enzymatic component | Subunit 8 | 50–70 |
| | Subunit 9 | 50–80 |
| | Subunit ε | 50–60 |
| Cytochrome oxidase polypeptides | 6, 7 and 8 | 50–80 |
| DNA-directed RNA polymerases | I, II and III | 70 (yeast) |
| Metallothioneins | | 50–100 |
| Ubiquiniol-cytochrome *c* reductase complex | Subunit 7·3 kDa | 73 |
| | Subunit 8·5 kDa | 85 |
| Ubiquitins | | 76 |

The likelihood of an ORF to be fortuitous, the result of a random distribution of nucleotides, increases dramatically (exponentially) for short lengths. The ratio between expressed ORFs and the total number of ORFs is therefore highly dependent on the length and very small for short ones. Very long ORFs, however, are extremely unlikely to occur by chance. Their length is usually maintained as a result of strong selection acting on the encoded protein, which eliminates any mutant shortened by a new stop codon.

The empirical rule has been stated that any ORF more than 150 base pairs (bp) long is likely to code for a protein (Sharp and Cowe, 1991). Accordingly, a more stringent (supposedly safe) length threshold of 300 bp has been applied for ORF selection, for example in the European yeast genome project (Oliver *et al.*, 1992). However, this 300 bp threshold is too high given the existence of important protein families proven to be expressed which are shorter than 100 amino acid residues. Among those there is a great variety of biological functions (see Table 1). So, blind application of the 300 bp residues cut-off, originally chosen for simplicity, is not advisable.

Another good reason for carefully inspecting short ORFs lies in sequencing errors that can generate frame shifts and putative stop codons and thus apparent—but erroneous—short ORFs. A sequencing error of only one bp per 1000 bp sequenced would introduce approximately one single nucleotide error per protein. This means that in a big genome sequencing project many ORFs may be shortened by false stop codons or by frame shifts followed by an early out-of-frame stop codon (Dujon, 1996).

Unless the 100 amino acid cut-off is lowered, sequence analysis would miss all of the expressed short or artificially shortened ORFs. However, lowering the cut-off requires the analysis of a large number of fortuitous ORFs. To complicate matters, the expression of an ORF is not at all assured even if the corresponding product length is above the 100 residue cut-off. For example, in yeast most of the apparently spurious ORFs are among the ORFs of length 100 amino acids or close above which encode proteins that have not been biochemically characterized (see Figure 1a), and/or do not have any homolog in the protein sequence databases (Figure 1b), as noted by Das *et al.* (1997). So, how can one lower the length threshold and filter out spurious ORFs at the same time?

## METHODS

A partial solution to the problem of filtering spurious ORFs is found by computational ORF verification using our knowledge of biologically

