YEAST
Yeast **15**, 1031–1034 (1999)

Total Number of Coding Open Reading Frames in the Yeast Genome

MARIA KOWALCZUK[1], PAWEL MACKIEWICZ[1], AGNIESZKA GIERLIK[1], MIROSLAW R. DUDEK[2] AND STANISLAW CEBRAT[1]*

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

At the end of 1996 we approximated the total number of protein coding ORFs in the *Saccharomyces cerevisiae* genome, based on their properties, as 4700–4800. The number is much smaller than the 5800 which is widely accepted. According to our calculations, there remain about 200–300 orphans—ORFs without known function or homology to already discovered genes, which is only about 5% of the total number of genes. Our results would be questionable if the analysed set of known genes was not a statistically representative sample of the whole set of protein coding genes in the *S. cerevisiae* genome. Therefore, we repeated our estimation using recently updated databases. In the course of the last 18 months, previously unknown functions of about 500 genes have been found. We have used these to check our method, former results and conclusions. Our previous estimation of the total number of coding ORFs was confirmed. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS — coding sequence; genome coding capacity; open reading frame; gene number

INTRODUCTION

It is easy to find in the *S. cerevisiae* genome all open reading frames (ORFs) defined as sequences beginning with a start codon and ending with a stop codon, independently of their length or overlapping. However, it is not trivial to determine the number of ORFs actually coding for proteins. There are nearly 7500 ORFs longer than 100 codons, but only about 2700 have an ascribed function. Approximately 1800 ORFs without known functions have homologues in the yeast or other genomes. Most of the remaining 3000 overlap. It is assumed that only one ORF of an overlapping pair may code for a protein and usually it is the longer one. Thus, the shorter one is considered non-coding. Coding capacity is also measured by the codon bias index (CBI) (Benetzen and Hall, 1982) and the codon adaptation index (CAI) (Sharp and Li, 1987). It is accepted that ORFs shorter than 150 codons with CAI<0.11 are non-coding (Dujon *et al.*, 1994). The rest are called 'orphans' because no functions or homologues have been found for them. A quarter of the yeast genes identified by traditional methods were orphans (Oliver *et al.*, 1992). The sequencing of the *S. cerevisiae* genome revealed a surprisingly high number of such sequences. The proportion of orphans grew quicker than the proportion of homologues, which is a paradox because the more genes known, the higher the proportion of homologues that should be found for them among newly sequenced ORFs. This phenomenon has been called the 'mystery of orphans' (Casari *et al.*, 1996, Dujon 1996).

Our previous paper (Cebrat *et al.*, 1997) provided an explanation for this. Approximation of the number of protein coding ORFs, based on their properties, yielded 4700–4800, a number much smaller than the widely accepted 5800–6000 (Goffeau *et al.*, 1996; Winzeler and Davis, 1997; Mewes *et al.*, 1997). If it is acknowledged that there are indeed only about 4800 genes in the yeast genome, the mystery of orphans disappears. There remain about 200–300 ORFs without known function or homology to genes already discovered, which is only about 5% of the total number.

*Correspondence to: S. Cebrat, Institute of Microbiology, Wroclaw University, ul. Przybyszewskiego 63/77, 51-148 Wroclaw, Poland. Tel: 48-071-3247 303; e-mail: cebrat@angband.microb.uni.wroc.pl.

Contract/grant sponsor: KBN, Poland; Contract/grant number: 6 P04A 030 14.

CCC 0749–503X/99/111031–04 $17.50

Received 1 August 1998
Accepted 3 March 1999

Our results are questionable only if already known genes are not a statistically representative sample of the whole set of protein-coding genes in the *S. cerevisiae* genome. If the genes used in our analysis were too homogenous a group, the newly discovered ones would enlarge the approximated number of genes. In the course of the last year, previously unknown functions of about 500 genes have been found. We have used these to check our method, former results and conclusions.

DATABASES AND METHODS

The DNA sequence for the previous database was downloaded from genome.stanford.edu on 23 September 1996, and information on gene functions from http://www.mips.biochem. mpg.de on 16 November 1996. The data from the latter are referred to in the paper as 'dbase 1'.

To create the new databases, we retrieved the required information from http://speedy.mips. biochem.mpg.de in October 1997 (referred to in this paper as 'dbase 2') and from http:// speedy.mips.biochem.mpg.de in June 1998 (referred to as 'dbase 3'). The analysis described by Cebrat *et al.* (1997) was repeated with the new data and new approximations of the number of genes were calculated.

