J. Appl. Genet. 38(4), 1997, pp. 365-372

Is there any mystery of ORPHANs ?

Stanisław CEBRAT[1], Mirosław DUDEK[2], Paweł MACKIEWICZ[1]

[1] Institute of Microbiology, [2] Institute of Theoretical Physics, Wrocław University, Wrocław, Poland

Abstract. We have analysed the coding capacity of ORFs longer than 100 codons found in the yeast genome. Comparing the parameters describing the DNA asymmetry in the set of known genes and the set of all ORFs>100 codons we have found that there are about 4700 coding ORFs in the yeast genome. Since for more than 2300 ORFs recognisable functions have been already found and for about 2000 ORFs homology to known genes has been identified – only about 400 ORFs can be considered as orphans – ORFs without any known function or homology. This finding means that there is no mystery of orphans – a paradox showing that the fraction of orphans has been growing with the growing number of genes with known functions in the yeast genome.

Key words: coding ORF, DNA asymmetry, gene finding, gene number, orphans, *Saccharomyces cerevisiae*.

Introduction

The July 1996 issue of Trends in Genetics was devoted to the *Saccharomyces cerevisiae* genome – the first fully sequenced eukaryotic genome. In two articles of this issue, the problem of orphans was discussed (CASARI et al. 1996, DUJON 1996). Orphans are open reading frames (ORFs) without any known function or homology to any other known gene. There is a paradox – in the previous genetic research, only a quarter of the yeast genes identified by traditional methods lacked homology to other genes already described in databases (OLIVER et al. 1992). Now, databases contain many more known genes and the fraction of identified ORFs without any homology to known genes seems to be much larger, too. It should be expected that the fraction of ORFs without

Received: March 1997.

Correspondence: S. CEBRAT, Institute of Microbiology, Wrocław University, Przybyszewskiego 63/77, 54-148 Wrocław, Poland. Email: cebrat@angband.microb.uni.wroc.pl.

any known function or homology should be rather shrinking when new genes are added to the databases. DUJON (1966) has described this phenomenon as the mystery of orphans.

It looks as if we have two worlds of genes – one with known functions or with function recognisable by available methods, and the other one with ORFs which escape any functional analysis. If Ockham was alive, he definitely would think about such an explanation.

Is there any other, simple explanation of the paradox of orphans? Why the previously found proportion of genes without homology was so small? Why such a considerable proportion of known ORFs lack any homology?

To solve the problem of orphans we propose to reconsider the accepted number of coding open reading frames.

Databases

Sequences for analysis were downloaded on September 23, 1996 from *genome.stanford.edu* Information on gene function and ORF homology and its presumed functions was downloaded on November 16, 1966 from *http://www.mips.biochem.mpg.de*. We have analysed the set of all ORFs longer than 100 codons (7440 ORFs), including all ORFs formerly discarded by the Saccharomyces Genome Database project (SGD). We have also analysed intergenic sequences. To avoid coding ORFs in the set of intergenic sequences we have analysed only intergenic regions longer than 100 triplets, outside ORFs longer than 70 codons (note that in this case the sum of nucleotides in ORFs and in the intergenic regions is lower than the total length of the genome).

Results and discussion

There are about 6200 ORFs in the yeast genome reported in the SGD. According to DUJON (1996), about 300-400 of these ORFs are questionable ORFs. It has been accepted that coding ORFs are:

– longer than 150 codons,
– from a pair of overlapping ORFs the longer one is coding,
– if the length of ORF is 100-150 codons, it is accepted as coding if its Codon Adaptation Index (CAI) is higher than 0.11 (DUJON et al. 1994).