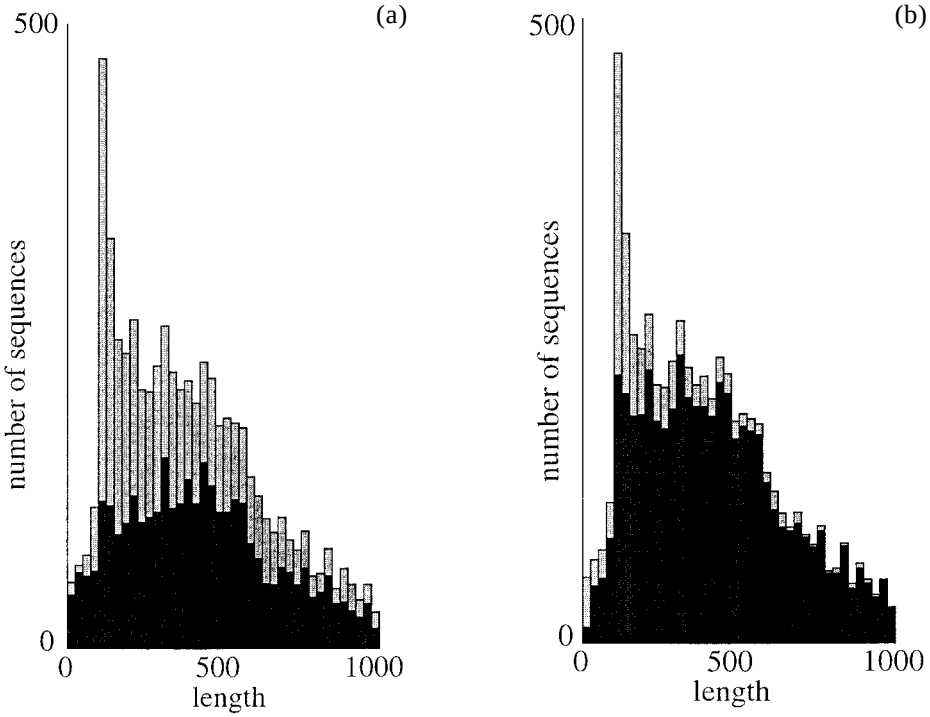



Figure 1. Evidence for spurious ORFs from the discrepancy between the length distribution of 'real' proteins (dark shading) and the distribution of putative proteins (light shading). (a) Distribution of yeast proteins in the protein databases according to their length up to 1000 amino acids. Each bar represents length increments of 25 amino acids. Dark shading indicates the number of sequences which have biological annotation. The total distribution has a sharp peak at a length of 100 amino acids, but not the distribution of annotated sequences, indicating possible artifacts in sequences near 100 amino acids length. The light (upper) part of the bar corresponds to the number of proteins not yet biologically characterized. Some of these are likely to be true proteins, i.e., ORFs actually transcribed and translated by the cell. For longer lengths (e.g. more than 500 amino acids) fortuitous ORFs are very unlikely and we can assume that the entire light-shaded bar at these lengths corresponds to true proteins. Assuming that the number of true proteins not yet characterized is proportional to the number of the already characterized proteins (dark shaded part) at all lengths, we have estimated that only approximately a 55% of the non-characterized proteins correspond to true proteins (10% of these are in the 100–120 amino acid range). (b) As in part (a) but here the shaded parts correspond to sequences that have at least one homolog protein in the database (with or without functional annotation). Note that the bar for lengths from 100 to 120 amino acids has more than half of its proteins without homologs, indicating a large number of fortuitous ORFs. Protein sequence similarity across species is evidence of the real existence of homolog proteins, because such similarity is the indirect result of functional constraints in evolution, while the products of non-real ORFs do not have the same evolutionary constraints. Similar results were independently obtained by Das *et al.* (1997).

characterized proteins. Biological sequences are related by evolutionary links and are the subject of evolutionary constraints:

(1) These evolutionary constraints maintain particular characteristics in coding regions that can differentiate them from non-expressed sequences. These properties can be learned from known examples (Fickett, 1982).
(2) Genes related by their evolutionary origin in different species maintain homology relationships often detectable by sequence comparison, while spurious ORFs are typically not seen to be conserved between species since there is no constraint to keep them. If ORFs from different species are found to be similar, their corresponding proteins have higher chances to be expressed.

We exploit both of these effects for ORF verification.

(1) First, we use a method of Termier and Kalogeropoulos (1996) that applies the first principle to computational ORF verification. (An alternative method has already been applied to the analysis of eight of the 15 yeast chromosomes by Barry *et al*. (1996).) For each ORF three characteristics are measured: the codon bias index (CBI; Bennetzen and Hall, 1982) that measures specific preferences for particular codon usage; the monopeptide score (MPS) related to amino acid composition; and the dipeptide score (DPS) related to the dipeptide usage. To calibrate the method, these scores were compared between a set of biologically characterized sequences and a set of ORFs extracted from an artificially generated DNA sequence. The comparison results in the definition of cut-off values above which a score is interpreted to be indicative of real protein-encoding genes.

(2) We add a second level of verification by searching for homologies. We compare the ORFs selected by the method of Termier and Kalogeropoulos (1996) and their translation products to all sequences stored in the DNA and protein databases, respectively, using the GeneQuiz software system (Casari *et al*., 1996). GeneQuiz automatically applies sequence analysis methods to protein sequences and uses a set of rules to interpret and filter the results with the aim of attaching maximal functional information to protein genes. Such an automatic system is essential if large amounts of data have to be analysed, as is the case in the analysis of complete cellular genomes.

We have applied this two-step ORF verification strategy to the yeast genome, the first eukaryotic organism fully sequenced (Goffeau *et al*., 1996; MIPS, *http://www.mips.biochem.mpg.de/*). In their original work, Termier and Kalogeropoulos (1996) studied yeast ORFs between 100 and 200 amino acid residues long. Here, we apply their method to shorter sequences which required a refinement of their procedure. The thresholds in the CBI ($t_{CBI}$), MPS ($t_{MPS}$) and DPS ($t_{DPS}$) were raised for shorter sequences to account for the increasing likelihood of short ORFs to be spurious:

$$t_{CBI}(l) = t^0_{CBI} + (100 - l)0{\cdot}005$$
$$t_{MPS}(l) = t^0_{MPS} + (100 - l)0{\cdot}005$$
$$t_{DPS}(l) = t^0_{DPS} + (100 - l)2{\cdot}5$$

where $t^0_{CBI} = 0{\cdot}15$, $t^0_{MPS} = 0{\cdot}05$, and $t^0_{DPS} = -25$ are the thresholds of the original work (Termier and Kalogeropoulos, 1996) and $l$ is the length of the ORF in amino acid residues.

## RESULTS

We extracted all ORFs of length 150 to 300 bp, corresponding to 50 to 100 amino acids, starting with a methionine codon and ending with a stop codon from the complete yeast genome sequence. This resulted in 5420 ORFs of which only 2329 were further considered as they were not completely included in a longer ORF on another frame. The codon bias, mono- and dipeptide criteria reduced the set further to 886 ORFs.

All these 886 ORFs were analysed using the GeneQuiz system. A large fraction of these had no significant sequence similarity to any functionally characterized protein. With the current approach these ORFs cannot be confirmed as expressed proteins. Of the remaining ORFs, most were already described in the database, either as entire proteins or as exons of spliced proteins. The final fraction of ORFs, with apparent homology to functionally characterized sequences, but not yet described in the databases, represents a number of interesting cases.

In Table 2 we present ten new ORFs predicted to be protein genes with functional annotation from GeneQuiz (three of them validated as well by the analysis of Barry *et al*. (1996) as likely protein-coding). None of these proteins were in the current versions of the sequence databases (not even as part of a longer protein). They constitute a very low fraction of the total number of ORFs analysed (10/886). Their detection as probably expressed proteins is beyond the capabilities of non-automated analysis.

In the following paragraphs we describe in detail three of these findings that illustrate why looking for short ORFs is worthwhile: an entire small protein, a protein broken by a putative stop codon, and a protein with frame shifts. We also briefly present evidence for the remaining seven cases.

