Yeast
Yeast 2002; **19**: 619–629.
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/yea.865

Research Article

How many protein-coding genes are there in the *Saccharomyces cerevisiae* genome?

Pawel Mackiewicz[1], Maria Kowalczuk[1], Dorota Mackiewicz[1], Aleksandra Nowicka[1], Malgorzata Dudkiewicz[1], Agnieszka Laszkiewicz[1], Miroslaw R. Dudek[2] and Stanislaw Cebrat[1]*
[1] *Institute of Microbiology, Wroclaw University, ul. Przybyszewskiego 63/77, 51-148 Wroclaw, Poland*
[2] *Institute of Physics, University of Zielona Góra, ul. Wojska Polskiego 69, 65-246 Zielona Góra, Poland*

Correspondence to: S. Cebrat, Institute of Microbiology, Wroclaw University, ul. Przybyszewskiego 63/77, 51-148 Wroclaw, Poland. E-mail: cebrat@microb.uni.wroc.pl; http://smORFland.microb.uni.wroc.pl

Received: 22 November 2001
Accepted: 4 February 2001

Abstract

We have compared the results of estimations of the total number of protein-coding genes in the *Saccharomyces cerevisiae* genome, which have been obtained by many laboratories since the yeast genome sequence was published in 1996. We propose that there are 5300–5400 genes in the genome. This makes the first estimation of the number of intronless ORFs longer than 100 codons, based on the features of the set of genes with phenotypes known in 1997 to be correct. This estimation assumed that the set of the first 2300 genes with known phenotypes was representative for the whole set of protein-coding genes in the genome. The same method used in this paper for the approximation of the total number of protein-coding sequences among more than 40 000 ORFs longer than 20 codons gives a result that is only slightly higher. This suggests that there are still some non-coding ORFs in the databases and a few dozen small ORFs, not yet annotated, which probably code for proteins.

Keywords: ***Saccharomyces cerevisiae*; gene number; hypothetical ORFs; questionable ORFs; coding probability; smORFs**

Introduction

There are about 7500 ORFs longer than 100 codons in the whole *S. cerevisiae* genome. The lower length limit of 100 codons was primarily accepted as an arbitrary compromise in order not to miss many real coding ORFs and to avoid a large number of spurious, non-coding ORFs (Sharp and Cowe, 1991; Oliver *et al*., 1992). The first elimination of spurious ORFs was done using the criterion that of two overlapping ORFs, only one is coding and usually it is the longer one. After this elimination, the number of ORFs annotated as presumably protein-coding sequences drops to about 6300 (Goffeau *et al*., 1996, 1997; Mewes *et al*., 1997). Seeing the overrepresentation of short ORFs (100–150 codons) in this set of ORFs, Dujon proposed elimination of ORFs shorter than 150 codons if their CAI <0.11 (Dujon *et al*., 1994, 1997; Dujon, 1996). However, this criterion has not been used in annotations of all chromosomes. At that time (1996–1997), the total number of annotated ORFs in *S. cerevisiae* databases was still about 6200–6300 and the number of ORFs with known phenotypes was about 2300. Based on the structural properties of the set of known genes, and assuming that this set was representative for the whole set of protein-coding sequences in the yeast genome, the total number of protein-coding intronless ORFs longer than 100 codons was then estimated to be about 4800 (Cebrat *et al*., 1997, 1998a; Kowalczuk *et al*., 1999; Mackiewicz *et al*., 1999). In this manuscript we have analysed the whole set of potentially coding sequences annotated in MIPS. The estimated total number of coding sequences cited in this manuscript ($\sim$5300) is consistent with our previous estimations; 4800 coding among annotated sequences in 1997, plus 86 added up to now in the database, plus 473 interrupted genes and genes shorter than 100 codons actually annotated in MIPS, giving 5359. The overrepresentation of short ORFs in the yeast databases has been pointed out in many other papers (Andrade *et al*., 1997; Das *et al*., 1997 Basrai *et al*., 1997; Mackiewicz *et al*.,

1999). Recent analyses of the yeast genome performed by Zhang and Wang (2000), and particularly the comparative analyses of the *S. cerevisiae* genome with other genomes (especially the 13 hemiascomycetous species closely related to *S. cerevisiae*) performed by the Genolevures program (see the series of papers in *FEBS Lett* **487**, 2000; e.g. Souciet *et al.*, 2000; Blandin *et al.*, 2000; Malpertuy *et al.*, 2000), have shown that many ORFs which are still annotated in *S. cerevisiae* databases, in fact do not possess orthologues (even among closely related species) and/or some of them do not possess compositional properties of known genes. This suggests that the total number of protein-coding ORFs in this genome is several hundred lower than originally assumed. Our estimations and the results of further genome analyses suggest that the choice of ORFs for the elimination from databases was correct, but the number of ORFs recently rejected is still too low, and some other ORFs should be eliminated from the databases as non-coding. These results are supported by the reannotations of the yeast genome recently done by other authors (Wood *et al.*, 2001).

Materials and methods

Databases

The sequences of the 16 *S. cerevisiae* chromosomes and ORFs annotations were downloaded from the MIPS database (**http://mips.gsf.de/proj/yeast/**) on 7 September 2001. The database listed 6422 ORFs, including 3374 genes with known phenotypes (grouped in class 1 in MIPS, referred below to known genes). Coding probabilities (scores) of particular ORFs counted by Zhang and Wang (2000) were kindly supplied by the authors. Classification and annotation of ORFs performed by the Genolevures program (Souciet *et al.*, 2000; Malpertuy *et al.*, 2000) were downloaded from **http://cbi.labri.u-bordeaux.fr/Genolevures/Genolevures.php3.** According to the Genolevures program, ORFs were grouped into the following classes: probably spurious ORFs; ORFs conserved in non-Ascomycetes; ORFs conserved only within Ascomycetes; and ORFs without homology in organisms other than *S. cerevisiae* itself.

