Nucleic Acids Research, 1999, Vol. 27, No. 17 **3503–3509**

Origin and properties of non-coding ORFs in the yeast genome

Pawel Mackiewicz, Maria Kowalczuk, Agnieszka Gierlik, Miroslaw R. Dudek[1] and Stanislaw Cebrat*

Institute of Microbiology, Wroclaw University, ul. Przybyszewskiego 63/77, 54-148 Wroclaw, Poland and
[1]Institute of Theoretical Physics, Wroclaw University, pl. Maxa Borna 9, 50-204 Wroclaw Poland

Received May 10, 1999; Revised June 25, 1999; Accepted July 8, 1999

ABSTRACT

In a recent paper we have estimated the total number of protein coding open reading frames (ORFs) in the *Saccharomyces cerevisiae* genome, based on their properties, at about 4800. This number is much smaller than the 5800–6000 which is widely accepted. In this paper we analyse differences between the set of ORFs with known phenotypes annotated in the Munich Information Centre for Protein Sequences (MIPS) database and ORFs for which the probability of coding, counted by us, is very low. We have found that many of the latter ORFs have properties of anti-sense sequences of coding ORFs, which suggests that they could have been generated by duplication of coding sequences. Since coding sequences generate ORFs inside themselves, with especially high frequency in the antisense sequences, we have looked for homology between known proteins and hypothetical polypeptides generated by ORFs under consideration in all the six phases. For many ORFs we have found paralogues and orthologues in phases different than the phase which had been assumed in the MIPS database as coding.

INTRODUCTION

There are 7472 open reading frames (ORFs) longer than 100 codons in the yeast genome. About 3000 of these ORFs overlap. Since known examples of overlapping genes in the genomes thus far sequenced are very rare, it has been assumed that a given sequence of DNA codes for only one polypeptide. Thus, the shorter one of a pair of overlapping ORFs is usually considered non-coding. Another strong parameter indicating that an ORF is coding is its sequence homology to any other coding sequence. Attempts to estimate coding probability were made using the Codon Bias Index (CBI) (1) and the Codon Adaptation Index (CAI) (2). In yeast genome analysis it has been accepted that ORFs shorter than 150 codons with CAI < 0.11 and without any homology to known genes are non-coding (3). Nevertheless, there are a lot of ORFs longer than 100 codons in the yeast genome for which functions or homologues have not yet been found. These ORFs were called orphans (4,5). Sequencing of the *Saccharomyces cerevisiae* genome revealed a surprisingly high number of such sequences. Furthermore, during sequencing the fraction of orphans grew quicker than the fraction of homologues, which is a paradox because the more known genes, the higher fraction of homologues and the lower fraction of orphans should be found among newly sequenced ORFs. This phenomenon was called the mystery of orphans (4,5). There are two ways of solving the orphan paradox: (i) orphans belong to a group of genes which for unknown reasons escape function identification; (ii) orphans are mainly non-coding ORFs generated by non-random mechanisms in the yeast genome.

Prior publications of ours (6,7) provided a partial explanation of the mystery of orphans. Approximation of the number of protein coding ORFs longer than 100 codons in the yeast genome, based on their properties, yielded 4700–4800, a number much smaller than the widely accepted 5800–6000 (8–10). If it is acknowledged that there are only ~4800 genes in the yeast genome, the mystery of orphans disappears. There remain ~300 coding ORFs without known function or homology to already discovered genes, which is only ~5% of the total number of genes. The problem of the origin of non-coding ORFs in the yeast genome remained unclear. The probability of generating an ORF longer than 100 codons in a DNA sequence is low and in a random DNA sequence of the yeast genome size the number of randomly generated ORFs should be much smaller than the number of orphans in *S.cerevisiae*. In this paper we supply more arguments that the number of coding ORFs in the yeast genome is much smaller than previously estimated and that many non-coding ORFs were generated in the past inside coding sequences in non-coding frames and later transferred into other genome regions by duplication mechanisms.

DATABASES AND METHODS

DNA sequences and information on gene functions in the *S.cerevisiae* genome were downloaded from http://speedy.mips.biochem.mpg.de in June 1998.