### *ORF coding for a small protein. Mitochondrial ribosomal protein homologous to prokaryotic L36*

An ORF with a translation product 93 amino acids long, P_0199094 (our notation: P for XVI, followed by a seven digit number for the starting nucleotide), was found on the W strand of the yeast chromosome XVI from positions 199094 to 199372 (see Figure 2a). This ORF was selected for

Table 2. Identified yeast proteins not present in the databases.

| Chr | From | To | s | l | CBI | MPS | DPS | Reason | Annotation of homolog | Homolog | l | Identifier |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I | 176844 | 177008 | C | 55 | *0·628 | *0·39 | *198 | Stop | PAU3 | PAU3_YEAST P25610 | 124 | |
| III | 113333 | 113626 | W | 98 | −0·161 | *0·59 | *55 | Frames | Cell division nuclear protein | DO34_YEAST P33309 | 311 | YCLSO1w |
| IV | 1013515 | 1013679 | C | 55 | *0·521 | −0·06 | −20 | Small | Salt-stress induced protein | EMBL Z25537 | 54 | |
| VIII | 115616 | 115894 | C | 93 | *0·244 | *0·17 | −87 | Small | Mitochondrial regulator of splicing | MRS5_YEAST P32830 | 109 | YHRS01c |
| X | 726833 | 727060 | W | 76 | 0·096 | *0·68 | *173 | Stop | Alcohol dehydrogenase | pir A55449 (*P. chrysosporium*) | 385 | |
| XI | 338830 | 339087 | C | 86 | −0·053 | *0·61 | −70 | Small | Expressed protein | YNZ7_CAEEL P45967 | 103 | YKLS05c |
| XIII | 808039 | 808188 | C | 50 | 0·029 | *0·41 | −63 | Small | Development related | SXLM_DROME P19340 | 48 | |
| XIV | 687243 | 687461 | C | 73 | 0·013 | 0·07 | *85 | Small | Ubiquitin | pir S55243 (*A. thaliana*) | 631 | |
| XVI | 20222 | 20380 | W | 53 | −0·055 | *0·43 | *146 | Frames | Aspartic proteinase 3 precursor | YAP3_YEAST P32329 | 569 | |
| XVI | 19904 | 199372 | W | 93 | 0·119 | *0·13 | −411 | Small | Ribosomal protein L36 | RL36_LACLA P27146 | 38 | |

Report on new proteins based on ORFs selected by sequence properties and for which GeneQuiz found evidence of functional homology in motif and/or sequence databases. Chr, yeast chromosome number; from–to, first and last nucleotide position of ORF relative to the beginning of the chromosome (MIPS June 1996 version, *http://www.mips.biochem.mpg.de/*); s, strand direction, W = Watson (one direction), C = Crick (opposite direction); l, length of the translated product in number of amino acids; CBI, codon bias index; MPS, monopeptide score; DPS, dipeptide score (Termier and Kalogeropoulos, 1996); reason why not found before either: false *stop* codon, product in different *frames*, or *small* protein ($l$<100 amino acids); closest homolog with some functional annotation given by SwissProt identifier and SwissProt accession number or other database accession number; *l*, length in amino acids of putative homolog. *Parameter value above the length dependent threshold used that led to the selection of the ORF. Identifier used by Barry *et al.* (1996) for the ORFs for which they presented evidence of protein-coding likelihood.

(a)

```
>P_0199094
 MFLQTLRLTM PRMFLHMKPS PITITRACTV PSLLSVAAPQ PALVAANRPL VFNRGFKVRT
 SVKKFCSDCY LVRRKGRVYI YCKSNKKHKQ RQG
```

(b)

```
P_0199094      56 FRVRTSVKKFGSDGYLVRRKGRVYIYGKSNKKHKQRQG 93
RL36_THETH      1 MKVRASVKRICDKCRVIRRHGRVYVIC.ENPKHKQRQG 37
RK36_ORYSA      1 MKIRASVRKIGTKCRIIRRRGRIRVIC.SNPKHKQRQG 37
RL36_HAEIN      1 MKVRASVKKMCRNCKIVKREGVMRVLC.SDPKHKQRQG 37
RK36_PEA        1 MKVAASVRKICEKCRLIRRRGRLLVIC.SNPKHKQRQG 37
RK36_EUGGR      1 MKIRSSVKKICNKCYLIRRKNNLLVVC.INNKHKQRQG 37
RL36_LACLA      1 MKVRPSVKPICEYCKVIRRNGRVMVIGPANPKHKQRQG 38
RK36_SPIOL      1 MKIRASVRPICEKCRLIRRRGRIIVIC.SNPKHKQRQG 37
RL36_MYCBO      1 MKVNPSVKPICDKCRLIRRHGRVMVIC.SDPRHKQRQG 37
RL36_BACST      1 MKVRPSVKPICEKCKVIRRRGKVMVIC.ENPKHKQRQG 37
RK36_MARPO      1 MKIRASVRKICENCRLIRRRRNIMVVC.SNPKHKQRQG 37
RK36_TOBAC      1 MKIRASVRKICEKCRLIRRRGRIIVIC.SNPRHKQRQG 37
RK36_ODOSI      1 MKVRPSVKKMCDKCRVIKRKGKIMVIC.PNAKHKQRQG 37
RL36_BACSU      1 MKVRPSVKPICEKCKVIRRKGKWMVIC.ENPKHKQKQG 37
RL36_ECOLI      1 MKVRASVKKLCRNCKIVKRDGVIRVIGSAEPKHKQRQG 38
RL36_CHLTR      1 MKVRASIKRICKDCKIIKRHGVNRVIC.INPKHKQRQG 37
RL36_MYCGE      1 MKVRASVKPICKDCRIIKRHRILRVIC.KTKKHKQRQG 37
RK36_ASTLO      1 MKVYSSVRKICKSCGLIRRHGKLFVRC.INSKHNQRQ. 36
RK36_EPIVI      1 MKIRASIRKICEKCRLICRRRHIIVIC.SNPRHKQRQG 37
RK36_PINTH      1 MKIRASIRRICGKCRPIRRRKRVMIIC.SNPRHKQKQG 37
```