The lists of disregarded spurious and very hypothetical ORFs identified by Wood *et al.* (2001) were downloaded from **http://www.sanger.ac.uk/Projects/S_cerevisiae**

Many authors have assigned some ORFs as spurious, hypothetical or questionable using different criteria, based on homology, length, CAI value, compositional properties or overlapping with other ORFs. We have collected a set of ORFs suggested as non-coding which have met at least one of the following criteria:

(i) Assigned as probably spurious by the Genolevures program.
(ii) Assigned as spurious or very hypothetical by Wood *et al.* (2001).
(iii) With ascribed coding probability lower than 0.5 according to Zhang and Wang (2000).
(iv) Classified as questionable ORFs in the MIPS database.

All the data have been matched with the ORFs database downloaded from MIPS, and the resulting tables are available at our website (**http://smORFland.microb.uni.wroc.pl/**). Moreover, in this database we have included coding probability scores and other compositional parameters counted for each ORF by our methods, as well as a detailed description of the methods. Parameters for a sample of ORFs are shown in Table 1.

We have also analysed the set of 41 005 ORFs, including the ORFs annotated in MIPS (6422) and 34 583 ORFs equal or longer than 20 codons and not overlapping with ORFs annotated MIPS classes 1–3 (sequences with known functions or similarities to known proteins).

Analysis of the DNA sequence asymmetry

Our method of discrimination between the coding and non-coding sequences is based on parameterization of the asymmetry between sense and antisense strands of ORFs. This method uses highly degenerated parameters of DNA composition, which leave out homology and arbitrarily chosen criteria (length, overlapping, etc.). Therefore, it is possible to find sequences with high coding probability which have no homology to any known protein. Four parameters described below characterize the compositional properties of each sequence. The set of ORFs with already ascribed functions is then compared with the set of all ORFs and, using a statistical method, the approximated number of coding sequences in the set of all ORFs is calculated. This number has been used for

Table 1. Sample ORFs from the yeast genome sorted according to their coding probability

Chr.	ORF's name	Start	Stop	Length	*S1*	*S2*	*V1*	*V2*	*D*	Cod. prob.	YZ score	MIPS	Genol.	T	Wood's annotation	Brief ID
11	YKL162c-a	146076	145927	50	−45.0	135.0	0.6	0.2	6.0	**0.00**	0.41	6	2	1	Very hypoth.	Questionable ORF—identified by SAGE
6	YFR036w-a	227437	228084	216	−177.0	−130.8	2.6	4.0	5.6	**0.04**	0.28	6	0	0		Questionable ORF
12	YLR444c	1023983	1023684	100	153.4	−153.4	1.3	2.7	5.0	**0.14**	0.39	6	0	0	Spurious	Questionable ORF
11	YKL030w	382136	382738	201	143.1	−165.1	2.5	1.1	4.8	**0.19**	0.33	6	0	0	Spurious	Questionable ORF
8	YHL042w	15665	16114	150	121.0	−153.4	1.0	1.1	4.3	**0.29**	0.47	4	3	0		Similarity to subtelom. encoded proteins
13	YML011c	247428	246898	177	44.0	140.2	3.1	0.6	4.2	**0.32**	0.58	5	2	0		Hypothetical protein
9	YIL152w	56545	57249	235	−52.9	−31.2	3.7	2.9	3.9	**0.39**	0.47	5	3	0		Hypothetical protein
1	YAL045c	57798	57493	102	18.4	−168.7	0.6	1.0	3.8	**0.42**	0.54	5	0	0	Spurious	Hypothetical protein
1	YAL008w	136912	137505	198	62.1	162.8	2.7	2.2	3.6	**0.46**	0.48	5	2	1		Hypothetical protein
14	YNL296w	76271	76582	104	68.2	−161.6	1.1	0.9	3.4	**0.50**	0.47	6	0	0	Spurious	Questionable ORF
4	YDR455c	1367670	1367365	102	99.5	−118.3	1.2	1.5	3.2	**0.55**	0.45	6	0	0	Spurious	Questionable ORF
15	YOL107w	112101	113126	342	29.1	−147.8	1.7	3.5	2.8	**0.66**	0.47	3	1	4		Weak similarity to human PL6 protein
2	YBR241c	704013	702550	488	80.8	−140.1	2.8	2.9	2.6	**0.71**	0.55	3	1	7		Similarity to glucose transport proteins
13	YML052w	170402	171307	302	73.2	−126.4	2.6	2.7	2.2	**0.79**	0.59	4	2	2		Similarity to YDL222c and YNL194c
7	YGL209w	95860	97005	382	−1.9	−56.3	3.1	3.7	2.0	**0.83**	0.59	1	1	4		C2H2 zinc-finger protein
14	YNL004w	623329	624615	429	58.4	29.4	6.4	4.4	1.9	**0.86**	0.65	2	1	4		Strong similarity to Gbp2p
15	YOL128c	79478	78354	375	28.2	−26.6	2.4	0.8	1.8	**0.87**	0.53	2	1	1		Strong similarity to kinase Mck1p
16	YPL143w	282121	282966	107	57.0	−63.4	2.3	0.6	1.8	**0.88**	0.60	1	1	5		Ribosomal protein L35a.e.c16
2	YBR053c	340710	339637	358	65.9	−3.0	3.4	1.0	1.8	**0.88**	0.60	3	2	1		Similarity to rat regucalcin
13	YMR174c	610364	610161	68	35.5	0.0	3.1	3.6	1.4	**0.93**	0.78	1	3	0		Protease A (ysca) inhibitor IA3
2	YBL003c	235754	235359	132	59.6	−61.7	2.9	1.3	1.3	**0.94**	0.69	1	1	7		Histone H2A.2
9	YIL129c	113237	106110	2376	50.9	−87.7	6.5	5.1	0.9	**0.98**	0.56	1	1	6		Transcriptional activator of OCH1
16	YPL023c	506308	504338	657	66.2	−54.3	4.3	2.2	0.9	**0.98**	0.61	1	1	3		Methylenetetrahydrofolate reductase
4	YDL112w	258914	263221	1436	56.6	−74.1	5.0	3.0	0.7	**0.99**	0.56	1	1	7		tRNA methyltransferase

