RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS
LABORATORY OF THEORETICAL GENETICS

PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE

Volume 2

BGRS'2000
Novosibirsk, Russia
August 7-11, 2000

ICG, Novosibirsk, 2000

NO MYSTERY OF ORFans IN GENOMICS - GENERATION OF ORFans IN THE ANTISENSE OF CODING SEQUENCES

Mackiewicz P., Kowalczuk M., Gierlik A., Szczepanik D., Nowicka A., Dudek M.R., *Cebrat S.

Institute of Microbiology, Wrocław University, Poland
e-mail: cebrat@angband.microb.uni.wroc.pl
*Corresponding author

Keywords: ORFan, Saccharomyces cerevisiae, gene number, coding probability, DNA asymmetry, DNA walk, random walk, long range correlation, antisense

Resume

Motivation:

Despite the growing number of known sequences coding for proteins or even completely sequenced genomes, the fraction of Open Reading Frames (ORFs) without known function or homology to other known coding sequences (so-called ORFans) is not diminishing. This phenomenon is known as Mystery of ORFans. There have been many attempts to explain this paradox but only one is in fact reasonable: a large fraction of ORFans do not code for proteins. Therefore, another problem arises: how these long, noncoding ORFs have been generated.

Introduction

Analyses of several completely sequenced genomes have revealed that many ORFs longer than 100 codons have no assigned functions or homologues. They make about one third of all ORFs in every genome. During sequencing of genomes the fraction of these ORFs (ORFans) grew much 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 paradox was called the ”mystery of orphans” [Dujon, 1996; Casari et al., 1996]. After researching updated databases for homologues, ORFans still exist in the number much higher than expected [Fischer & Eisenberg, 1999].

Because of problems with finding homologues for ORFans and classifying them to known protein families, many authors consider ORFans fast evolving proteins or sequences unique to an organism or to a closely related group of organisms. Assuming this, ORFans should form new unknown protein superfamilies of unique function and structure. If it was true, almost every ORFan should define a new superfamily and the number of protein superfamilies ought to be several times larger than earlier estimations [Fischer & Eisenberg, 1999].

We have approximated the total number of coding ORFs longer than 100 codons in the most intensively studied genome, yeast *Saccharomyces cerevisiae*. Based on the analysis of asymmetry between coding and non-coding strands, we have found no more than 4700-4800 coding ORFs in this genome [Cebrat et al., 1997; Cebrat et al., 1998a; Kowalczuk et al., 1999]. It is about 1000 less than 5800-6000 which is the total number of ORFs annotated in data bases [Goffeau et al., 1996; Mewes et al., 1997]. The result indicates that about 1000 ORFs considered ORFans in the yeast genome data bases should be eliminated as non-coding.