Figure 2. A short ORF containing a small protein: predicted eukaryotic homolog of prokaryotic ribosomal protein L36. (a) P_0199094 is a 93 amino acids long predicted translation product from an ORF on strand W of the yeast chromosome X from positions 199094 to 199372. (b) Alignment of the C-terminal 38 amino acids of yeast P_0199094 to prokaryotic (SwissProt identifier RL36_) and chloroplast (SwissProt identifier RK36_) ribosomal 50S subunit proteins of the L36 family. Shading according to residue conservation. Sequence similarity to human and mouse ESTs, also detected, is not shown.

further analysis by GeneQuiz because of its amino acid composition (high MPS score).

GeneQuiz detected a clear sequence similarity of P_0199094 to the L36 ribosomal protein of *Lactococcus lactis* (SwissProt RL36_LACLA). All typical sequence motifs of ribosomal 50S subunit L36 proteins are present (prosite PS00828 (Bairoch and Bucher, 1994) and block BL00828 (Henikoff and Henikoff, 1994)). Subsequent multiple sequence alignment of the translation product to 19 homologs of L36 (using the MAXHOM alignment program; Sander and Schneider, 1991) confirmed these very high similarities. L36 proteins have a length of 36–38 residues and share the sequence motif of Figure 2b. The similarity of the P_0199094 C-terminal region to the L36 family proteins, measured in percentage of identical residues, ranges between 47% and 60%. This similarity is a little lower than the average level of identity among the L36 proteins of 64%, but indicative of functional homology.

The 50S ribosomal subunit is characteristic of prokaryotes or chloroplasts, but it has not been described in eukaryotes: the L36 protein family so far extends only over prokaryotes and chloroplasts. It is likely that this protein is one of the many yeast mitochondrial ribosomal proteins coded in the nucleus (Dang and Ellis, 1990). More evidence for the existence of a nuclear coded L36 homolog is given by significant sequence similarity of P_0199094 to human and mouse DNA from expressed sequence tags (ESTs; data not shown).

We note that P_0199094 has no relevant similarity to other proteins in the database except for the C-terminal 38 amino acids. Given the fact that the other members of the family are 36–38 amino acids long, this could indicate that the first, unmatched part of the translation product of P_0199094 is not part of the expressed protein. More likely, the full length protein may have a dual function. Interestingly, such dual function has been already experimentally proven for another nuclear coded yeast mitochondrial ribosomal protein, MrpS28, also larger than its *Escherichia coli* homolog (Huff *et al.*, 1993).

### *ORF with an erroneous stop codon. Alcohol dehydrogenase*

An ORF coding for a product 76 amino acids long, J_0726833, was found on the W strand of yeast chromosome X at positions 726833 to 727060. The ORF was selected because of its amino acid and dipeptide composition (MPS and DPS scores).

GeneQuiz reported clear sequence similarity of J_0726833 to the N-terminal end of several alcohol

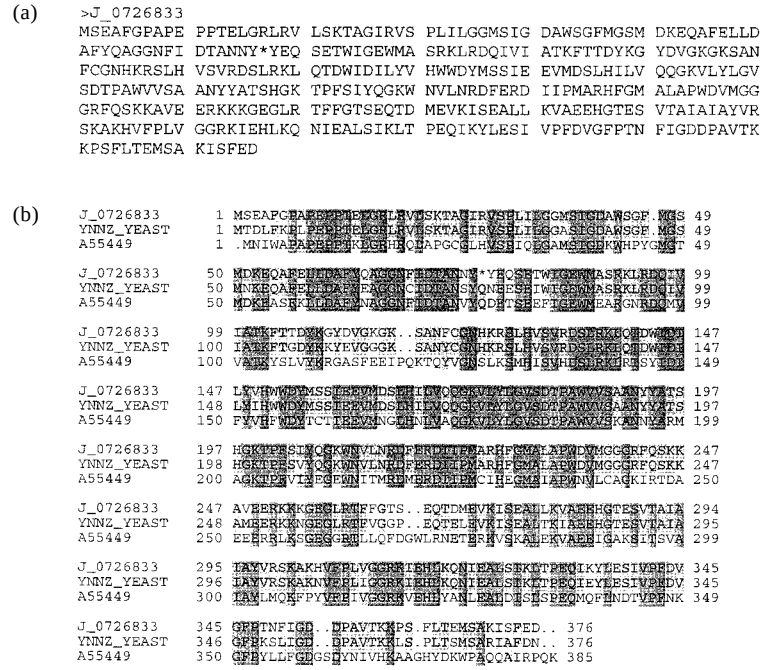



Figure 3. Reconstruction of a full-length protein by correction of an erroneous stop codon: predicted alcohol dehydrogenase. (a) Protein sequence constructed from artificial extension up to the following stop codon of an ORF encoding an apparent translation product 76 amino acids long on the W strand of the yeast chromosome X from positions 726833 to 727060. The first stop codon is shown by an asterisk at position 77. The construction has 376 amino acids and represents a predicted full-length protein sequence. The correct amino acid at the position of the first codon, probably Q, is not known. (b) Alignment of the extended ORF to yeast hypothetical protein YNNZ_YEAST (88% identity), and plant alcohol dehydrogenase (PIR accession number A55449; 53% identity), confirming the extension in (a).

dehydrogenases, the closest being an aryl-alcohol dehydrogenase (385 aa, PIR A55449) from another fungus (*Phanerochaete chrysosporium*). The N-terminus of yeast protein YNNZ_YEAST (376 aa), of unknown function, has even higher sequence similarity to J_0726833.