The distributions in the two-dimensional space of the two sets of ORFs found in the yeast genome were prepared according to the method described by us in detail (6; also at our Web site http://smORFland.microb.uni.wroc.pl). In short, we have measured a parameter describing base composition for each

*To whom correspondence should be addressed. Tel: +48 071 3247 303; Fax: +48 71 3252 151; Email: cebrat@angband.microb.uni.wroc.pl

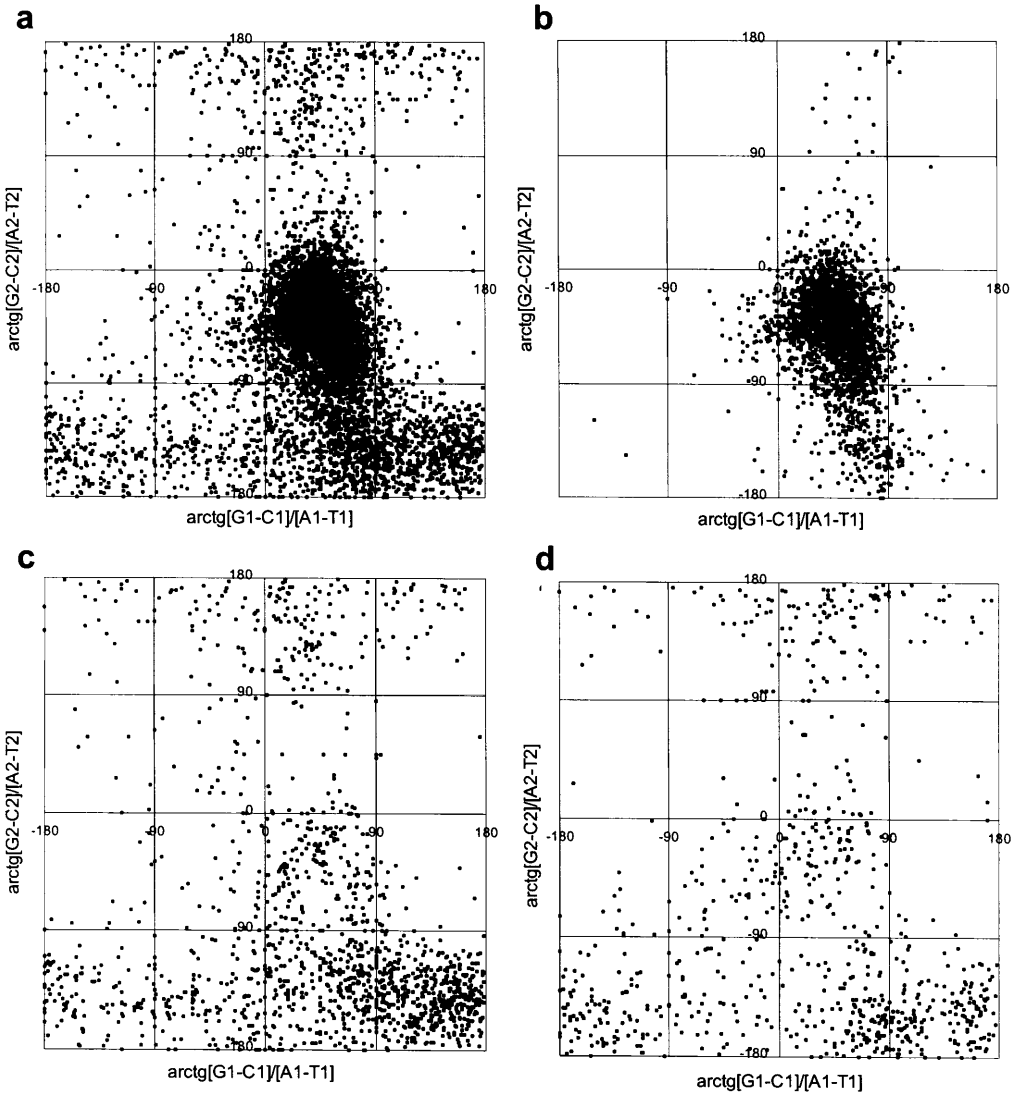

Figure 1. Distributions of different sets of ORFs on the torus surface projection. (**a**) Distribution of all 7472 ORFs longer than 100 codons; (**b**) distribution of all genes with known phenotypes; (**c**) ORFs discarded by MIPS as non-coding, obtained by subtracting all ORFs annotated in the MIPS database from the set of all ORFs longer than 100 codons found by us in the chromosome sequences [ORFs represented in (a)]; (**d**) set of ORFs obtained by subtracting all ORFs with described phenotypes (the 2513 genes known the day of data downloading) from the set of ORFs annotated by MIPS and subtracting a further 2600 ORFs drawn randomly from the rest of the ORFs according to their coding probability counted by us previously. This set represents ORFs which are annotated in MIPS, presumably coding according to the criteria of the database authors but with low coding probability according to our criteria.

