International Journal of Modern Physics C, Vol. 10, No. 4 (1999) 635–643

SOME HINTS ON OPEN READING FRAME STATISTICS — HOW ORF LENGTH DEPENDS ON SELECTION

AGNIESZKA GIERLIK, PAWEŁ MACKIEWICZ,
MARIA KOWALCZUK, and STANISŁAW CEBRAT*
Institute of Microbiology, University of Wrocław, ul. Przybyszewskiego 63/77
54-148 Wrocław, Poland
**E-mail*: *cebrat@microb.uni.wroc.pl*

MIROSŁAW R. DUDEK
Institute of Theoretical Physics, University of Wrocław, pl. Maxa Borna 9
50-204 Wrocław, Poland

Received 5 February 1999
Revised 9 February 1999

Coding sequences of DNA generate Open Reading Frames (ORFs) inside them with much higher frequency than random DNA sequences do, especially in the antisense strand. This is a specific feature of the genetic code. Since coding sequences are selected for their length, the generated ORFs are indirect results of this selection and their length is also influenced by selection. That is why ORFs found in any genome, even much longer ones than those spontaneously generated in random DNA sequences, should be considered as two different sets of ORFs: The first one coding for proteins, the second one generated by the coding ORFs. Even intergenic sequences possess greater capacity for generating ORFs than random DNA sequences of the same nucleotide composition, which seems to be a premise that intergenic sequences were generated from coding sequences by recombinational mechanisms.

Keywords: Open Reading Frame; Coding Density; ORF Generation; ORF Length; Genetic Code Properties.

1. Introduction

One of the main functions of a DNA molecule is coding for amino-acid sequences of proteins. Proteins play many different roles in every organism, determining its inherited properties. In the overwhelming number of cases of known proteins, there is a colinearity of nucleotide sequence of adenines (A), thymines (T), guanines (G) and cytosines (C) in DNA, and amino-acid sequence in the polypeptide. However, in higher eukaryotes and in some instances in other organisms and viruses, a nucleotide sequence coding for one polypeptide is interrupted by noncoding ones (introns), while in eubacteria, a sequence coding for one polypeptide is almost

exclusively uninterrupted. In coding sequences, information is coded by triplets — each three-nucleotide sequence means one amino-acid. The proper "reading frame" is determined by the start codon ATG meaning "start translation" (for amino-acid sequence). The end of the coded protein is marked by one of the three stop-translation codons — TAA, TAG or TGA. There are some small differences resulting from redefinition of codon meanings during the phylogenesis of some genomes, changing the composition of start and stop codons, but this will not influence significantly our problems discussed in this paper.

A sequence of trinucleotides beginning with the start codon and ending with a stop codon is called an Open Reading Frame (ORF). The length of an ORF is defined by a value k, which represents the number of triplets of nucleotides between start and stop codons, usually taking into account the start codon and not the stop codon, that corresponds to the number of presumably coded amino-acids. At this point there is the first possible misunderstanding between physicists and biologists or even among biologists themselves. Sometimes, for a biologist, an ORF is a presumably coding sequence (gene) by definition. ORFs whose coding probabilities are very low are not even listed in databases. That is why usually there are no ORFs shorter than 100 codons listed in databases, unless it was found experimentally that they are really coding. In the yeast (*Saccharomyces cerevisiae*) databases, the criteria are even more stringent; it has been accepted that an ORF is not listed if its length is below 150 codons and the Codon Adaptation Index (CAI) below 0.11.[1] CAI is a parameter indicating how the codon composition of ORF follows the preferences of codon usage in the genome.[2]