The striking difference in length between J_0726833 and the homologous dehydrogenases led us to inspection of the DNA downstream of this ORF. We extended the ORF past one stop codon up to the next stop codon. The construction results in an ORF of 376 aa, interrupted by a first stop codon at position 77 (see Figure 3a). The extended sequence has very high similarity along its entire length to four alcohol dehydrogenases and another four hypothetical proteins (see Figure 3b).

By ignoring a probable false stop codon, we have detected an evolutionary relationship of this piece of DNA with well-characterized proteins. If this ORF really codes for a functional protein we have an indication of a sequencing error that interrupted the ORF with a putative stop codon. Alternatively, this gene might have been inactivated and may actually be a pseudogene. A definitive decision between those two alternatives can only be made by a sequencing and/or expression experiment.

### *A frame-shifted ORF. Homolog to DOM34 yeast protein*

An ORF coding for a product apparently 98 amino acids long, C_0113333, was found on the W strand of the yeast chromosome III from positions 113333 to 113626. The ORF was selected according to its amino acid and dipeptide composition (MPS and DPS scores).

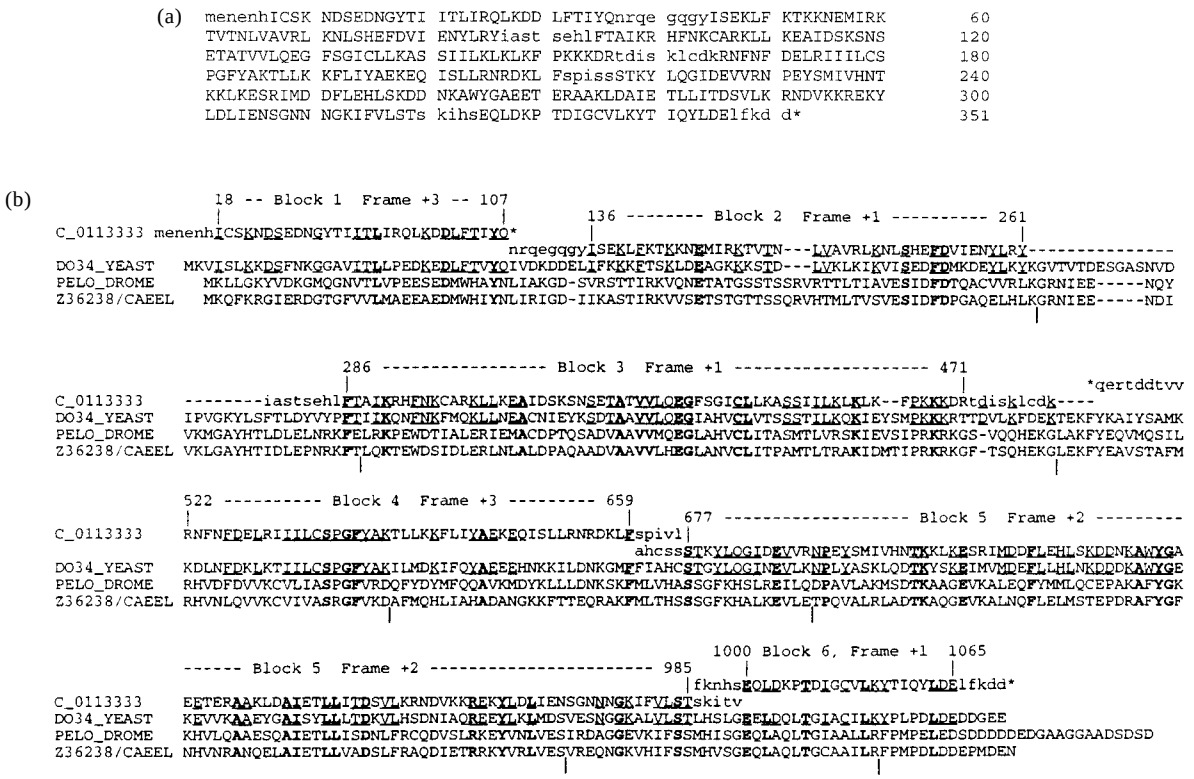



Figure 4. Multiple frame shifts from sequencing errors: predicted homology of yeast protein DOM34 or possible pseudo gene. (a) Predicted protein sequence if there have been sequencing errors in the DNA region of C_0113333. Capital letters stand for regions with sequence similarity to other proteins. (b) Alignment of DO34_YEAST, PELO_DROME (*Drosophila melanogaster*) and Z36238 (*Caenorhabditis elegans*), against translation products in the three W frames around ORF coding for C_0113333 in yeast chromosome III (from positions 112600 to 116765). Consecutive ordering of the hits gives six homology blocks. C_0113333 forms part of the fifth one. The homology was extended around the blocks using the translation products (small letters). There are at least four frame shifts. It is not possible to discern whether they are real or originate from sequencing errors. Some of them could be due to introns. In order to find indications of their existence, we compared the exon organization of the *Caenorhabditis elegans* sequence (EMBL Z36238) with the homology blocks found in C_0113333. Vertical bars at the bottom of the alignment mark the borders of the Z36238 sequence. No correlation in exon structure is evident. Identities across the four sequences are marked with bold characters. Identities between the yeast chromosome III translation products and the closest homolog DO34_YEAST are marked with underlined characters.