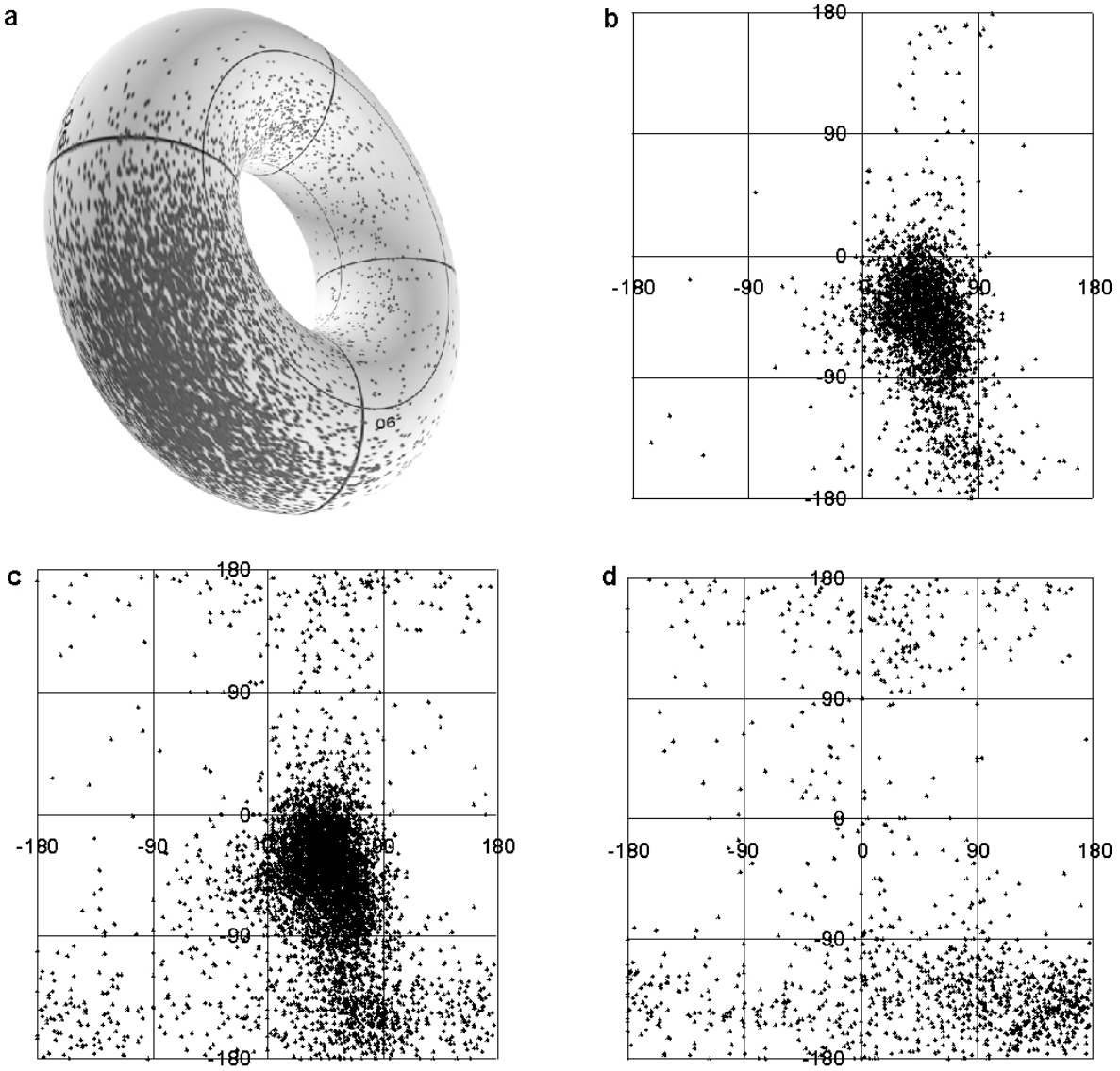
Thus, we suggest to use the Okham razor to solve the problem of ORFans - we just claim that the overwhelming fraction of ORFans do not code for proteins.

It is difficult to accept such high number of non-coding long ORFs in the yeast genome if we assume after Senapathy (1986), Sharp and Cowe (1991) that there is a small chance of occurring of long ORFs in a random sequence of the same size as the real yeast genome. Nevertheless, Cebrat and Dudek, (1996) and Cebrat et al., (1998b) have shown that the genetic code and coding sequences have specific properties of generating long ORFs, especially in the antisense strand.

In this paper we have shown that many ORFans in the yeast genome do not code for proteins and have been generated in protein coding sequences, mostly in their antisense strand.

Methods

We have parameterised coding sequence composition counting arctan([G-C]/[A-T]) for the first and the second codon positions separately. Each sequence is represented by a point on the surface of torus, with the values of these two parameters as co-ordinates. The distributions of different sets of ORFs from the yeast genome are presented in Fig. 1. For details of the method see our papers [Cebrat et al., 1997; Cebrat et al., 1998a] and our web page (http://smorfland.microb.uni.wroc.pl).