Chr., chromosome number; length, length in codons; parameters of DNA asymmetry: *S1*, *S2*, values of angles (in degrees) and *V1*, *V2*, normalized lengths of vectors, for the first and the second codon positions, respectively; *D*, the Euclidean distance of the ORF from the centre of known genes' distribution in the four–dimensional space of parameters *S1*, *S2*, *V1*, *V2*; Cod. prob., coding probability; YZ score, coding probability according to Zhang and Wang (2000); MIPS, class in MIPS database: 1, known protein; 2, strong similarity to known protein (higher than one–third of FASTA self–score); 3, similarity to known protein (lower than one–third of FASTA self–score); 4, similar to unknown protein; 5, no similarity; 6, questionable ORF; Genol., class in Genolevures program (Souciet *et al.*, 2000; Malpertuy *et al.*, 2000): 0, probably spurious ORF; 1, ORF conserved in non–Ascomycetes; 2, ORF conserved in Ascomycetes only; 3, ORF without homology in organisms other than *S. cerevisiae* itself; T, total number of yeast species in which at least one homologue of the *S. cerevisiae* ORF has been identified by Genolevures program; Wood's annotation, reannotations of ORFs done by Wood *et al.* (2001). The complete data for all ORFs annotated in MIPS are available at **http://smORFland.microb.uni.wroc.pl**

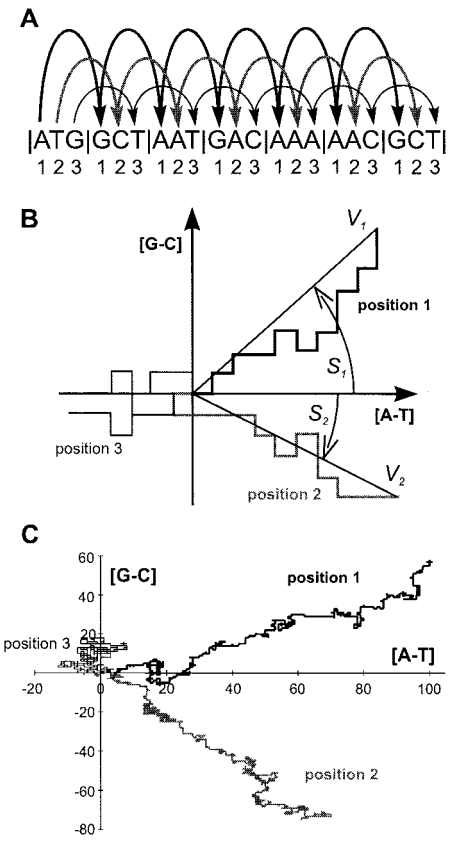


Figure 1. The method of DNA walks on coding sequence. (A) Three DNA walks performed separately for each position in codons beginning from start codon of ORF. (B) Two-dimensional representation of the above three DNA walks. Two pairs of parameters describing asymmetry of the first and the second positions in codons have been shown: *S*, angle between a vector representing the walk and *x* axis; *V*, length of this vector. (C) Two-dimensional representation of three DNA walks performed for the gene SNF12 (YNR023w) coding component of the SWI–SNF global transcription activator complex

calculating the coding probability for each ORF in the genome.

Parameters of asymmetry

To visualize the asymmetry of DNA sequence and to show the biological meaning of the parameters used, DNA walks are performed (Cebrat *et al.* 1998a; Cebrat and Dudek, 1998). Imagine that a virtual walker starts its walk on the ORF sequence from the first nucleotide of the start codon. Next, it jumps to the first nucleotide in the second codon, and so on (Figure 1A). It stops at the first nucleotide of the last codon of the analysed ORF. These walks (or jumps) are translated into a plot in a two-dimensional space where the walker goes one unit up if the visited nucleotide is guanine, down if the visited nucleotide is cytosine, right if adenine and left if thymine (Figure 1B). Then, the walker does its walk for the second and the third codon positions. Since there are very strong and specific compositional trends in each position in coding sequences, the plots 'drawn' by the walker are also specific for each position in the codons and do not resemble Brownian motion (Figure 1C).

In this study, we have used two pairs of parameters describing the walks obtained, which in fact are measures of the asymmetry in the composition of the first and the second positions in the codons of ORF sequences:

$$S = \arctan([G-C]/[A-T]) \quad (1)$$

$$V = \sqrt{[A-T]^2 + [G-C]^2}/\sqrt{N} \quad (2)$$

where A, T, G and C are the numbers of respective nucleotides in the first or second positions in codons, and N is the length of the analysed ORF in codons.

The parameter S (Figure 1B) represents the angle between the x axis and the vector, determined by the beginning and the end of the corresponding walk (equation 1). It represents the relation between the relative abundance of purines over pyrimidines in complementary pairs of nucleotides. To avoid infinite values of slopes, we have used a measure of angle rather than tangent. Furthermore, the function of arctangent in many instances 'normalizes' distributions. The parameter V (Figure 1B) describing the asymmetry was the length of the vector described by the beginning and the end of the walk (equation 2). The length of vector representing a sequence was divided by the square root of the length of this sequence in codons. Such normalization shows the relation between the length of this vector and the average vector for random DNA sequence of the same length.