ORF, arctan[(A-T)/(G-C)], for the first and second base position in codons separately. We have found that genes form a very compact set. A comparison of the distribution of genes and of all ORFs was the basis of the method of estimation of the total number of coding ORFs in the yeast genome and coding probability for ORFs, which was also previously described (7). The most important premises of the methods are briefly described in Results and Discussion.

ORFs with low estimated probability of coding were computer translated into theoretical peptides in all six frames. Amber and ochre stop codons appearing in some frames were translated into tyrosine and opal stop codons into tryptophan. Resulting polypeptide sequences were used to search databases for homologues using the FASTA search program (11).

RESULTS AND DISCUSSION

ORF distributions on the torus projection

To show some qualitative differences or resemblances between different sets of ORFs, we have prepared distributions of these sets of ORFs on the finite torus surface projection (Fig. 1). For each ORF longer than 100 codons found in the yeast genome, the numbers of each of the four nucleotides in each position in codons were counted. Next, the values of arctan[(G-C)/(A-T)] for the first and the second positions in codons were plotted against each other. The result was a distribution of points representing individual ORFs on the torus surface. In Figure 1a the distribution of all 7472 ORFs longer than 100 codons is shown. In Figure 1b the same distribution for all genes with

known phenotypes is presented. In Figure 1c ORFs discarded by MIPS as non-coding are shown. This set of ORFs was obtained by subtracting all ORFs annotated in the MIPS database from the set of all ORFs longer than 100 codons found by us in the chromosome sequences (ORFs represented in Fig. 1a). It represents mainly shorter ORFs from pairs or from larger clusters of overlapping ORFs or some ORFs shorter than 150 codons with CAI < 0.11 (3). The latter rule was not respected rigidly, since we have found more than 350 ORFs still fulfilling these criteria in MIPS. In general, discrimination of shorter ORFs of overlapping pairs is a good method of eliminating non-coding ORFs. Nevertheless, there are some pairs of overlapping ORFs in which the coding ORF is the shorter one. Furthermore, some ORFs shorter than 150 codons with CAI < 0.11 have identified phenotypes. Another set of ORFs (Fig. 1d) was obtained by subtracting all ORFs with described phenotypes (the 2513 genes known the day of data downloading) from the set of ORFs annotated by MIPS and subtracting a further 2600 ORFs drawn randomly from the rest of the ORFs according to their coding probability counted by us previously (6). This set represents ORFs which are annotated in MIPS, presumably coding according to the criteria of the database authors but with low coding probability according to our criteria. Note that the used method of random selection leaves some coding ORFs in this set and eliminates from it some ORFs with lower coding probability. Nevertheless, Figure 1c resembles Figure 1d quite well, especially if one compares these two sets, for contrast, with Figure 1b, representing ORFs with described phenotypes.

Nucleotide composition of different sets of ORFs

It is well known that there is a prevalence of purines in the first positions in codons in all genomes thus far sequenced (6,12–15; see also http://smORFland.microb.uni.wroc.pl for detailed information). In Table 1 we have shown which fractions of ORFs fulfil the composition rule for the first positions in three different sets of ORFs from the yeast genome: *genes* represents ORFs with known phenotypes; *A>3* represents ORFs with the distance to the distribution centre of genes in Figure 1b larger than 3 SD (according to us the probability of coding for proteins by these ORFs is of the order of 0.001); *all ORFs eliminated by MIPS* shows ORFs longer than 100 codons not included in the MIPS database.

Table 1. Fractions of ORFs fulfilling the composition rule of the first positions in three different sets of ORFs from the yeast genome.

	A>T	G>C	A>T and G>C
Genes	0.97	0.96	0.92
A>3	0.46	0.59	0.24
ORFs eliminated by MIPS	0.55	0.74	0.40

Above 92% of ORFs with known phenotypes in the yeast genome have more A than T and more G than C in the first positions of codons. In contrast, only 24% of ORFs with very low coding probability fulfil these rules. It is an even much lower fraction than among ORFs eliminated by MIPS. For a random sequence the expected fraction of ORFs fulfilling the rules that the first positions are richer in both purines is 25%.