C_0113333 was functionally annotated by GeneQuiz as a cell division related protein because of its clear sequence similarity to yeast DOM34 (SwissProt identifier DO34_YEAST), a protein of 311 amino acids involved in mitosis and meiosis, and to pelota protein from *Drosophila melanogaster* (SwissProt identifier PELO_DROME, 395 amino acids long) also involved in cell division.

As in the previous case, we observe an important difference in size between the ORF translated from the yeast DNA and the closest homologs. Whereas the DOM34 and pelota genes code for their proteins in one continuous reading frame, the gene for a *Caenorhabditis elegans* homolog (EMBL Z36238, 381 amino acids long) is interrupted by seven introns. So we checked whether C_0113333 could correspond to an exon of a larger protein gene.

Investigation of the DNA downstream and upstream of the ORF showed a series of matches in the three W frames to the closest homolog, the DOM34 yeast protein, with at least six regions of clear matches (see Figure 4a). These data suggest that this ORF on chromosome III may be a protein gene, with its sequence interrupted by frame shifts at several places in the gene. Yeast chromosome III was the first to be completely sequenced and analysed with bioinformatics tools (Oliver *et al.*, 1992) and has been re-analysed

several times since then (Bork *et al*., 1992a,b; Koonin *et al*., 1994). Nonetheless, the existence of this protein gene apparently escaped the attention of experts, until the recent analysis by Barry *et al*. (1996) on eight yeast chromosomes, which reported in print for the first time the DNA region corresponding to C_0113333 as a likely coding region (labelled YCLS01w). Here we propose a possible sequence for the putative protein (see Figure 4b).

Rather than containing simple sequencing errors, this DNA region might code for a pseudogene. It is plausible that the DOM34 gene has been duplicated in yeast. The second copy might then have been knocked out by multiple frameshifts. Re-sequencing of this region of DNA is therefore suggested.

*Other predicted new proteins*

A_0176844 is an ORF on chromosome I from positions 176844 to 177008 on the C strand. The translation product has 55 amino acids. The most similar sequence found in the databases was PAU3 protein from yeast (SwissProt PAU3_YEAST), a protein of the seripauperin family (proteins similar to serine-rich proteins, but with poor serine content) 124 amino acids long. Extended sequence similarity to this homolog results if the translation product is extended past the first stop codon (Figure 5a), suggesting the presence of either a pseudo gene or a sequencing error. A total of six seripauperin proteins (named PAU1 to PAU6) are already known in yeast. This sequence could be called PAU7.

M_0808039 is an ORF on the C strand of chromosome XIII from positions 808039 to 808188 that encodes a very small product only 50 amino acids long (the lower length-limit of our analysis). The closest sequence in the databases is, interestingly, a similarly tiny protein of 48 amino acids from *D. melanogaster* (SwissProt SXLM_DROME) that controls sexual development, being active only in female specimens. The level of identity between these two sequences is not very high (29% of identical amino acids; Figure 5b). This fact, combined with the short length of the aligned sequences, makes this a borderline case. If the existence of this protein is proven, it may have an important role in the yeast cell cycle.

K_0338830 is located on the C strand of chromosome XI from positions 338830 to 339087. The putative protein has a length of 86 amino acids. Alignment to both the translation product of a mouse EST (EMBL MM0253) and to *C. elegans* hypothetical protein (SwissProt YNZ7_CAEEL) sequences, although in a short stretch, reveals the possibility of a repeat (see Figure 5c) with an interesting pattern of conserved cysteines.

N_0687243 is an ORF on chromosome XIV, on the C strand from positions 687243 to 687461. Its translation product is probably a ubiquitin 73 amino acids long, as shown by its similarity to the highly conserved pattern of similar ubiquitin sequences and repeats (see Figure 5d). A much closer sequence was found from a translation product from a mouse cDNA. Most ubiquitins from mammals, plants and fungi are extremely conserved and yeast ubiquitin (UBIQ_YEAST) is not an exception. Surprisingly, N_0687243 has a fairly low 20% amino acid identity to these. However, we found a closer sequence (64·4% identical), a translation product from mouse cDNA. These two, together, appear to be homologs of a near subfamily of proteins related to ubiquitins.

P_0020222 is on the W strand of chromosome XVI from positions 20222 to 20380. Its translation product would have 53 amino acids. However, the most similar sequence with characterized function is much longer: a yeast aspartic proteinase precursor (SwissProt YAP3_YEAST). Another related sequence from yeast, a hypothetical aspartic proteinase 3 precursor, 537 amino acids long (SwissProt YIV9_YEAST) is also much longer. The alignment of both of these longer proteins extends upstream of the P_0020222 ORF, past the putative initiating methionine, in the same frame for about 40 amino acids (Figure 5e). However, we did not find sequence similarity further upstream to the 300 N-terminal amino acids of the two yeast proteinases (not even in other translation frames). Because of the apparent lack of an upstream region of P_0020222 homolog to the N-terminal region of these larger proteins, we doubt that P_0020222 actually is translated.