While analyzing ORFs, one should keep in mind that the information carried by ORFs is not translated into proteins in one step. First, the nucleotide sequence of DNA is transcribed by DNA-dependent RNA polymerase and then the product of transcription — mRNA — is translated into an amino-acid sequence. Translation mechanisms recognize start and stop codons in the mRNA molecule, the reading frame being determined by the start codon and triplet structure of the coding sequence. For transcription mechanisms, the triplet structure of the nucleotide sequence does not exist. Signals which are quite different from start and stop translation are used instead. They are usually only topologically related. There are some other very important mechanisms which are fundamental for saving ORFs' coding information and which do not recognize their triplet structure or even do not recognize that the sequences are protein coding at all. They are: replication, replication-associated mutational pressure, and recombination. Transcription-associated mutational pressure does not see the codon structure but could preferentially affect coding sequences.[3–5] Nevertheless, there is one very important mechanism responsible for the stability of coding information — it is selection. Selection is responsible for both the length of ORFs and their nucleotide composition. Since the coding role of nucleotides in the three positions in codons is different, the effect of mutational pressure and selection on these positions is

different, resulting in varying representations of the four nucleotides at these positions in various genomes.[6–10]

The structure of the genetic code itself plays a very important role in the stability of genes against mutations and selection especially its degeneration. Keeping all this in mind, we can expect that both the nucleotide sequences belonging to ORFs and their length are not random. When we add the evidently highly sophisticated and nonrandom influence of the genetic code, to all these mechanisms affecting structure of ORFs then we can see how careful we should be when interpreting the results of statistical analyses of ORFs.

2. Size Distribution of ORFs

If we have a random DNA sequence with the frequencies of A, T, G and C equal to p_A, p_T, p_G, p_C then the probability of generating an ORF with k triplets in one DNA strand can be estimated by the following expression (see also Ref. 11):

$$P(k) = \frac{3}{2}(p_A p_T p_G)(2p_A p_T p_G + p_A^2 p_T)(1 - 2p_A p_T p_G - p_A^2 p_T)^{k-1} \qquad (1)$$

which uses the information about the frequencies of A, T and G only, because the start codons ATG and one of the three stop codons TGA, TAG or TAA uniquely determine the ORF. In Eq. (1), the overall factor 3 reflects the triplet structure of the genetic code. The factor is divided by 2 because coding sequences make sense only in one direction. It is evident that even in very long DNA sequences, the probability of random generation of ORFs longer than 100 codons in one DNA strand is relatively low and the number of ORFs diminishes exponentially with their length.[12] The frequency of long ORFs in natural genomes is quite different from that in random DNA sequences, indicating that there is no simple relation between nucleotide composition of the DNA molecule and the length of ORFs. To show this, in Fig. 1 we present the distribution of ORFs versus their length for both the whole yeast genome (about 12.5 M base pairs = bp) and the random DNA sequence (computer-generated) of the same nucleotide composition, p_A, p_T, p_G, p_C, and length. There are many more long ORFs in the natural genome than in the corresponding random sequence, which is simply the result of competitive processes of selection and mutations. The difference already becomes significant when starting from ORFs of ~ 32 triplets. However, it is well known that there is a statistically negligible number of ORFs representing genes shorter than 100 triplets.[13] Thus, the question arises as to how selection mechanisms could be responsible for the short ORFs existing in the genome. The answer is: indirectly.

Coding sequences, due to specific triplet structure of the genetic code, can generate ORFs inside themselves with much higher frequencies than one could conclude from their nucleotide composition.[14] Two stop codons — TAA and TAG — have the first two nucleotides in palindromes, which means that in the complementary strand, the two nucleotides are read in the same order, i.e., TA. If any purine