Figer 1. Distribution of ORFs of the yeast genome on the torus projection. a - distribution of all ORFs on the torus, b - distribution of ORFs with known phenotype on the torus projection, c - distribution of all ORFs annotated in MIPS data base, d - distribution of baby ORFs (generated inside coding sequences, see text for a more detailed explanation).

Results

ORFs with known phenotypes form a compact set of points on the torus surface - about 98% of sequences are situated on about 11% of the torus projection (Fig. 1b). When all ORFs annotated in the yeast data bases are plotted using this method, they form a much more dispersed set of points, but still not evenly dispersed (Fig. 1c).

Our previous studies [Cebrat & Dudek, 1996; Cebrat et al., 1998b] have shown that coding sequences preferentially generate noncoding overlapping ORFs (called *baby*) in the antisense in phases 3/3 and 2/2 (numbers indicate the positions in codons which overlap). In fact there are many (about 1500) such generated ORFs in the yeast genome. Their distribution is shown in Fig 1d.

Comparing all plots on Fig. 1 it is visible that many ORFs annotated in the yeast data bases are located in regions occupied by *baby* ORFs. On the other hand, these regions are very poor in known genes. It suggests that some ORFs (possessing properties of *baby* ORFs) have been generated in the same way as *baby* and probably do not code for proteins. Many of these ORFs may have arisen by ancient duplications of coding sequences nesting noncoding ORFs. Duplicated sequences accumulated mutations which eventually eliminated the proper reading frames of the original genes, leaving generated *baby* ORFs.

To prove that, we have translated the antisense of 2840 ORFs (without determined functions or distinct homologues, grouped in MIPS data base in classes 3-6). Then, we have searched protein databases for homologues using FASTA search program. We have found significant homologues for 757 ORFs with E value < 0.01 and for 603 ORFs with E value < 0.001 [Mackiewicz et al., 1999].

Table 1.Fractions of ORFs for which antisense homologues were found, depending on the distance to centre of distribution of ORFs with known phenotypes; A_d - distance to the centre, N - number of homologues found, N_f - number of homologues for which the generating phase was properly predicted.

A_d	Number of sequences	N	%	N_f	$\%_f$
1-2	886	111	12	5	4
2-3	351	65	18	38	58
>3	587	298	51	217	73

We have grouped analysed ORFs according to their distance from the centre of distribution of known genes on the torus projection. This distance is anti-correlated with the ORFs' coding probability. For half of ORFs with low coding probability we have found homologues for their antisense (Table 1). For about 70 % of these ORFs we have predicted properly (based on our base content parameters) frame in which they had been generated.

Almost 80% of generated ORFs arose in antisense, 50% of which in the sixth phase - overlapping 3/3 and 28% in the fourth phase - overlapping 2/2, which is in agreement with our previous observations on generating overlapping ORFs [Cebrat et al., 1998b].

Discussion

One would argue that the set of about 3000 ORFs with recognised phenotypes (Fig. 1b) is not representative for all coding sequences in the yeast genome - for unknown reasons. Thus, one has to accept an implication of his argument that ORFs coding for unknown protein superfamilies have very specific properties - they resemble the antisense of coding sequences at least in their nucleotide composition of specific positions in codons, since they are dispersed non-evenly on the torus surface, grouping preferentially in the regions where generated ORFs are grouped. Defenders of the larger number of coding ORFs in the yeast genomes could argue that, still for unknown reasons, perhaps structural constraints of DNA molecule, ORFans (their double strand structure) have to posses the overall nucleotide structure of normal, known coding sequences and the only difference between them and the already known genes is in the phase they are coding in and which strand is coding.

If we agree with such arguments another question would rise: should we expect homologues between the antisense of known genes and presumed product of ORFans? It is hard to assume that it would be very easy to adopt the antisense information for producing functional proteins. In fact we have found a few such homologues between known coding sequences, but they are very rare cases [Cebrat et al., 1998b]. If there are no phyllogenic relations between them, there should be no ORFans homologous to antisense of coding sequences. As we have proved, it is not the case, a lot of ORFans have homologues in the antisense of ORFs with known phenotypes.