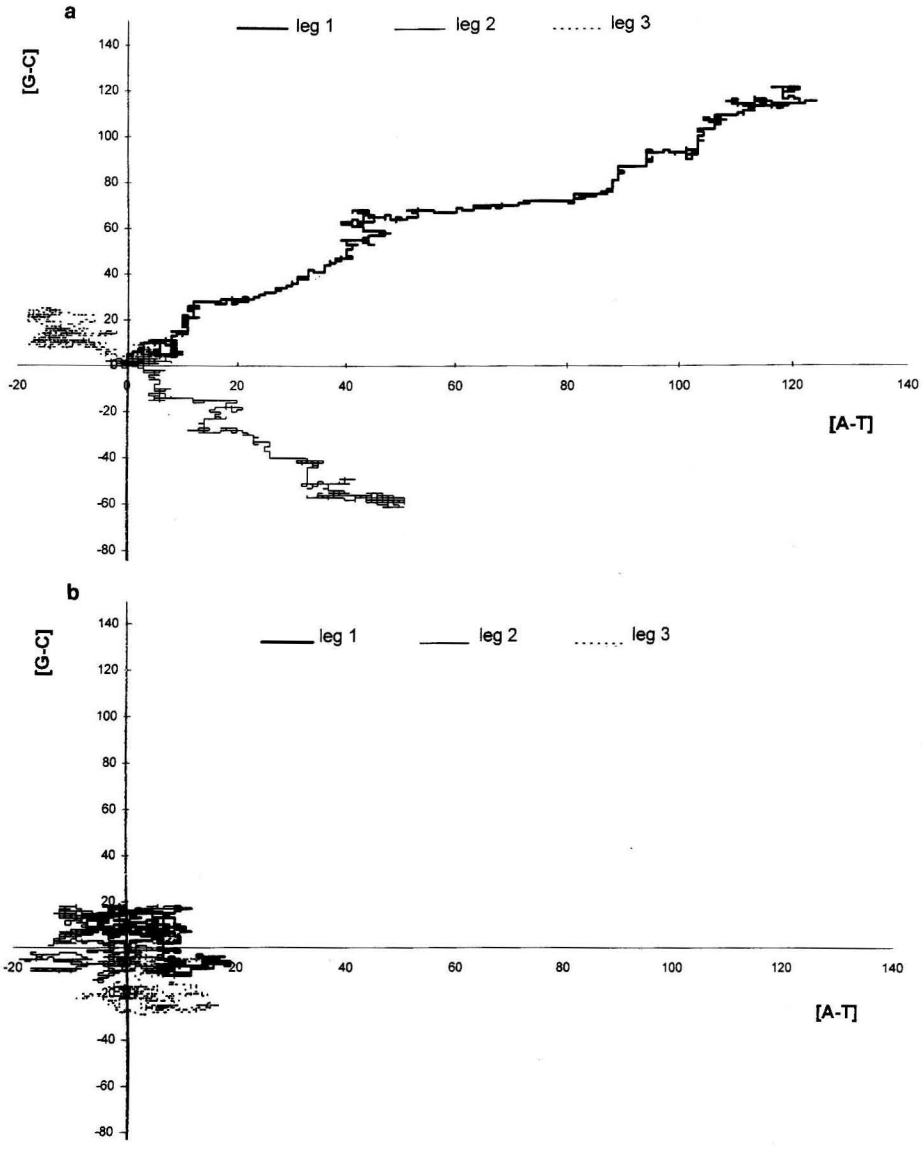

Fig. 1. Examples of spiders: a – for a coding of 841-codon-long sequence; b – for an intergenic 902-triplets-long sequence. Numbers 1, 2, 3 denote legs for the first, the second and the third positions in triplets.

We have found some criteria based on asymmetry sense/antisense strands of coding sequences, which can identify coding ORFs (CEBRAT et al. 1997a, b). The effect of the asymmetry in occupation of codon positions by nucleotides could be very well seen in Fig. 1 where we have presented the result of a DNA walk in two-dimensional space with the bases (A,T) on opposite ends of one