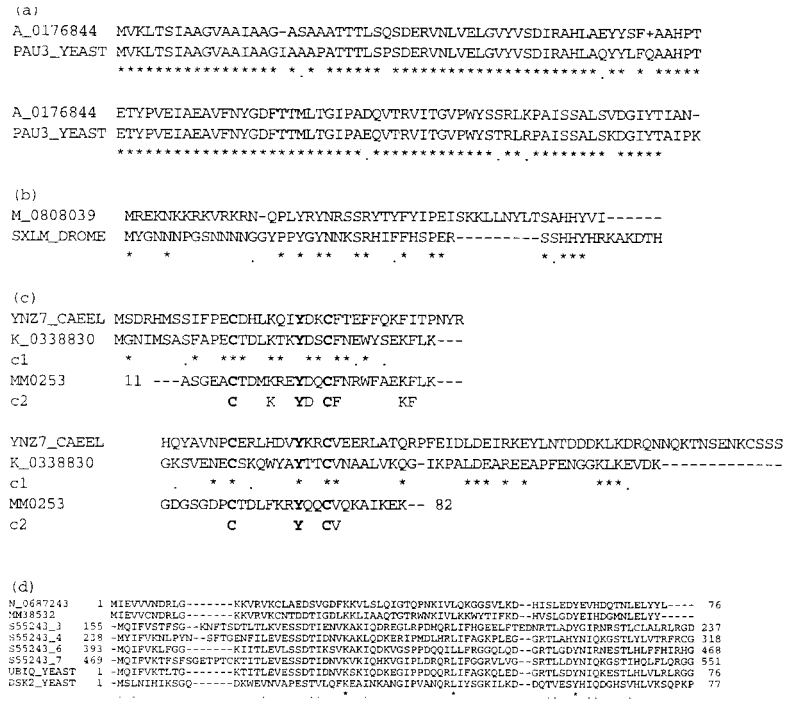
Our last two findings involve part of ESTs not present as proteins in the databases.

D_1013515 from the C strand of chromosome IV, at positions 1013515 to 1013679, is already identified as a part of a yeast EST (EMBL T36719). It translates into a product 55 amino acids long with clear sequence similarity to other ESTs from barley, and to both an EST and a hypothetical protein from *C. elegans*, the latter also relatively small (80 amino acids). The function of those proteins is unknown.

H_0115616 is from strand C of chromosome VIII located from positions 115616 to 115894. The 93 amino acid long translation product has significant sequence similarity to another yeast protein, a mitochondrial regulator of splicing (SwissProt MRS5_YEAST) of a similar size.

## DISCUSSION

We have used the yeast genome to show, by example, that the analysis of short ORFs and of ORFs shortened by apparent sequencing errors can lead to the identification of new protein



products. We have shown how some of these ORFs can be revealed using a combination of methods: first, selection of ORFs based on the analysis of ORF properties, followed by a powerful database search for sequence similarity.

We started with 2329 potential ORFs of length 150 to 300 bp from the complete yeast genome sequence, and selected 886 of those based on ORF compositional properties. Of these we predicted 135 coding regions using sequence similarity. Ten of these correspond to proteins not present in the databases. The number of findings is very small compared to the total number of sequences analysed. Such a discovery task can therefore only be performed with a highly efficient and automated procedure. The two-step selection procedure presented here fulfilled these requirements.

Although the proportion of new findings is small compared to the total number of yeast proteins (0·16%), they can contribute to the understanding of the organism and open surprising new aspects. The ribosomal L36-protein serves as an example: a protein that was not known to be present in eukaryotes is now predicted to exist in yeast.

The most stringent criterion in the presented ORF verification procedure is a homology search in public sequence databases (Bork *et al.*, 1995; Gelfand *et al.*, 1996). With the dramatic growth of sequence databases significant sequence similarity to database proteins can already be found for almost all proteins. Yet, because of incomplete database information, there remains a margin of unique, rare or previously unobserved real protein genes which cannot be verified by homology search.

The most popular current method to select ORFs from DNA sequences requires a length above the strict cut-off of 300 bp. This simple criterion ignores shorter genes, and, at the same time, results in many false positives of length 100 amino acid and just above. The procedure described here is quite expensive in terms of computational effort. This effort, however, is small relative to experimental efforts necessary to confirm or discard an ORF as a protein-encoding region.

There are, of course, other sequence signals characteristic of transcribed ORFs that may be useful in ORF verification, such as transcriptional and translational regulatory signals, the signals the cell actually reads as it makes proteins. We expect a clearer distinction between coding and non-coding ORFs if these signals are detected by sequence analysis, especially for genes for which currently no homolog is known.

In several cases, conflicts between homologous sequences have indicated likely sequencing errors. Such conflicts can only be resolved at the level of the actual sequencing data. Therefore, we encourage the use of bioinformatics in conjunction with sequencing experiments. This close interaction can lead to a reduction in sequencing errors by