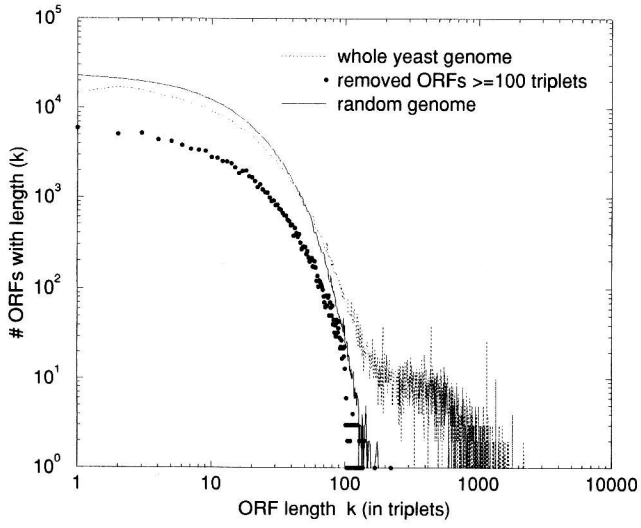

Fig. 1. Dependence of the number of ORFs on their length k (in triplets) in three cases: yeast genome (16 chromosomes), sequence obtained after removing from yeast genome all ORFs with size $k > 100$ triplets, and computer-generated random genome of the same composition as in yeast genome.

(A or G) follows the TA dinucleotide, a stop codon is generated. However, there are no stop codons inside a coding sequence. Thus, stop codons are not generated in the complementary strand by this mechanism. The smaller number of stop codons implies the existence of longer ORFs. Furthermore, there are very strong preferences for the phases in which these ORFs are generated.[13–15] To show the result of ORF generation by coding sequences, we cut off all ORFs longer than 100 triplets from the yeast chromosomes. The remaining sequences were spliced together, and ORF distribution in this new artificial sequence was prepared again. As seen in Fig. 1, the sharp cut-off at the position of 100 triplets is accompanied by significant diminishing of the number of shorter ORFs (a few ORFs longer than 100 triplets, which we can observe in Fig. 1, are generated at splicing sites). The effect of the cut-off would be much stronger if all ORFs overlapping the longer ORFs were cut off. This result means that a lot of shorter ORFs were discarded with the longer ones. These discarded ORFs are nested in the longer ORFs. In our previous study,[14,16] we have called these longer ORFs "mummy" ORFs and the shorter ones "baby" ORFs, and we have shown that there are genetic code properties which determine topological relations between mummy and baby ORFs.

3. Coding Sequences Generate ORFs

Let us assume that the first three nucleotides in a chromosome determine the first reading frame for the whole DNA strand. Since this frame has little to do with the real reading frame, we call it a phase. The DNA molecule could and should be read

Table 1.

Codon	No. of codons ATG, TAG, TAA, TGA generated by splicing ORFs ($k >= 100$) of phase (1) of all 16 yeast chrs., read in phases (1)–(6)						No. of codons in phase (1) of 16 yeast chrs.
	(1)	(2)	(3)	(4)	(5)	(6)	(1)
ATG (start)	11170	17244	3719	7218	9033	10014	73897
TAG (amber stop)	281	9283	7223	6989	11158	3454	52088
TAA (ochre stop)	571	11699	17092	13533	11603	5717	88379
TGA (opal stop)	411	12823	21396	10477	11539	9395	81350
Sum of stop codons	1263	33805	45711	30999	34881	18566	221817
Number of codons/stop	412	15	11	17	15	28	18

in six different phases[13–15] — three in one strand and three in the complementary, anti-parallel strand, read in the opposite direction. Every ORF lies in one and only one of the six phases. Imagine now that we read all the 16 yeast chromosomes in their first phase and spliced (joined the stop of an ORF to the start of the next one) all ORFs longer than 100 codons found in that phase. Next we read the resulting artificial sequence in all its six phases and counted the frequency of start and stop codons in all phases (data presented in Table 1).

The first phase simply represents the statistics of the spliced ORFs. From these data, the effect of generating ORFs inside coding sequences could be deduced. The analyzed sequence has exactly the same nucleotide composition ([G+C]/[A+T]) in all phases and has quite different ability of generating long ORFs. Furthermore, in the same phase, the occurrence of three codons (ATG, TAG and TGA) with the same nucleotide composition are quite different. For the third phase of the spliced sequence, the numbers are 3719, 7223 and 21396, respectively. The first codon in Table 1 means "start translation" and the last two codons mean "stop translation." Note that the distribution and frequencies of start and stop codons determines the length of ORFs. No statistical methods are needed to see the difference which is set by selection of coding sequences on the length of ORFs lying in other phases. One may conclude that looking for a simple relation between nucleotide composition of genomes and the length of ORFs or their coding density makes no sense. Selection and genetic code properties play too dominant a role to accept a hypothesis that the length of ORFs in the genome depends simply on the (G+C) fraction in this genome.