If anybody still defended the position of the large number of coding ORFans in the genomes, they would have to accept another, perhaps very plausible hypothesis that duplication and exploiting the antisense for a new function is a very common way of new gene evolution. But in the view of the data shown above, one has to accept the implication: such "inverted genes" have very specific functions, because they preferentially escape the traditional methods of finding the gene phenotypes - if not, we could find genes with known phenotypes in both classes, "normal phase" and "inverted". Futhermore, they diverge much faster, evolving into huge number of genes coding for new protein superfamilies.

Our question is: why to not cut off all these beings with the Okham razor and accept the thesis that a lot of ORFs in the genomes have been generated inside coding sequences and by the common recombination events were translocated into other genome regions where they can accumulate mutations very fast but they are still visible as the "antisense pseudogenes"? In this thesis, there are no assumptions *(beings)* like these:

Two third of ORFs in the yeast genome with known phenotypes make a very unrepresentative set of all genes in this genome.

Protein coding ORFans for unknown reasons have structural properties of ORFs generated spontaneously in the highest frequency in the antisense of known coding sequences.

Many (coding!) ORFans developed a product function by simple reading their information in the antisense of another coding sequence.

The rate of divergence of antisense ORFans is much faster than that of normal genes.

The generation and evolution of ORFans is unidirectional - "normal genes" are primordial.

References

1. G. Casari, A. de Druvar, C. Sander and R. Shneider, "Bioinformatics and the discovery of gene function" Trends Genet. **12**, 244 (1996).
2. S. Cebrat and M.R. Dudek, "Generation of overlapping open reading frames" Trends Genet. **12**, 12 (1996).
3. S. Cebrat, M.R. Dudek and P. Mackiewicz, "Sequence asymmetry as a parameter indicating coding sequence in *Saccharomyces cerevisiae* genome" Theory in BioSciences **117**, 78 (1998a).
4. S. Cebrat, M.R.Dudek, P. Mackiewicz, M. Kowalczuk and M. Fita, "Asymmetry of Coding versus Noncoding Strand in Coding Sequences of Different Genomes" Microb. & Comp. Genom. **2**, 259 (1997).
5. S.Cebrat, P. Mackiewicz and M.R. Dudek, "The role of the genetic code in generating new coding sequences inside existing genes" Biosystems **42**, 165 (1998b).
6. B. Dujon, "The yeast genome project, what did we learn" Trends Genet. **12**, 263 (1996).
7. D. Fischer and D. Eisenberg "Finding families for genomic ORFans" Bioinformatics **15**, 759 (1999).
8. A. Goffeau, B.G. Barrel, H. Bussey, R.W. Davis, B. Dujon, H.Feldmann, F.Galibert, J.D. Hoheisel, C. Jacq, M. Johnston et al., "Life with 6000 genes" Science **274**, 546 (1996).
9. M. Kowalczuk, P. Mackiewicz, A. Gierlik , M.R. Dudek and S. Cebrat, "Total Number of Coding Open Reading Frames in the Yeast Genome" Yeast **15**, 1031 (1999).
10. P. Mackiewicz, M. Kowalczuk, A. Gierlik, M.R. Dudek and S. Cebrat, "Origin and properties of noncoding ORFs in the yeast genome" Nucleic Acids Res. **27**, 3503 (1999).
11. H.-W.Mewes, K.Albermann, M.Bähr, D.Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S.G. Oliver, F. Pfeiffer and A. Zollner, "Overview of the yeast genome" Nature **387**, 7 (1997).
12. P. Senapathy "Origin of eukaryotic introns: a hypothesis, based on codon distribution statistics in genes, and its implications". Proc. Natl. Acad. Sci. USA **83**, 2133 (1986).
13. P.M. Sharp and E. Cowe, "Synonymous codon usage in *Saccharomyces cerevisiae*" Yeast **7**, 657 (1991).