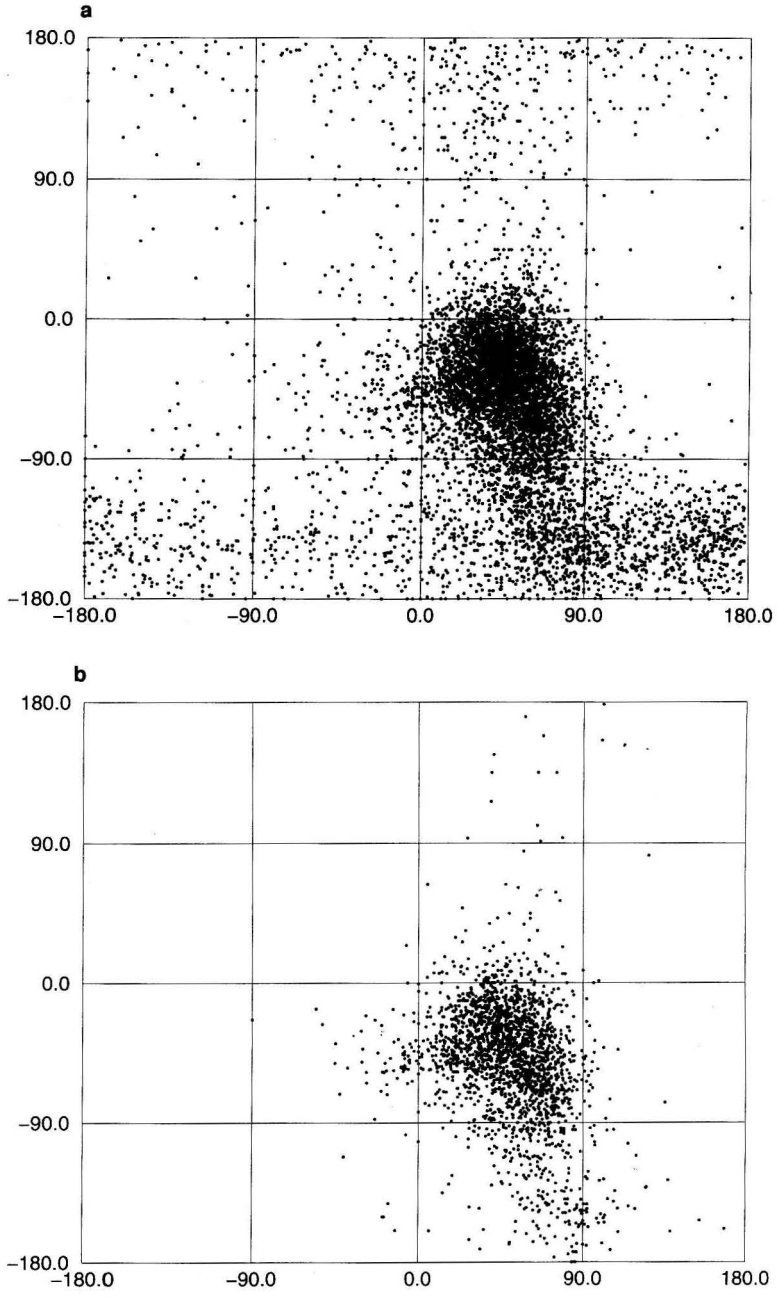

Fig. 2. Plots representing relations between slopes of leg 1 and slopes of leg 2 for:
a – all ORFs (7440) > 100 triplets, b – 2205 ORFs with identified functions

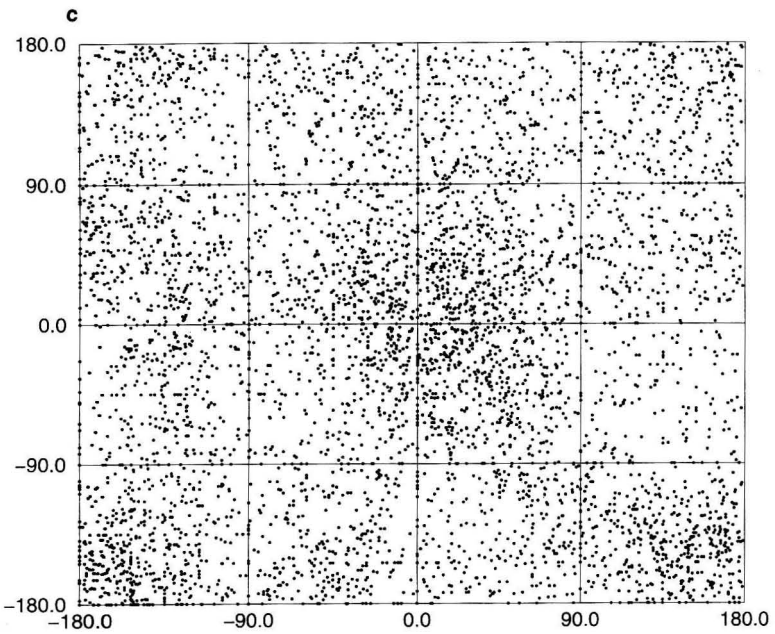

Fig. 2. c – 5137 intergenic sequences > 100 triplets.