4. Coding Density

The fraction of nucleotides found in protein coding sequences of a particular genome is called coding density. To determine the coding density of any genome, it is important to estimate the coding probability of ORFs or to accept the same, or

Table 2.

Genome	Length of Sequence [bp]	Fraction of Nucleotides in ORFs with cut-off at $k = 100$	$k = 500$	$k = 1000$	$k = 3000$
E. coli	4639221	1.65	0.87	0.55	0.04
Bacillus subtilis	4214814	1.54	0.77	0.49	0.06
Borrelia burgdorferi	910724	1.12	0.82	0.55	0.04
Mycobacterium tuberculosis	4411529	2.12	1.12	0.64	0.09

at least roughly reasonable, features of coding sequences. Since the cases when one nucleotide sequence codes for two different proteins are extremely rare, usually the shorter ORF of an overlapping pair is discarded as noncoding. In some cases, it is possible to decide to which ORF the overlapping fragment belongs. If two overlapping ORFs are divergent (when they lie on opposite strands and overlap with start regions), it is possible that the overlapping fragment belongs to one ORF but both ORFs are coding. The same situation is possible for ORFs in tandem, but not for convergent ORFs (overlapping with terminal regions). Of two convergent ORFs, usually only one is coding. It doesn't make much sense to accept the length of ORFs as the only criterion in counting coding density, especially when the lower limit of length is really low (i.e., 33 codons). Li in his recent paper[17] compared coding densities for bacterial and yeast genomes. He presented the results as the number of ORFs per length of the analyzed sequence instead of the fraction of nucleotides in coding sequences (or presumably coding ORFs). If he had calculated the fraction of nucleotides, he would have noticed that, e.g., in the *Escherichia coli* genome (*E. coli*), the number of nucleotides in ORFs longer than 100 bp is much higher than the total length of the genome. This means that most of the nucleotides in the *E. coli* genome are in more than one ORF (ORFs are mostly overlapping). In the *E. coli* genome as well as in *Bacillus subtilis* and *Mycobacterium tuberculosis*, the number of overlapping ORFs is so high that after setting the low limit for the ORF length k, most of the nucleotides are shared by more than one ORF (Table 2).

Only some of these ORFs are coding. Thus, there is no point in doing coding density statistics on such sets of ORFs. Furthermore, genomes differ in their properties of generating overlapping ORFs. There are relatively less overlapping ORFs in *Borrelia burgdorferi*, *Treponema pallidum* or *Haemophillus influenzae* genomes. Li in his recent paper[17] counted coding densities of genomes for different lower limits of ORF length. It would seem reasonable to elevate the limit to 1000 bp for example, because the probability that such long ORFs are not coding is relatively low, but it is difficult to conclude about coding density from the results which take into consideration only some ORFs, if we do not know the length distribution of coding ORFs of the compared genomes. It is known that prokaryotes cannot pro-

cess synthesis of very long proteins and longer ORFs are under-represented in their genomes in comparison with eukaryotes, which Li[17] succeeded in showing.

5. ORFs Generated in Intergenic Sequences

It should not be concluded from the above results that the higher probability of ORF generation by protein coding sequences is a feature of genes only. We have found that in the yeast genome, a lot of intergenic sequences share this feature with coding sequences. This feature of intergenic sequences can be deduced also from Fig. 2, which presents the results shown in Fig. 1 but instead of the absolute number of ORFs, the number of ORFs of a given length normalized by the total number of ORFs in the genome have been shown (here, the total number of ORFs are different in random and natural genomes). When ORFs longer than 100 codons are cut off the genome, the distribution of the length of ORFs shorter than 100 codons stays the same as for the whole genome and is significantly different than for a random DNA sequence of the same nucleotide composition. We have found about 9000 more ORFs 32–100 codons long in the yeast genome than in the random DNA of the same nucleotide composition. On the other hand, there is experimental evidence that the number of coding ORFs of this length is statistically negligible (we estimate this number to be 3%–4%). It is known that even in intergenic sequences, triplet structure can be detected by looking for correlations in DNA molecules as has been discussed in papers by Voss,[18] Peng *et al.*,[19] and Buldyrev *et al.*,[20] In Fig. 3, we have shown how the amplitude of the peak corresponding