Estimation of number of protein-coding sequences

Having each ORF described by four independent parameters (*S1*, *S2*, the values of angles; and *V1*, *V2*, the normalized lengths of vectors, for the

first and the second codon positions, respectively), we prepared distribution of all ORFs in four-dimensional space. To compare the distribution of coding sequences to the distribution of all ORFs, we counted for genes with known functions the average values, M, and the standard deviations, SD, for each of the four parameters, P. The difference between M and P of the given parameter for each ORF was divided by SD (equation 3). The obtained value, N, was a coordinate of the ORF in the four-dimensional space and was used for counting the Euclidean distance, D, of the ORF from the centre of known genes distribution:

$$N_i = |P_i - M_i| / SD_i \quad (3)$$

where P_i is a value of parameter i for a given ORF [parameters $S1$, $S2$, are values of angles (in degrees) and $V1$, $V2$, are normalized lengths of vectors for the first and the second codon positions, respectively], M_i is the average value of parameter i for genes with known functions, SD_i is SD of parameter i for genes with known functions, and N_i is the normalized distance of an ORF to the centre of known genes distribution of parameter i:

$$D = \sqrt{\sum_{i=1}^{4} N_i^2} \quad (4)$$

After the data were grouped into classes, the modal value for distribution of D for genes with known functions was estimated. Then the distribution functions of D for sets of all ORFs and of known genes were prepared. Assuming that all ORFs with D lower than the mode for known genes were coding, the distribution function for genes with known functions was respectively rescaled by multiplying them by ratio ORF_m/G_m, where ORF_m is the number of ORFs in the distance lower than the mode for known genes, and G_m is the number of known genes in the same distance. From the rescaled diagram, the approximated total number of coding ORFs was counted for 5322 (Figure 2).

This method of approximation by definition overestimates the total number of coding ORFs, because there is an assumption that all ORFs closer to the centre of distribution than the modal value for known genes are coding. The distance of the point describing a given ORF to the centre of known genes distribution, and the approximate number of coding ORFs, enable calculation of the coding probability for each ORF (Table 1), which is a fraction of all ORFs for which one could expect to find phenotypes. For complete data for all ORFs, see **http://smORFland.microb.uni.wroc.pl/.**

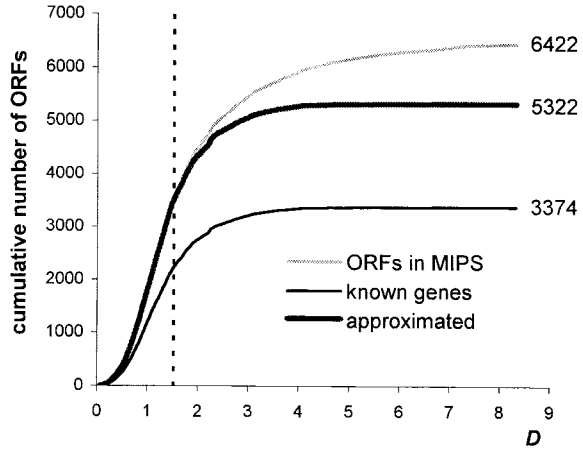
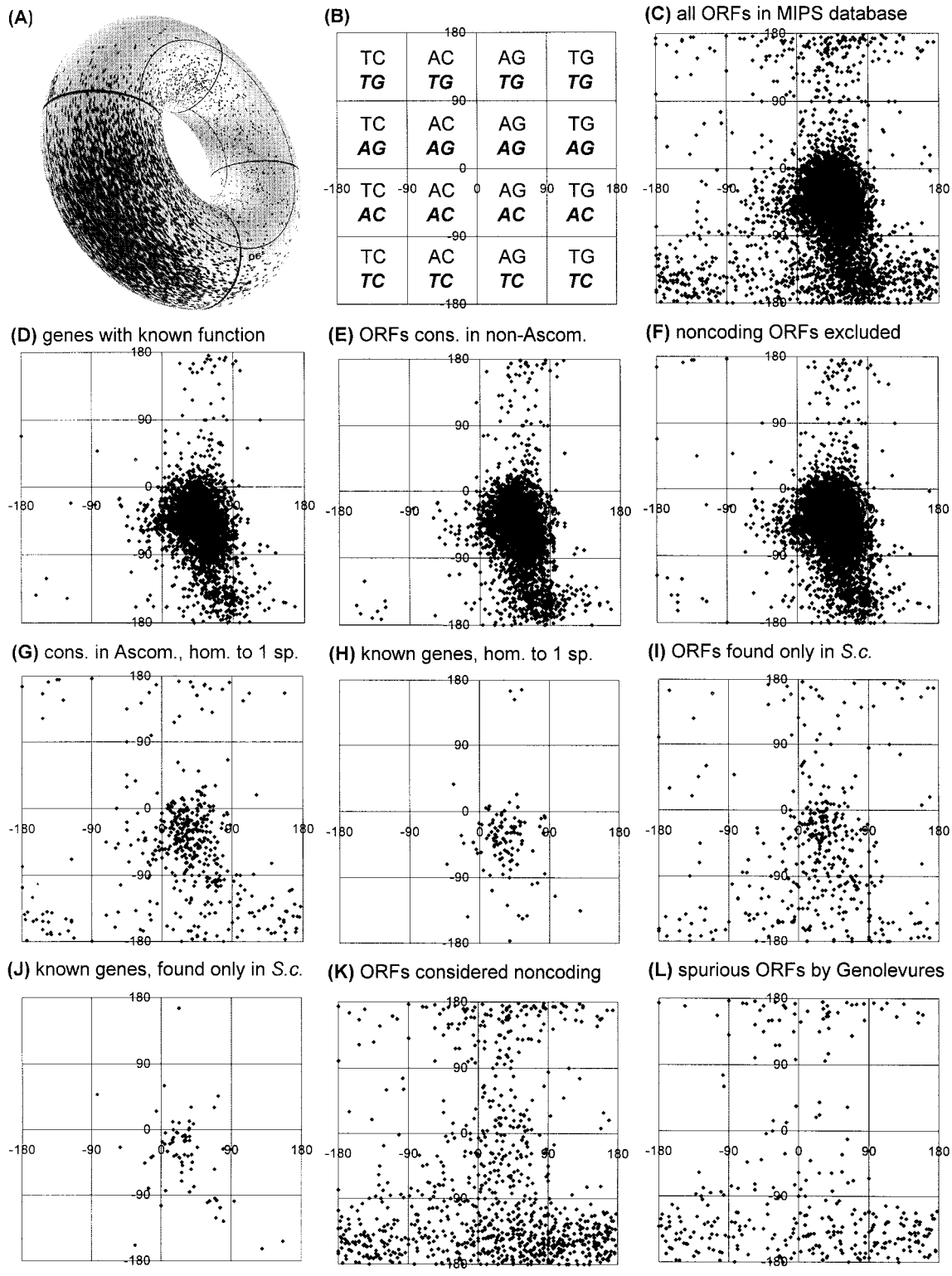