Specific asymmetry in the occurrence of purines and pyrimidines is observed in the first and second positions in codons in protein-coding sequences (Cebrat *et al.*, 1997). The asymmetry was measured by the ratio (G–C)/(A–T), calculated for the first and second positions in codons on the coding strand of ORFs. The values of arctg(G–C)/ (A–T) for the respective positions were plotted against each other for all ORFs longer than 100 codons (7472 ORFs of dbase 3). The result was a distribution of points representing individual ORFs on the torus surface. The same kind of plot was done for ORFs with known function, and the co-ordinates of the centre of that distribution were determined by the average values for the two positions in codons. For each ORF, distance A from the centre of the set of genes was calculated, using standard deviation (SD) to normalize both parameters. The distribution of all ORFs was compared with the distribution of genes.

To estimate the number (N) of coding ORFs, we calculated the ratio between the number of genes inside the space determined by a given A (G_{in}) and the number of genes outside that space (G_{out}). Next, we counted all ORFs inside the same space (ORF_{in}) and assumed that they were genes. Presuming that the ratios for genes and all ORFs are equal, we calculated the number of coding ORFs outside that space and added the two numbers, according to the algorithm:

$$N_{\mathrm{i}} = ORF_{\mathrm{in}} + [(G_{\mathrm{out}}/G_{\mathrm{in}})(ORF_{\mathrm{in}})]$$

We plotted values N_{i} against their respective A values and obtained a plot in which the approximate number of protein-coding ORFs is determined by the value on the y axis at the intersection with the extrapolated line ($A=0$). At each end of the plot, 10% of values N_{i} were ignored because the error of the method increases approaching the centre of the distribution. Each ORF located close to the centre decides about rejecting or accepting as coding a considerable number of ORFs from outside.

RESULTS AND DISCUSSION

The distribution of all ORFs in the torus projection is shown in Figure 1a. Simple comparison of this distribution with previous distributions calculated for the set of genes already-known in 1996 (Figure 1b) and for the 500 genes identified during the last 18 months (Figure 1c) shows that a lot of ORFs are located in the regions of the plot where not a single gene has been found and that new genes are located in the same regions as previously known genes. The results of estimations for the three databases are shown in Figure 2. These estimations were done using all ORFs found in the yeast genome, without any previous discrimination, including all overlapping ORFs, even smaller ones inside larger ORFs with already-known coding functions. The agreement between these results is high and the estimated number of coding ORFs does not exceed 4800. We have also performed the approximation for ORFs annotated in MIPS (dbase 3). During the preparation of these data by MIPS, a lot of presumably non-coding ORFs were discarded. If we assume that these decisions were correct, the result of the approximation should be the same as for whole sets of ORFs. In fact, the estimated number of coding ORFs was only 20 lower than for the whole dbase 3 (note that the difference between the total number of ORFs and the number of ORFs annotated in MIPS base is about 1600). Thus, preliminary elimination of

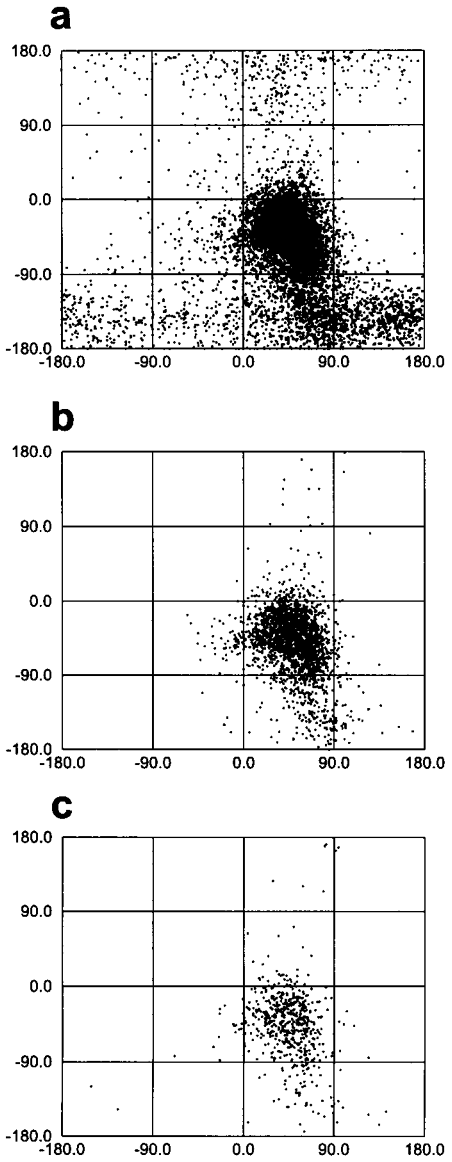

Figure 1. Distribution of the yeast genome ORFs on the torus projection. Each ORF is represented by a point with co-ordinates: x [arctg(G–C)/(A–T)] counted for the first positions in codons; y [arctg(G–C)/(A–T)] counted for the second positions in codons: (a) distribution of all 7472 ORFs longer than 100 codons found in the yeast genome dbase 3; (b) 2205 ORFs with known function from dbase 1; (c) 500 genes with functions described during the last 18 months.