We have noted previously that long protein coding sequences generate ORFs inside themselves, especially in the antisense sequences, with much higher probability than happens in random sequences (16,17). Following the rule that the coding strand is richer in purines, the 'coding' strand of an ORF generated in the antisense sequences has to be richer in pyrimidines. That is why we suppose that some ORFs with low coding probability were generated inside protein coding genes and that duplication mechanisms are responsible for transferring them into the intergenic space.

Furthermore, the distribution of length for ORFs with MIPS annotation '1' (with known phenotype) is quite different than that for ORFs with low coding probability (Fig. 2). Note that our method of coding probability estimation does not take into account the length of ORFs; the only length condition is that an ORF should be longer than 100 codons. Thus, if our method cannot distinguish non-coding ORFs properly, there should be no correlation between coding probability and the length of ORFs. In fact, the average length of ORFs with annotation '1' is ~560 codons, while for non-coding ORFs according to our method it is ~200 codons (Fig. 2).

There is another conclusion which could be drawn from Figure 2. The distribution of ORF length annotated in MIPS implicates a growing number of coding ORFs in classes near 100 codons, which suggests that there are a lot of coding ORFs shorter than 100 codons. In the distribution of ORFs with relatively higher coding probability counted by us (with SD < 2), the numbers of coding ORFs in the length classes near 100 diminish. Inversely, the fraction of ORFs with low coding probability grows rapidly when their length diminishes from 200 to 100 codons, suggesting that the number of coding ORFs shorter than 100 codons is not very high.

Correlation between the number of presumably coding ORFs and the number of identified transcripts

If the predicted number of coding ORFs in the yeast genome is correct, the number of presumably coding ORFs in a group of ORFs should be correlated with the number of transcripts found for such a group. To test this we have grouped ORFs into classes of 200, according to their distance from the centre of gene distribution (Fig. 1b) and counted the expected number of coding ORFs in each class. Next we counted the number of transcripts found in the same class of ORFs (data obtained from: http://www.sagenet.org/yeast). The correlation coefficient between the two sets of data, presumably coding ORFs and ORFs with identified transcripts, was 0.96. The correlation coefficient counted for randomly grouped sets of ORFs was of the order of 0.1. Since the distance of an ORF from the distribution centre is the measure of its coding probability, the found correlation means that the probability of finding transcripts for ORFs is correlated with our estimation of its coding probability.

Generated ORFs in the yeast intergenic space

It was found previously that intergenic sequences in different genomes have significant traces of coding history (18,19). It is known that even in intergenic sequences triplet structure can be detected by looking for a correlation in DNA molecules. We have found that yeast intergenic sequences also share this feature with coding sequences. Furthermore, in intergenic yeast sequences it is much easier to find long ORFs than in random DNA sequences.