Figure 2. Distribution functions of D, the Euclidean distance from the distribution centre of genes with known phenotypes, for a set of all ORFs annotated in MIPS and for a set of genes with known phenotypes. Approximated number of all coding ORFs was received by scaling-up the distribution function for known genes assuming that all ORFs with D lower than the mode for known genes (vertical dashed line) were coding. See text for a more detailed description

Results and discussion

Analyses of ORFs distribution on the torus surface

Parameters $S1$ and $S2$, which are the angles of vectors described by DNA walks for the first and the second positions in codons, are the measure of DNA asymmetry for each ORF. They may be used for producing the distribution of ORFs in two-dimensional space that is an unfolded surface of a torus (Cebrat *et al.*, 1997). In such representation, each sequence is described by a point on the finite surface of the torus (Figure 3A). In Figure 3 we have shown the distributions of different sets of ORFs. The location of ORF in this plot is related to base composition of the first and the second codon position (Figure 3B), e.g. ORFs in the sector between coordinates (0°, −90°), (90°, −90°), (0°, 0°) and (90°, 0°) are relatively rich in guanine and



adenine in the first codon positions and rich in adenine and cytosine in the second codon positions. It can be seen that ORFs with known functions form a very compact set in this space (Figure 3D). All ORFs annotated in the MIPS database form a much more dispersed set of points (Figure 3C). In other plots we have shown the distributions of different sets of ORFs, annotated in MIPS but classified according to different criteria. ORFs conserved in non-Ascomycetes form a set resembling the distribution of genes with known functions (Figure 3E). Conversely, ORFs without homology in organisms other than *S. cerevisiae* itself form much dispersed sets, not resembling the distribution of coding sequences (Figure 3I). Even the distribution of ORFs conserved in Ascomycetes that have homology to only one of the 13 hemiascomycetous species closely related to *S. cerevisiae* is quite different from the distribution of coding sequences. This suggests that in this set many of ORFs are not coding (Figure 3G). On the other hand, genes with assigned functions classified in these two sets are distributed similarly to other known genes (Figure 3J and 3H, respectively). In Figure 3K we have collected ORFs which, according to at least one author, are considered non-coding (see Materials and methods). According to our criteria, coding probabilities of these ORFs are very low and they form a set quite different from the set of genes with known functions. If we eliminate these ORFs from the database, the number of remaining ORFs will be 5492 (see the distribution in Figure 3F). This number seems to be still overestimated because some ORFs, especially in the set of ORFs without homology in organisms other than *S. cerevisiae* itself, may be considered non-coding. If we subtract the number of these ORFs (132) from 5492, we receive the estimated number of coding ORFs, 5360.

Evaluation of approximations of the total number of coding ORFs

The first sets of ORFs annotated in databases were over-represented by definition. Authors assumed that it was better to include some false positives than eliminate false negatives from databases. Dujon proposed the elimination of ORFs shorter than 150 codons if their CAI <0.11 (Dujon *et al.*, 1994, 1997; Dujon, 1996), which resulted in a gene number of 5885 (Goffeau *et al.*, 1996).

Approximations done by Zhang and Wang (2000) suggested that there were no more than 5645 protein-coding ORFs. These authors constructed a ranking of ORFs according to their coding probabilities on the basis of some compositional properties of known genes and ORFs generated in intergenic sequences. Nevertheless, two phenomena could influence their results: first, intergenic sequences in the yeast genome are not random and do not represent a sequence that generates ORFs randomly with a probability indicated simply by its nucleotide composition; and second, many non-coding, questionable ORFs have a very strong triplet structure and even asymmetry resembling coding sequences, because they have been generated inside coding sequences, usually in their antisense (Cebrat *et al.*, 1998b; Mackiewicz *et al.*, 1999). We have found that, according to parameters used by

Figure 3. Distribution of different sets of ORFs from the *S. cerevisiae* genome on the torus (A) and on the torus unfolded surface (C–L). The location in this plot is related to base composition of the first and the second codon position. Compositional properties of each sector are shown schematically in (B). The two upper standard letters in a sector indicate the relative excess of given nucleotides in the first codon position, and the two bottom letters in bold italics indicate the relative excess in the second codon position, e.g. ORFs in the quadrant sector between coordinates (0°, −90°), (90°, −90°), (0°, 0°) and (90°, 0°) are relatively rich in guanine and adenine in the first codon positions and rich in adenine and cytosine in the second codon positions. (A) All ORFs longer than 100 codons. (C) All ORFs annotated in MIPS database. (D) Genes with known functions. (E) ORFs conserved in non-Ascomycetes. (F) ORFs remaining after excluding all ORFs non-coding according to at least one author. (G) ORFs without known functions conserved in Ascomycetes which have homology only to one of the 13 closely related to *S. cerevisiae* hemiascomycetous species. (H) Genes with known functions in the set shown in (G). (I) ORFs without known functions and without homology in organisms other than *S. cerevisiae* itself. (J) Genes with known functions in the set shown in (I). (K) ORFs that, according to at least one author, are considered non-coding. (L) ORFs assigned as spurious by the Genolevures program. Each ORF is represented by a point with coordinates: x, $S_1 = \arctan([G_1 - C_1]/[A_1 - T_1])$, counted for the first positions in codons; y, $S_2 = \arctan([G_2 - C_2]/[A_2 - T_2])$, counted for the second positions in codons

Zhang and Wang (2000), many overlapping ORFs entirely included in longer ones are considered coding (data not shown). That is why their approximations could be overestimated. In addition, the gene number estimation done by Wood *et al.* (2001), i.e. 5570, seems to be overestimated, as the authors themselves stated.