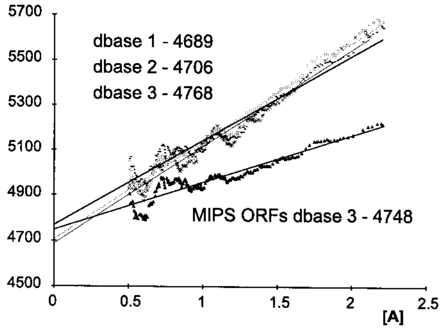

Figure 2. Approximation of the total number of protein coding ORFs in the yeast genome. x=Distance from the centre of gene distribution; y=number of ORFs inside the surface with the distance lower than x-value.

these 1600 ORFs has not significantly changed the total number of presumably coding ORFs. Another way to verify the method is to exclude presumably non-coding ORFs, which should not change the approximation. We excluded the shorter ORFs without known function of 153 convergent overlapping pairs. The approximated number of protein coding sequences is one gene lower than the result obtained from the whole database.

In the set of about 500 genes with newly found coding functions, only two genes (YIL049 and YMR274) are located in a region where no genes in dbase 1 were found. Disruption of both genes is not lethal. The four results shown in Figure 2 are convergent and they clearly indicate that there are considerably fewer coding ORFs and fewer orphans in the *S. cerevisiae* genome than formerly expected. Five hundred new genes have not changed the approximate number of genes significantly.

One could argue that our method of approximation produces an underestimation because all ORFs close to the centre of the distribution follow the rules of nucleotide composition of genes, therefore they must be genes. Even if we assume that this is true and accept the hypothesis that all ORFs at a distance from the centre of distribution lower than 1 SD are protein coding—the estimated total number of coding ORFs is below 5100. This result is higher than our previous estimations, but it is overestimated by definition because we have assumed that there are no random ORFs at all in the yeast genome with parameters characteristic for coding sequences.

We have also compared the sets of genes from the three analysed databases by non-parametric Kruskal–Wallis, Mann–Whitney U and Kolmogorov–Smirnov tests (Sokal and Rohlf, 1981). The results allow us to accept the hypothesis that differences between the three sets are not significant, and that the set of already known genes is significantly different from the set of the rest of ORFs annotated in the MIPS base. Our recent studies have shown that an overwhelming number of ORFs annotated 'class 6' in the MIPS database (i.e. ORFs without known function or homology to known coding sequences) 'translated' in a different phase than had been assumed for them as coding in MIPS databases, find homologous proteins in the yeast and/or other genomes (Gierlik *et al.*, 1999). This confirms our hypothesis that many non-coding ORFs have been generated by coding sequences (Cebrat and Dudek, 1996).

CONCLUSION

The number of protein-coding ORFs in the yeast genome is of the order of 4800. Thus, the number of orphans in that genome is of the order of 300, which is acceptable and should not be considered paradoxically high. Accepting the estimated number of coding ORFs in the yeast genome, it is possible to count the probability of coding for each ORF without identified function.

ACKNOWLEDGEMENTS

We would like to thank Professor A. Goffeau for many long discussions. This work was supported by a Grant from KBN, No. 6 P04A 030 14.

REFERENCES

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

Casari, G., de Druvar, A., Sander, C. and Schneider, R. (1996). Bioinformatics and the discovery of gene function. *Trends Genet.* **12,** 244–255.

Cebrat, S. and Dudek, M. R. (1996). Generation of overlapping reading frames. *Trends Genet.* **12,** 12.

Cebrat, S., Dudek, M. R., Mackiewicz, P., Kowalczuk, M. and Fita, M. (1997). Asymmetry of coding versus non-coding strand in coding sequences of different genomes. *Microb. Comp. Genom.* **4,** 259–268.

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

Dujon, B. *et al.* (106 co-authors). (1994). Complete DNA sequence of yeast chromosome XI. *Nature* **369,** 371–378.

Gierlik, A., Mackiewicz, P., Kowalczuk, M., Dudek, M. R. and Cebrat, S. (1999). Some hints on open reading frame statistics—how ORF length depends on selection. *Int. J. Modern Phys.* **C** (in press).

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). Life with 6000 genes. *Science* **274**(5287), 546.

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). Overview of the yeast genome. *Nature* **387,** 7–8.

Oliver, S. G. *et al.* (146 co-authors). (1992). Complete DNA sequence of yeast chromosome III. *Nature* **357,** 38–46.

Sharp, P. M. and Li, W. H. (1987). The codon adaptation index: a measure of directional synonymous codon usage bias and its potential applications. *Nucleic Acids Res.* **15,** 1281–1295.

Sokal, R. R. and Rohlf, F. J. (1981). *Biometry: The Principles and Practice of Statistics in Biological Research.* 2nd edn. W.H. Freeman, New York.

Winzeler, E. A. and Davis, R. W. (1997). Functional analysis of the yeast genome. *Curr. Opin. Genet. Dev.* **7**(6), 771–776.