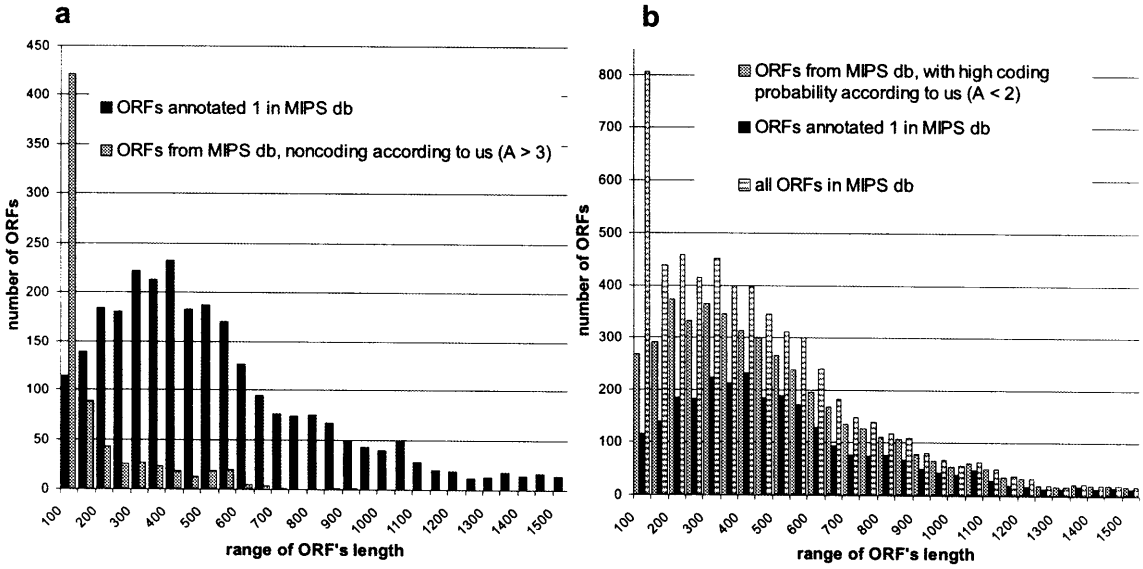

Figure 2. Distribution of the length of ORFs >100 codons in the yeast genome. (**a**) Comparison of two sets of ORFs, black columns representing ORFs annotated in MIPS as '1' and shaded columns representing ORFs annotated in MIPS as '2'–'6' but with very low probability of coding according to our method; (**b**) distributions of ORFs annotated '1' (black) and with high coding probability according to our method (shaded) among all ORFs annotated in the MIPS database.

Nucleotide composition of ORFs shorter than 100 codons found inside intergenic sequences indicates that they were generated by coding sequences in the past. To show this, we have used a computer generated DNA sequence (CGS) of the same length and nucleotide composition as the yeast genome. Next we compared the distribution of ORFs depending on their length in the CGS and in the real yeast genome (Fig. 3a). In the yeast genome the number of longer ORFs is much higher than in the CGS, which is a trivial statement, since the long ORFs are the result of selection. However, when all ORFs longer than 100 codons were excluded from the genome, many shorter ORFs were also eliminated (Fig. 3b). That is because these eliminated ORFs had been nested in longer ones. In fact, the rest of the sequence is mainly non-coding intergenic sequence. However, in this sequence we have still observed far more longer ORFs than in the CGS. The two distribution lines for the CGS and for intergenic sequences of the same length and composition cross at the position of ORFs 32 triplets long (Fig. 3c). Starting from this point, intergenic sequences contain more ORFs. The difference in the number of ORFs longer than 32 triplets is ~9000. Our results suggest that most of the 9000 short ORFs are remains of coding sequences and have been generated by them.

Coding sequences and the genetic code have specific properties of generating long ORFs in the antisense (complementary) strand of genes (16). There are ~3000 overlapping ORFs (longer than 100 codons) in the yeast genome. Many ORFs in the intergenic space were probably generated in the past in non-coding frames of protein coding sequences and later transferred into other genome regions by duplication mechanisms. We may assume that the coding sequences nesting ORFs (we have called these smaller ORFs baby ORFs) were at least partially duplicated. Duplicated sequences accumulated mutations which eventually eliminated the reading frames of the original gene, leaving baby ORFs. That is why ORFs in the intergenic space, like baby ORFs, are richer in pyrimidines than in purines. Then, it should be possible to find homology between generated ORFs and their maternal gene.

Homologues of products of 'non-coding' frames

All yeast ORFs were divided by MIPS (http://www.mips.biochem.mpg.de/) into six classes, according to the properties of the encoded protein: (i) known function or phenotype; (ii) strong similarity to known protein; (iii) similarity or weak similarity to known protein; (iv) similarity to unknown protein; (v) no similarity; (vi) questionable ORF.

Similarities have been measured by FASTA scores. A questionable ORF is defined by a combination of the following attributes: low CAI value; partial overlap with a longer ORF or with an ORF with known phenotype; no similarity to other ORFs.

Many properties of ORFs located long distances from the distribution centre of ORFs with known phenotypes (Fig. 1) indicate that they form significantly distinct sets of ORFs. Most of these ORFs belong to classes described in the MIPS database with a number higher than '3'. Since these ORFs read in their phases (determined by the start codon in MIPS) do not fit the parameters of coding sequences described by us, we have assumed that there exists another frame in which the original 'generating coding sequence' coded for the protein, which means that some ORFs with low coding probability were in the past generated in non-coding frames of coding sequences. Thus, we have looked for homologues of these ORFs, not only for translation products in their assumed reading frames, but in all six frames. To find such homology we have translated ORF sequences in six frames and performed a FASTA search of all six theoretical peptide sequences versus the full databases (11). The generated amber and ochre stop