axis and (G,C) on the other to represent DNA sequences (BERTHELSEN et al. 1992). The example in Fig. 1a corresponds to a gene of 841 codons with a known function, whereas Fig. 1b corresponds to an intergenic sequence of 902 triplets. We have used a significant modification of Berthelsen walk – in fact, for each sequence we have performed three DNA walks, independently for each nucleotide position in triplets. The first walker starts from the first nucleotide position of the first codon and then jumps every third nucleotide until the end of the examined sequence is reached. Similarly, the second and the third walkers start from the second and third nucleotide positions of the first codon, respectively. Every jump of a walker is associated with unit shift in two-dimensional space with the bases (A,T) and (G,C) depending on the type of nucleotide being visited. The jumps are: (0,1) for G, (1,0) for A, (0,–1) for C and (–1,0) for T. Hence, each DNA walk represents "history" of nucleotide composition of the first, the second and the third position of codons along the DNA sequence. The three walks together have been called a spider and a single walk has been called a leg of the spider. The spider in Fig. 1a is typical for coding sequences. Spiders representing coding ORFs usually have the first legs in the first quarter of the plot, the second legs at the fourth quarter and the third legs resembling Brownian walk. In the case of a typical intergenic sequence, as in Fig. 1b, all three legs resemble the trace of Brownian motion.

The simplest method of sequence analysis in terms of spiders is to estimate the co-ordinates (x,y) of the ends of spider legs. In our case x = [A-T] and y = [G-C] and the brackets denote the number of nucleotides. These criteria are based on the A/T and G/C ratios of the first and the second positions in codons. Simply, we have measured the ratio (G-C)/(A-T) for the first and second position in codons. To avoid infinite values of tangents, we have calculated the arctg (G-C)/(A-T). Next, we have plotted for each ORF the value for the first position in the codon versus the value for the second position. The results for all ORFs and for ORFs with known functions are plotted in Figs. 2a and 2b, respectively. Just to imagine how the plot representing the intergenic sequences looks like – Fig. 2c is presented. Using standard deviation for the normalisation of both parameters (x,y), we have determined the centre of gene distribution in the plot seen in Fig. 2a (details in CEBRAT et al. 1997b).

We have represented each individual ORF by values A_{s1} and A_{s2} for the first and second leg respectively, where:

$$A_{s1} = \frac{(A_1 - \overline{A}_1)}{SDEV_{A_1}} \quad \text{and} \quad A_{s2} = \frac{(A_2 - \overline{A}_2)}{SDEV_{A_2}} \ ,$$

where: A_1 and A_2 are the values of slopes in degrees of the first and second legs, respectively; $\overline{A}_1$ and $\overline{A}_2$ are the average values of slopes of the first and second legs, respectively, for all ORFs with known function; SDEV is a standard deviation for slopes of first and second legs for all ORFs with known function.

Next we have counted for each ORF the values:

$$A = \sqrt{A_{s1}^2 + A_{s2}^2} \ .$$

In this equation A is a distance in two-dimensional space from the centre of the set representing ORFs with known functions, measured in SD units.

To estimate the number of coding ORFs we have calculated the number of genes inside the space determined by A, for a given value of A, and we have estimated the ratio between the number of genes inside the determined space to the number of genes outside the space. Next, we have counted the number of ORFs from the set of all yeast ORFs inside the same surface and assuming that they are presumably coding, we have counted, from the ratio for genes, how many of the presumably coding ORFs should be expected outside the space. We have added these two values and the sums plotted in Fig. 3 as an approximate number of coding ORFs versus distance [A]. We have repeated this calculation for A values of 2000 genes with the largest A values. We have stopped the calculation for surfaces so small that one gene inside the space