Figure 5. Evidence for several short ORFs. Unless indicated, (−) indicates a gap in the alignment, and (*) and (.) indicate matching and similar amino acids, respectively, in the corresponding position of the alignment. (a) The translation product of A_0176844 was extended by 67 amino acids past one stop codon (+) up to the next. The resulting sequence is 87·9% identical to PAU3_YEAST over its entire length (122 amino acids). (b) Full-length alignment of the translation product of M_0808039 to SXLM_DROME, a 48 amino acids *Drosophila melanogaster* protein related to development. (c) Alignment of K_0338830 with the translation product of a mouse EST (frame +2, MM0253) and a *C. elegans* hypothetical protein (YNZ7_CAEEL). The pattern C-x5-Y-x2-C (where xn stands for a sequence of any n amino acids) appears twice in the three sequences with a separation of about 20 amino acids. c1, Consensus line for K_0338830 and YNZ7_CAEEL; c2, completely conserved positions in the three sequences. The alignment is displayed in two blocks to show the alignment between the first and second repeat. Bold letters are used for the positions conserved between the repeats. (d) MAXHOM alignment of the translation product of ORF N_0687243, 73 amino acids long, with several ubiquitins, which have a similar size, together with a very related mouse cDNA (EMBL MM38532) translated from bp 81 to 296. We included in the alignment four repeats out of the eight present in a poli-ubiquitin from *Arabidopsis thaliana* (PIR identifier S55243). The ubiquitin sequences and repeats used for the alignment were selected such that no pair had more than 70% identical amino acids. (e) Alignment of the translation product around P_0020222 (unique frame, tp line) to an aspartic proteinase precursor from yeast (YAP3_YEAST), 537 amino acids long. The consensus line (c) indicates the residues conserved in the three aligned sequences, and (.) marks the positions common between P_0020222 and the hypothetical proteinase precursor YIV9_YEAST (569 amino acids long). (+) are stop codons in the translation product. The unextended ORF P_0020222 starts at the MFVL (second block). Note the conserved DFGG pattern preceded by a region with high identity to YIV9_YEAST in the upstream region. (f) The translation product of D_1013515 is similar to a protein sequence (EMBL accession number U55854, gene C04G6.5, 58 amino acids long) and a hypothetical protein (YOT0_CAEEL), both from *Caenorhabditis elegans*, and to an EST from barley (EMBL Z25537, 53 amino acids long). c, Consensus line indicating the conserved residues and positions with hydrophobic residues, either ACIFLMVW (:). There are two boxes of conserved residues surrounded by hydrophobic positions. (g) Alignment of the translation product of H_0115616 to a mitochondrial regulator of splicing from yeast (MRS5_YEAST). The percentage of conserved positions is 34·5% over the first 88 amino acids of the alignment.

indicating regions of likely sequencing errors (Bork *et al*., 1995; Voss *et al*., 1997), followed by inspection of the original sequencing pieces.

Detection of small proteins and improvement in sequencing accuracy are the two rewards of the application of more elaborate systems for ORF extraction and analysis, contributing to the overall goal of having a correct and complete set of proteins for each of the fully sequenced organisms.

## REFERENCES


Bairoch, A. and Bucher, P. (1994). PROSITE: recent developments. *Nucl. Acids Res.* **22,** 3583–3589.

Barry, C., Fichant, G., Kalogeropoulos, A. and Quentin, Y. (1996). A computer filtering method to drive out tiny genes from the yeast genome. *Yeast* **12,** 1163–1178.

Bennetzen, J. L. and Hall, B. D. (1982). Codon selection in yeast. *J. Biol. Chem*. **257,** 3026–3031.

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

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

Bork, P., Ouzounis, C., Casari, G., *et al*. (1995). Exploring the *Mycoplasma capricolum* genome: a minimal cell reveals its physiology. *Mol. Microbiol.* **16,** 955–967.

Casari, G., Ouzounis, C., Valencia, A. and Sander, C. (1996). GeneQuiz II: automatic function assignment for genome sequence analysis. In *Proceedings of the First Annual Pacific Symposium on Biocomputing*. World Scientific, Hawaii, U.S.A, pp. 707–709.

Dang, H. and Ellis, S. R. (1990). Structural and functional analyses of a yeast mitochondrial ribosomal protein homologous to ribosomal protein S15 of *Escherichia coli*. *Nucl. Acids Res.* **18,** 6895–6901.

Das, S., Yu, L., Gaitatzes, C., *et al*. (1997). Biology's new Rosetta stone. *Nature* **385,** 29–30.

Dujon, B. (1996). The yeast genome project: what did we learn? *Trends in Genetics* **12,** 263–270.

Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucl. Acids Res.* **10,** 5303–5318.

Gelfand, M. S., Mironov, A. A. and Pevzner, P. A. (1996). Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA* **93,** 9061–9066.

Goffeau, A., Barrell, B. G., Bussey, H., *et al*. (1996). Life with 6000 genes. *Science* **274,** 546–567.

Henikoff, S. and Henikoff, J. G. (1994). Protein family classification based on searching a database of blocks. *Genomics* **19,** 97–107.

Huff, M. O., Hanic-Joyce, P. J., Dang, H., Rodrigues, L. A. and Ellis, S. R. (1993). Two inactive fragments derived from the yeast mitochondrial ribosomal protein MrpS28 function in trans to support ribosome assembly and respiratory growth. *J. Mol. Biol.* **233,** 597–605.

Koonin, E. V., Bork, P. and Sander, C. (1994). Yeast chromosome III: New gene functions. *EMBO J.* **13,** 493–504.

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

Sander, C. and Schneider, R. (1991). Database of homology-derived structures and the meaning of sequence alignment. *Proteins* **9,** 56–69.

Termier, M. and Kalogeropoulos, A. (1996). Discrimination between fortuitous and biologically constrained open reading frames in DNA sequences of *Saccharomyces cerevisiae*. *Yeast* **12,** 369–384.

Voss, H., Benes, V., Andrade, M. A., *et al*. (1997). DNA sequencing and analysis of 130 kilobases from yeast chromosome XV. *Yeast* **13,** 655–672.