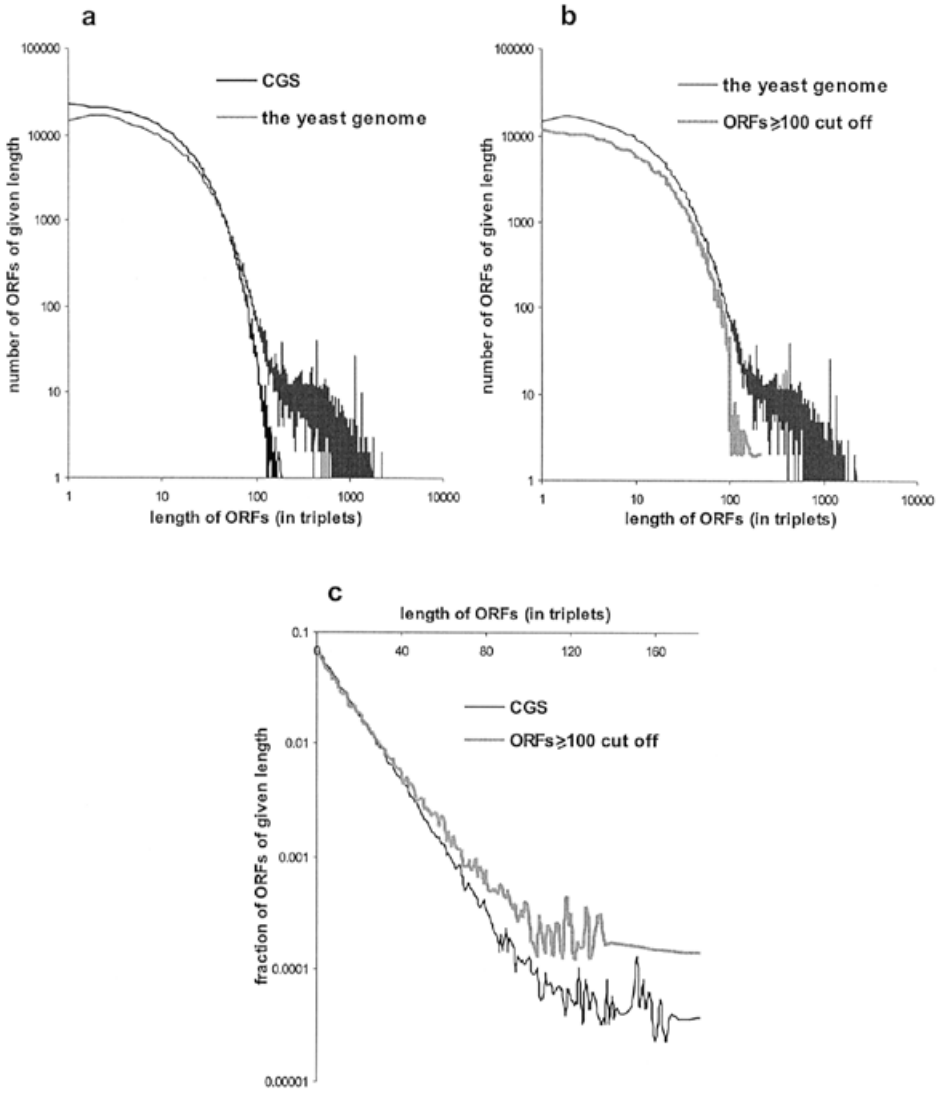

Figure 3. Distribution of ORF length in the yeast genome sequences and the sequence generated by computer (CGS). (**a**) Distribution of ORFs depending on their length in the CGS and in the real yeast genome; (**b**) distribution of ORFs in the whole yeast genome and in the sequence after all ORFs longer than 100 codons were excluded; (**c**) distribution lines for intergenic sequences and for the CGS of the same length and composition [note the scale of (c) has been changed to show the differences in ORF length distributions for shorter ORFs].

codons (TAA and TAG) were translated into tyrosine while opal (TGA) was translated into tryptophan.

We have analysed 2840 ORFs annotated classes 3–6 in MIPS. The excluded ORFs belong to classes 1 and 2, with known function and/or strong similarities to known proteins.