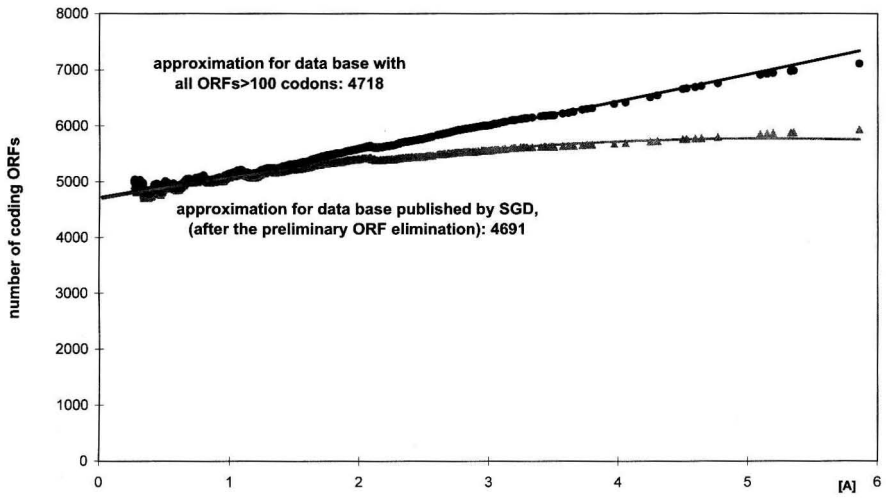

Fig. 3. Approximation of the total number of coding ORFs in the yeast genome, in the set of all 7440 ORFs > 100 codons and in the set of 6095 ORFs published in the SGD

decided about dozens of ORFs outside the space (the error of estimation for smaller A values is growing). We have obtained a plot shown in Fig. 3. The extrapolated line crosses the axis at 4718 for our database (7440 ORFs).

We have repeated the same estimation for the set of ORFs published in the SGD. In this database about 1300 ORFs have been eliminated by preliminary criteria used by the SGD project. For this database the estimated number of coding ORFs was 4691. The smaller number for SGD probably means that the preliminary elimination of some ORFs was not justified. Nevertheless, the par of both estimations is striking.

These estimations mean that since there are about 2300 ORFs with known functions and about 2000 ORFs with known homology to other genes, there are only about 400 coding ORFs without any identified function or homology. Then, there is about 8.5% of orphans without any identified function or homology, what should be expected from previous studies of OLIVER et al. (1992), when the first set of ORFs with the first fully sequenced yeast chromosome has been analysed in respect to homology of ORFs to known genes. In the SGD programme, it has been accepted 5800 coding ORFs (according to criteria described above). Thus, among these ORFs about 1500 orphans should be expected (all ORFs minus genes and ORFs with known homology), which

means that there are still more than a quarter of orphans in the set of yeast ORFs. Has the Ockham razor cut the mystery of orphans?

To accept this hypothesis, the synchronisation of transcript map with the map of ORFs should be revisited or at least reconsidered.

Acknowledgements: This work was supported by The State Committee for Scientific Research, grant number 1016/S/IMi/97.

REFERENCES

BERTHELSEN Ch.L., GLAZIER J.A., SKOLNICK M.H. (1992). Global fractal dimension of human DNA sequences treated as pseudorandom walks. Phys. Rev. A45: 8902-8913.

CASARI G., de DRUVAR A., SANDER C., SCHNEIDER R. (1996). Bioinformatics and the discovery of gene function. Trends in Genetics 12: 244-255.

CEBRAT S., DUDEK M.R., ROGOWSKA A. (1997a). Discrimination of Open Reading Frames as protein coding sequences. J. Appl. Genet. 38: 1-9.

CEBRAT S., DUDEK M.R., MACKIEWICZ P., KOWALCZUK M., FITA M. (1997b). Asymmetry of coding versus non-coding strands in coding sequences of different genomes. Microbial & Comparative Genomics 2: 259-268.

DUJON B. and 106 co-authors (1994). Complete DNA sequence of yeast chromosome XI. Nature 369: 371-378.

DUJON B. (1996). The yeast genome project, what did we learn. Trends in Genetics 12: 263-270.

OLIVER and 146 co-authors (1992). Complete DNA sequence of yeast chromosome III. Nature 357: 38-46.