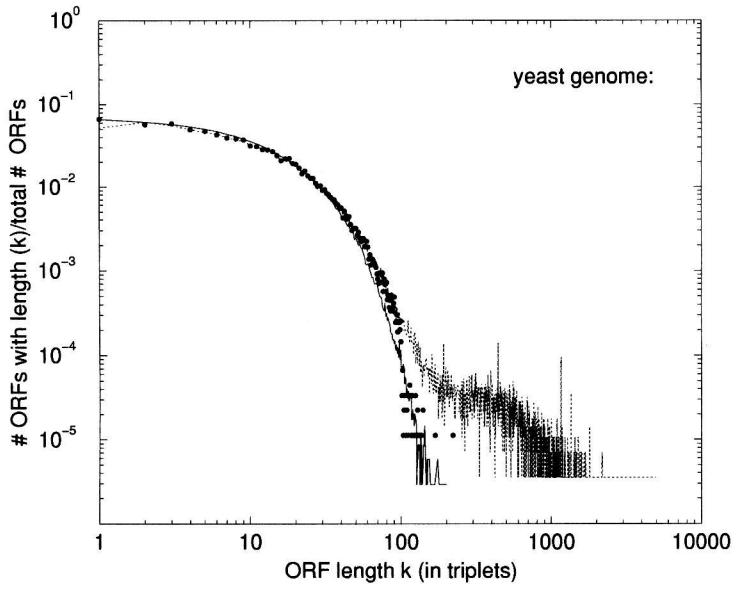

Fig. 2. The same as in Fig. 1, but the number of ORFs with length k are divided by the total number of ORFs.

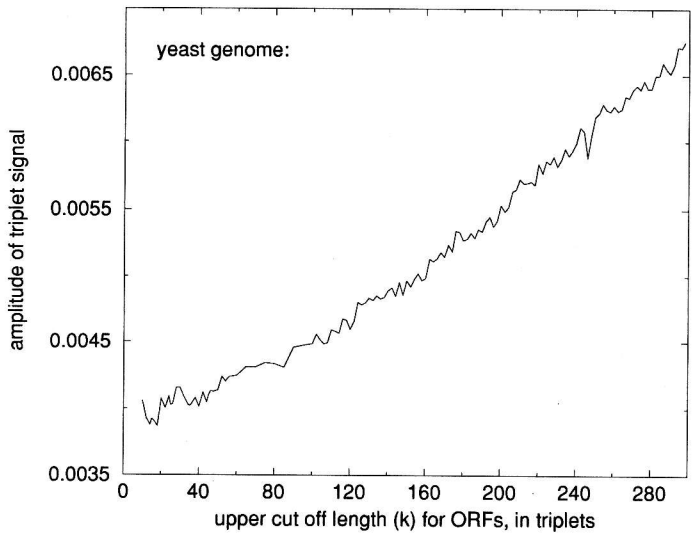

Fig. 3. Power spectrum for intergenic sequences in yeast genome has been calculated by FFT with a window size of 512 bases. We have shown how the amplitude of the FFT signal at one particular frequency $f_0 = 1/3$ bp^{-1} depends on the upper cut-off k of the size of the longest ORFs included in intergenic sequences.