The frame assumed by MIPS as coding was considered *frame 1* by us. For each analysed ORF we have chosen the best homologue out of all the homologues found in the remaining five frames (2–6). We have accepted an E value < 0.001 as indicating homology between the analysed ORF product and the sequence found in the databases (Fig. 4). Homologues of the products of 783 ORFs have been found translated in phases 2–6 with an E value < 0.001 (947 for E < 0.01). However, this set of ORFs contains both generated and generating ORFs. We have assumed arbitrarily that an ORF was generated if its class annotation in MIPS is higher than that of its homologue or if its length is smaller than the length of the found homologue. In this way we have selected 603 generated ORFs for which we have found homologues with E < 0.001 (757 with E < 0.01). Among these ORFs, 426 do not have better homologues in the frame which was assigned by MIPS as coding (frame 1). Sequences homologous to products of translation in frames 2–6 were both paralogues and orthologues. Many homologues found by this technique were simply the longer ORFs of the pair of overlapping ORFs, which means that it should be decided which one of the two overlapping ORFs should be discarded from the database or at least to which ORF the overlapping sequence should be assigned.

Some homologues have long stretches of the same amino acid or short repetitions. It is difficult to prove the phylogenetic

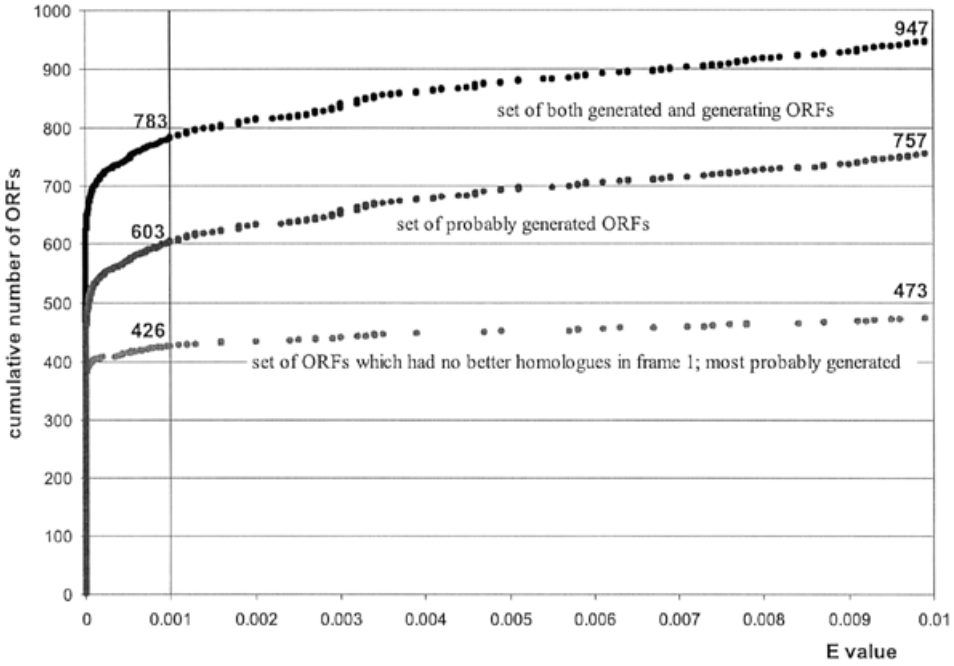

Figure 4. Cumulative numbers of yeast ORFs translated in phases 2–6 for which homologues of their hypothetical products were found.

relation between such pairs of proteins but their close similarities may suggest that some of the yeast ORFs were generated by specific non-stop codon sequences (i.e. some regulatory sequences) and do not have protein coding function. We have found that many ORFs located in subtelomeric parts of chromosomes (with specific repetitions) have homologues in different frames to other ORFs in these regions and to yeast delta and core X elements. Many ORFs annotated in the MIPS database are localised in this region, but in our opinion these sequences play a structural role rather than coding. There are many premises suggesting their structural role: (i) they consist of 72 bp repetitions in which 36 or even 12 bp repetitions are recognisable; (ii) inside these repetitions some positions are extremely conservative (100% consensus); (iii) when these sequences are analysed in triplets as coding sequences, many conserved positions are at the third positions of codons while some, not conserved, are at the first or second codon positions; (iv) 72 bp repetitions are organised in higher order and, for example, the external Y' element of the left arm of chromosome 12 has two 144 bp repetitions with 135 matched positions; (v) Y' elements from different chromosomes differ in length and these differences are always a multiple of 72 bp (isn't it one turn around the nucleosome?).