A lot of information on probability of coding by questionable ORFs in the yeast genome has been supplied by the Genolevures program, coordinated by Dujon and published in a series of papers in *FEBS Letters* **487** (e.g. Souciet *et al.*, 2000; Blandin *et al.*, 2000; Malpertuy *et al.*, 2000). The predicted gene numbers were estimated on 5651 and 5547 (mean from two estimation made by two statistical methods). Based on studies of homology with the *S. cerevisiae* genome, other Ascomycetes (especially the 13 hemiascomycetous species closely related to *S. cerevisiae*) and non-Ascomycetes, the authors have eliminated several hundred spurious ORFs from the yeast databases. Figure 3L shows how the distribution of the set of these eliminated ORFs is different from the set of genes with known functions (Figure 3D), suggesting that the elimination was correct. Some ORFs having one or even two hypothetical homologues in other Ascomycetes could also be non-coding. Furthermore, Wood *et al.* (2001) have found that some *S. cerevisiae* ORFs with homologues in hemiascomycetes are in fact non-coding but are generated in the alternative frame of real genes, which is a very common mechanism of generating false reading frames in genomes (Mackiewicz *et al.*, 1999). Using asymmetry parameters, we have approximated the total number of protein-coding sequences among sequences annotated in MIPS (including genes with introns and ORFs shorter than 100 codons) as 5322 (Figure 2). This predicted gene number in the yeast genome is in agreement with the number of about 5360 that we have received after eliminating the ORFs that are non-coding according to other authors' criteria. Nevertheless, our approximations are higher than the lower limit of estimations obtained by Malpertuy *et al.* (2000), i.e. 5175, since they have assumed the most restrictive conditions for qualifying ORFs as coding on the basis of homology study.

The approximations performed with the whole set of 41 005 ORFs longer than 20 codons have given a number a few dozen higher, suggesting that there are still some coding ORFs among ORFs shorter than 100 codons.

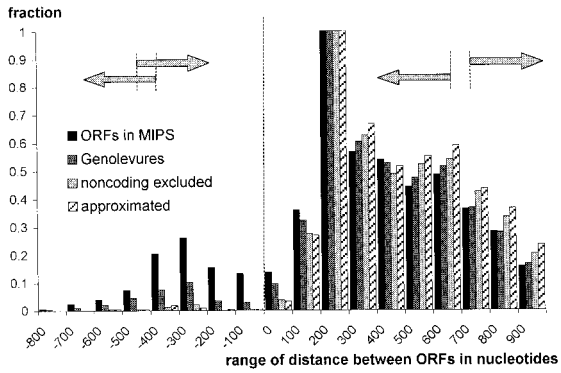


Figure 4. The distribution of the distances start to start between divergent pairs of different sets of ORFs in the yeast genome: all ORFs annotated in MIPS database (ORFS in MIPS), ORFs remaining after elimination spurious ORFs by Genolevures program (Genolevures), ORFs remaining after excluding all ORFs non-coding according to at least one author (non-coding excluded), ORFs assigned as coding according to our criteria (approximated). To compare sets, fraction of ORFs in the given range of distance in the analysed set was normalized by the modal value. Dashed line indicates border between negative and positive values of distance. Negative values of distance indicate overlapping regions of ORFs

Analysis of genome structure

There are some other premises which suggest that the number of spurious ORFs eliminated by the Genolevures program should be higher. We have studied the distribution of distances between the divergent ORFs in the yeast genome. It is widely accepted that the number of overlapping divergent coding ORFs should be negligible, if any. Such overlapping genes with known functions have not been detected in the yeast genome so far (Wood *et al.*, 2001). In Figure 4 we have shown the distribution of the distances, start-to-start, between divergent pairs of different sets of ORFs. To compare sets, the diagrams show the fraction of ORFs in the given range of distance in the analysed set, normalized by the modal value. There are still many overlapping ORFs in the MIPS database, while there are single overlapping frames in the set of ORFs coding according to our criteria and in the set of ORFs after rejection of sequences suggested to be non-coding by the different criteria of other authors. Elimination of ORFs by the Genolevures program led by Dujon improved this situation considerably, but after that elimination, overrepresentation of ORFs with short distances between

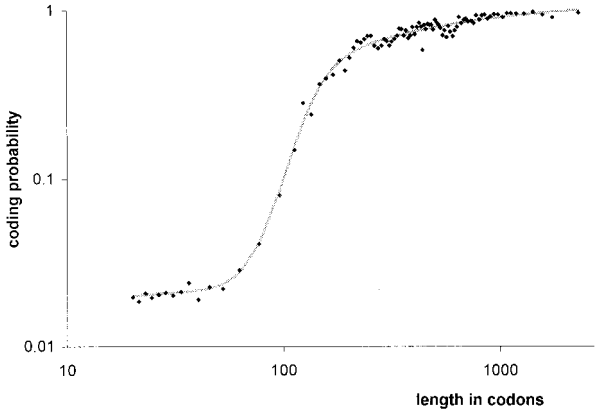


Figure 5. The distribution of coding probabilities of ORFs in relations to their length. Note that both axes are in logarithmic scale. The plot was obtained by analysis of the set of 41 005 ORFs, including the ORFs annotated in MIPS (6422) and 34 583 ORFs equal or longer than 20 codons and not overlapping with ORFs annotated MIPS classes 1–3 (sequences with known functions or similarities to known proteins)

them is still observed. In some instances, it could be possible that start codons in one or in both ORFs should be shifted down the coding sequence. The same conclusions could be drawn after analysis of distances between convergently and tandemly orientated ORFs (data not shown).