to the frequency $1/3$ bp^{-1} in the Fast Fourier Transform (FFT) power spectrum calculated for intergenic sequences depends on the presence of longer ORFs. We notice that the signal $1/3$ bp^{-1} is significant even in sequences from which all ORFs longer than 50 triplets were cut off. Nucleotide composition of short ORFs found inside intergenic sequences indicates that they were generated by coding sequences in the past.[20] We have also found that there are about 750 ORFs longer than 100 codons (10% of all ORFs longer than 100 triplets) in the yeast genome which were generated by coding sequences and transferred into intergenic space by recombination processes. Many of these ORFs are listed in Munich Information Center For Protein Sequences (MIPS) databases as coding or presumably coding. In fact, they resemble coding sequences but not in their own reading frame. Read in a different frame, they give a "translation product" homologous to other genes. This property of generating ORFs in intergenic sequences can be explained by recombination mechanisms. In fact, the overwhelming fraction of intergenic sequences originated from coding sequences. Thus, the whole statistics of ORF length is strongly affected by selection and only slightly reflects the nucleotide composition of a genome.

6. Conclusions

The specific composition of coding sequences and their strong asymmetry together with the specific genetic code properties imply that the coding sequences generate Open Reading Frames (ORFs) inside them with much higher frequency than random DNA sequences do. Hence, not all of the ORFs represent genes.

The length of coding sequences is under direct selection. Thus, the length of ORFs generated by them is also influenced by selection. Our conclusion is that during statistical analyses of ORFs, it is necessary to divide them into two sets: One coding for proteins, and the other one generated by the coding ORFs. Otherwise, statistical analyses of overlapping ORFs produces additional correlations in the analyzed data of no biological sense. The problem is very acute in the case of the study of coding densities of genomes. Including ORFs generated by coding sequences in those studies produces paradoxical results which indicate that a lot of nucleotides code for more than one peptide. We stress that even intergenic sequences possess strong capacity of generating ORFs. It is greater than in random DNA sequences of the same nucleotide composition. This seems to be a premise that intergenic sequences were generated from coding sequences by recombinational mechanisms.

Acknowledgment

This work was supported by a KBN grant No. 6 PO4A 030 14.

References

1. B. Dujon *et al.*, *Nature* **369**, 371 (1994).
2. P. M. Sharp and W.-H. Li, *Nucl. Acids Res.* **15**, 1281 (1987).
3. M. P. Francino, L. Chao, M. A. Riley, and H. Ochman, *Science* **272**, 107–109 (1996).
4. M. P. Francino and H. Ochman, *Trends Genet.* **13**, 240 (1997).
5. J. M. Freeman, T. N. Plasterer, T. F. Smith, and S. C. Mohr, *Science* **279**, 1827 (1998).
6. G. Gutierrez, L. Marquez, and A. Marin, *Nucl. Acids Res.* **24**(13), 2525 (1996).
7. S. Cebrat, M. R. Dudek, P. Mackiewicz, M. Kowalczuk, and M. Fita, *Microbial & Comparative Genomics* **2**(4), 259 (1997).
8. J. Mrazek and S. Karlin, *Proc. Nat. Acad. Sci. USA* **95**, 3720 (1998).
9. J. Wang, *J. Biomol. Struct. Dyn.* **16**, 51 (1998).
10. C. T. Zhang and R. Zhang, *Nucl. Acids. Res.* **19**(22), 6313 (1991).
11. S. Cebrat, M. R. Dudek, and A. Rogowska, *J. Appl. Genet.* **38**(1), 1 (1997).
12. P. Senapathy, *Proc. Nat. Acad. Sci.* **83**, 2133 (1986).
13. B. Dujon, *Trends Genet.* **12**, 263 (1996).
14. S. Cebrat and M. R. Dudek, *Trends Genet.* **12**, 12 (1996).
15. B. Dujon and A. Goffeau, *Trends Genet.* **12**, Poster (1996).
16. S. Cebrat, M. R. Dudek, and P. Mackiewicz, *Theory Bioscienc.* **117**, 78 (1998).
17. W. Li, *submitted to Computer & Chemistry*, private communication.
18. R. Voss, *Phys. Rev. Lett.* **68**, 3805 (1992).
19. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, *Nature* **356**, 168 (1992).
20. S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley, *Phys. Rev. E* **51**, 5084 (1995).