Thirty-six homologues were found in virus genomes, suggesting horizontal transfer of non-coding ORFs.

The most pronounced results have been obtained for ORFs annotated in MIPS class '6' (questionable ORFs). In this group were classified ORFs without any known homology. Thus, by definition, we have not found homologues of these ORFs in databases if translated in their own reading frame. But homologues were found for almost 80% of these sequences translated in other reading frames (mostly in the antisense). For other groups of ORFs annotated in MIPS we have also found homologues in other reading frames, but with lower frequency.

In a large majority of the cases of homology we found (i.e. for 70% of ORFs annotated in MIPS as class '6'), we were able to properly predict in which frame the homology should be expected using our method of content coding sequence analysis.

About 50% of generated ORFs originated in the sixth phase of genes and 28% in the fourth phase, as predicted previously (17). Most short overlapping baby ORFs originated in the sixth phase.

In general, we have found ~700 ORFs in the MIPS database which have homologues 'not in frame'. That is why we conclude that most ORFs annotated in the MIPS database which according to our gene identification method are not coding are in fact ORFs generated in the past by coding sequences. Thus, most of the so-called orphans are not coding sequences.

DATA AVAILABILITY

Data concerning the ORF co-ordinates in the yeast genome, their distances from the centre of yeast gene distribution on the torus surface as well as their coding probability counted according to our method with a lot of other technical information are available at our Web site: http://smORFland.microb.uni.wroc.pl

ACKNOWLEDGEMENT

This work was supported by The State Committee for Scientific Research, grant no. 6 P04A 030 14.

REFERENCES

1. Benetzen,J.L. and Hall,B.D. (1982) *J. Biol. Chem.*, **257**, 3026–3031.
2. Sharp,P.M. and Li,W.H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.
3. Dujon,B., Alexandraki,D., Andre,B., Ansorge,W., Baladron,V., Ballesta,J.P., Banrevi,A., Bolle,P.A., Bolotin-Fukuhara,M., Bossier,P. *et al*. (1994) *Nature*, **369**, 371–378.
4. Casari,G., de Druvar,A., Sander,C. and Schneider,R. (1996) *Trends Genet.*, **12**, 244–255.
5. Dujon,B. (1996) *Trends Genet.*, **12**, 263–270.
6. Cebrat,S., Dudek,M.R., Mackiewicz,P., Kowalczuk,M., Fita,M. (1997) *Microbial Comp. Genomics*, **4**, 259–268.
7. Kowalczuk,M., Mackiewicz,P., Gierlik,A., Dudek,M.R. and Cebrat,S. (1999) *Yeast*, in press.
8. Goffeau,A., Barrel,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., Louis,E.J., Mewes,H.W., Murakami,Y., Philippsen,P., Tettelin,H. and Oliver,S.G. (1996) *Science*, **274**, 546.
9. Winzeler,E.A. and Davis,R.W. (1997) *Curr. Opin. Genet. Dev.*, **7**, 771–776.
10. Mewes,H.W., Albermann,K., Bähr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G., Pfeiffer,F. and Zollner,A. (1997) *Nature*, **387** (suppl.), 7–8.
11. Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
12. Zhang,C.T and Zhang,R. (1991) *Nucleic Acids Res.*, **19**, 6313–6317.
13. Gutierrez,G., Marquez,L. and Martin,A. (1996) *Nucleic Acids Res.*, **24**, 2525–2528.
14. Mrazek,J. and Karlin,S. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 3720–3725.
15. Wang,J. (1998) *J. Biomol. Struct. Dyn.*, **16**, 51–57.
16. Cebrat,S. and Dudek,M.R. (1996) *Trends Genet.*, **12**, 12.
17. Cebrat,S., Mackiewicz,P. and Dudek,M.R. (1998) *Biosystems*, **45**, 165–176.
18. Buldyrev,S.V., Goldberger,A.L., Havlin,S., Mantegna,R.N., Matsa,M.E., Peng,C.-K., Simons,M. and Stanley,H.E. (1995) *Phys. Rev. E*, **51**, 5084.
19. Voss,R. (1992) *Phys. Rev. Lett.*, **68**, 3805–3808.