Length distribution of the coding ORFs

In our method, to avoid the influence of an ORF's length on its coding probability estimation, we have divided the length of the vector representing an ORF by the square root of the length of the ORF (see Materials and methods). Thus, the value of the parameter represents the measure of a statistically significant trend in the sequence asymmetry. That enables us to estimate the coding probability even for very short ORFs. Furthermore, it also makes the analysis of the length distribution of ORFs reasonable. In Figure 5 we have shown the distribution of coding probabilities of ORFs longer than 20 codons in relation to their length. This probability is very low but higher than zero even for the class of ORFs 20–30 codons long, and it is still low for ORFs just over 100 codons. In Figure 6 we have compared the length distributions of different sets of ORFs. Four sets have almost identical distributions (Figure 6A): 1, ORFs with already known phenotypes, annotated class 1 in MIPS; 2, ORFs conserved in non-Ascomycetes; 3, ORFs remaining after excluding all ORFs that are non-coding according to at least one author; 4, 5322 ORFs with the highest coding probability according to us. The other three sets of ORFs (Figure 6B) are: 5, suggested to be non-coding at least by one author; 6, without homology in organisms other than *S. cerevisiae* itself; 7, ORFs conserved in Ascomycetes which have homology to only one of the 13 hemiascomycetous species closely related to *S. cerevisiae* have overrepresented groups of shorter ORFs, especially of the class just above 100 codons long. It is characteristic that the set of ORFs conserved in Ascomycetes only is overrepresented also by short ORFs, suggesting that there are still some non-coding ORFs in this set, indicating again that it is possible to find homologues even for non-coding ORFs in related species. On the other hand, the length distribution of known genes belonging to sets 6 and 7 resemble the distribution of other known genes (data not shown).

There is a possible mechanism explaining why non-coding ORFs, even those that have homologues in other species, could exist in genomes. It could involve intragenomic recombinations which generate fragments of coding sequences with ORFs in antisense (Mackiewicz *et al.*, 1999). One can argue that ORFs specific for a narrow group of species could be so specific that they escape our criteria of coding estimation, but why are these ORFs shorter? Some mathematical approaches have indicated that, if the prevailing number of questionable ORFs code for proteins, then proteins coded by them would form new protein families—almost each one for itself (Fisher and Eisenberg, 1999). It was hard to accept the hypothesis that the number of about 1200 protein families known should be doubled by a simple assumption that all ORFs annotated in *S. cerevisiae* databases are coding. Moreover, we would have to allow overlapping of genes with their 5′ regions.

Our explanation of the existence of these ORFs takes into account that coding sequences and the genetic code itself have specific properties of generating long ORFs in alternative frames, particularly in the antisense strand (Cebrat and Dudek, 1996; Cebrat *et al.*, 1998b). Many of these ORFs may have arisen by duplications of coding sequences nesting overlapping non-coding ORFs. Duplicated sequences accumulated mutations which eventually generated stops in the 'maternal' reading frames of the original genes, leaving the generated non-coding

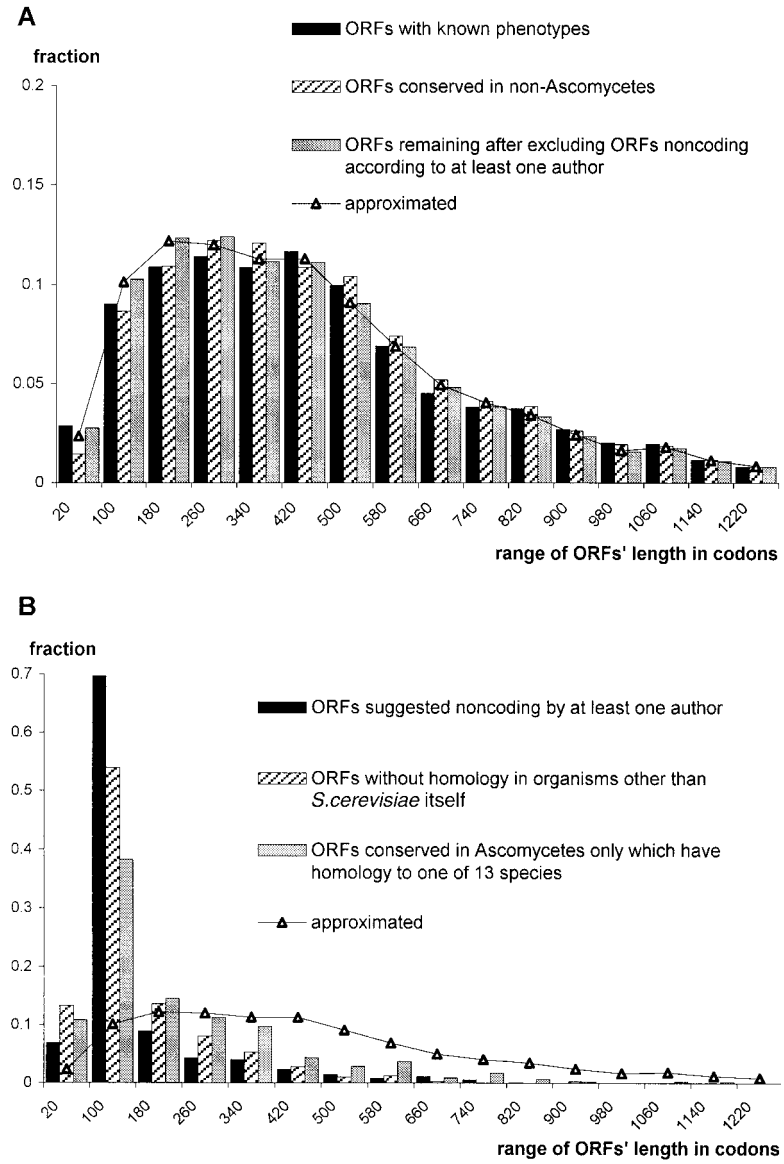


Figure 6. The length distributions of different sets of ORFs. (A) ORFs with known phenotypes; ORFs conserved in non-Ascomycetes; ORFs remaining after excluding all ORFs non-coding according to at least one author; 5322 ORFs with the highest coding probability according to us (approximated). (B) ORFs suggested as non-coding at least by one author; ORFs without homology in organisms other than *S. cerevisiae* itself; ORFs conserved in Ascomycetes which have homology to only one of the 13 closely related to *S. cerevisiae* hemiascomycetous species; 5322 ORFs with the highest coding probability according to us (approximated), shown here in the same scale for comparison

ORFs. That is why these ORFs have compositional properties of non-coding overlapping frames entirely nested inside known genes. We have found about 700 ORFs in the MIPS database whose putative protein products have homologues in frames different from the frame that had been assumed in the database as coding (Mackiewicz *et al.*, 1999). We have identified ORFs with compositional properties of non-coding overlapping frames also in other genomes (Mackiewicz *et al.*, 2002). There is a lot of evidence that the yeast genome underwent many duplications in the past (e.g. Wolfe and Shields, 1997), which is why this scenario seems to be plausible.

Although some antisense ORFs exhibit triplet structure and sequence asymmetry characteristic for coding sequences, they possess completely different codon usage compared to real genes. The probability that they really code for proteins is very low. In fact we have noticed the property of the genetic code of generating antisense ORFs (Cebrat and Dudek, 1996) and we have found only three genes out of more than 200 sequences in the yeast genome which could arise in this way (Cebrat *et al*., 1998b). The problem of (non)coding in non-coding frames and coding of overlapping ORFs was recently widely discussed by Boldogkoi (2000).

Combined data of coding probabilities counted by us for each ORF and some data published by other authors are available on our website (**http://smORFland.microb.uni.wroc.pl/).**

Acknowledgements

This work was supported by the State Committee for Scientific Research, Grant Nos 6 P04A 025-18 and 6 P04A 016-20. P.M. and M.K. were supported by the Foundation for Polish Science.

References

Andrade MA, Daruvar A, Casari G, Schneider R, Termier M, Sander C. 1997. Characterization of new proteins found by analysis of short open reading frames from the full yeast genome. *Yeast* **13**: 1363–1374.

Basrai MA, Hieter P, Boeke JD. 1997. Small open reading frames: beautiful needles in the haystack. *Genome Res* **7**: 768–771.

Blandin G, Durrens P, Tekaia F, *et al*. 2000. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Lett* **487**: 31–36.

Boldogkoi Z. 2000. Coding in the non-coding DNA strand: a novel mechanism of gene evolution? *J Mol Evol* **51**: 600–606.

Cebrat S, Dudek MR, Mackiewicz P, Kowalczuk M, Fita M. 1997. Asymmetry of coding versus non-coding strand in coding sequences of different genomes. *Microb Comp Genom* **2**: 259–268.

Cebrat S, Dudek MR, Mackiewicz P. 1998a. Sequence asymmetry as a parameter indicating coding sequence in *Saccharomyces cerevisiae* genome. *Theory BioSci* **117**: 78–89.

Cebrat S, Dudek MR. 1996. Generation of overlapping open reading frames. *Trends Genet* **12**: 12.

Cebrat S, Dudek MR. 1998. The effect of DNA phase structure on DNA walks. *Eur Phys J* **3**: 271–276.

Cebrat S, Mackiewicz P, Dudek MR. 1998b. The role of the genetic code in generating new coding sequences inside existing genes. *Biosystems* **42**: 165–176.

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

Dietrich F, Voegeli S, Brachat S, *et al*. 2001. Evolution of the *S. cerevisiae* genome: lessons learned from the genome analysis of the fungus *Ashbya gossypii*. 2001. XXth International Conference on Yeast Genetics and Molecular Biology, Prague, August 26–31 (late abstracts).

Dujon B, Albermann K, Aldea M, *et al*. 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XV. *Nature* **387**: 98–102.

Dujon B, Alexandraki D, Andre B, *et al*. 1994. Complete DNA sequence of yeast chromosome XI. *Nature* **369**: 371–378.

Dujon B. 1996. The yeast genome project: what did we learn? *Trends Genet* **12**: 263–270.

Fischer D, Eisenberg D. 1999. Finding families for genomic ORFans. *Bioinformatics* **15**: 759–762.

Goffeau A, Aert R, Agostini-Carbone ML, *et al*. 1997. The yeast genome directory. *Nature* **387**: 5–105.

Goffeau A, Barrel BG, Bussey H, *et al*. 1996. Life with 6000 genes. *Science* **274**: 546–567.

Kowalczuk M, Mackiewicz P, Gierlik A, Dudek MR, Cebrat S. 1999. Total number of coding open reading frames in the yeast genome. *Yeast* **15**: 1031–1034.

Mackiewicz P, Kowalczuk M, Gierlik A, Dudek MR, Cebrat S. 1999. Origin and properties of non-coding ORFs in the yeast genome. *Nucleic Acids Res* **27**: 3503–3509.

Mackiewicz P, Kowalczuk M, Gierlik A, *et al*. 2002. No mystery of ORFans in genomics—generation of ORFans in the antisense of coding sequences. *Biophysics* (in press).

Malpertuy A, Tekaia F, Casaregola S, *et al*. 2000. Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. *FEBS Lett* **487**: 113–121.

Mewes H-W, Albermann K, Bahr M, *et al*. 1997. Overview of the yeast genome. *Nature* **387**: 7–8.

Oliver SG, van der Aart QJM, Agostoni-Carbone ML, *et al*. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357**: 38–46.

Sharp PM, Cowe E. 1991. Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* **7**: 657–678.

Souciet J-L, Aigle M, Artiguenave F, *et al*. 2000. Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett* **487**: 3–12.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708–713.

Wood V, Rutherford KM, Ivens A, Rajandream M-A, Barrell B. 2001. A re-annotation of the *Saccharomyces cerevisiae* genome. *Comp Funct Genom* **2**: 143–154.

Zhang C-T, Wang J. 2000. Recognition of protein-coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res* **28**: 2804–2814.

