J. Appl. Genet. 38(1), 1997, pp. 1-9

Asymmetry in nucleotide composition of sense and antisense strands as a parameter for discriminating open reading frames as protein coding sequences

Stanisław CEBRAT[1], Mirosław R. DUDEK[2], Adelina ROGOWSKA[1]

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

Abstract. Coding properties of yeast chromosomes were analysed and a strong asymmetry was found in nucleotide composition of sense and antisense strands. This property generates two very simple parameters – [A]/[T] and [G]/[C] of the sense strand – which could be used for discrimination of open reading frames as coding sequences with very high, statistically described level of significance. The paper contains a description of the method of ellipse of concentration in the two parameter space, which can close coding sequences inside, leaving a big fraction of noncoding sequences outside the ellipse.

Key words: coding function, ellipse of concentration, DNA asymmetry, open reading frames (ORFs), *Saccharomyces cerevisiae*.

Introduction

Sequencing programmes collected many open reading frames (ORFs), i.e., nucleotide sequences starting with the start codon (ATG) and ending with one of stop codons (TAA, TGA, TAG), but for most of them functions have not been identified yet.

The first step in looking for ORF function is its classification as a protein coding sequence. There are a lot of criteria which try to discriminate ORFs as coding sequences or indicate that some ORFs play a coding role in organism (see FICKETT 1996, for review). However, the most strong criterion is the length of ORF. Since there are very few proteins shorter than 100 aminoacids, usually

Received: June 1996.
Correspondence: S. CEBRAT, Institute of Microbiology, Wrocław University, Przybyszewskiego 63/77, 51-148 Wrocław, Poland. Email: cebrat@ungband.microb.uni.wroc.pl

the ORFs shorter than 100 codons are disqualified as coding sequences. The criterion of length seems to be acceptable, because the probability of generating the ORF longer than 100 codons in stochastic sequence is very low. In particular, if we assume that the nucleotides are distributed randomly and DNA phases are independent, then the probability to find a codon inside the ORF of size k (k has the meaning of the ORF length measured in codons) within the DNA strand could be calculated by the following formula:

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

where p_A, p_T, p_G are the frequencies of the A, T, G occurrence within a DNA strand and the overall factor reflects the presence of the DNA phases and codon structure. Then for $k > 100$ and the length of the stochastic chromosome being equal to the length of the yeast chromosome II one obtains only a few ORFs longer than 100 codons. One could expect, that it should be possible to obtain the length distribution of coding ORFs in a natural chromosome by subtraction the statistically predicted ORF distribution from the experimental one. DUJON et al. (1994) tried to correct the estimation of stochastic ORF frequency in chromosome XI by taking into consideration the frequency of dinucleotides in this chromosome. However, this expectation is false, because the nucleotide occurrence in chromosome is highly correlated (PENG et al., 1992, VOSS 1992, 1993) instead of being random.

We have observed that in each DNA phase the nucleotide composition of the sequences belonging to the ORFs shorter than 70 codons differs from the composition of the sequences in the ORFs longer than 130 codons. We have also found that there exist very strong relations between DNA phases (CEBRAT, DUDEK 1995). In particular, the selection for long ORF in one phase elevates the probability of ORFs generation in one related phase of opposite strand and simultaneously decreases this probability in other phases (CEBRAT, DUDEK 1996). This is an intrinsic feature of genetic code. The specific relations between phases (at least in coding regions) implicate that the probability of occurrence of long ORFs in chromosome with 60%-70% of sequences sequestered by coding sequences is much higher than in stochastic sequence of the same nucleotide (or even codon) composition (CEBRAT, DUDEK 1996).

Usually, ORFs with the relatively high CAI (Codon Adaptation Index, SHARP, LI 1987) or CBI (Codon Bias Index, BENNETZEN, HALL 1982) are classified as coding sequences. However, this method can discriminate also coding ORFs because these indices are low for genes with a slow rate of translation of their transcripts. The other strong criterion for ORF classification is its homology with the ORF of known function. However, even if we apply

the criteria of length and homology then still at least one third of ORFs found in yeast genome will stay without known function. Therefore, one should first estimate the probability that the given ORF is coding before laborious and costly genetic studies are undertaken. Below we show that the ratio of numbers of purines to pyrimidines can be significant for the first approximation of the probability.

Results and discussion

Our first observation was that the sense strand of ORFs > 130 codons cumulates purines. There is a very strong relation between the number of codons being the sum of all ORFs exceeding 130 codons in the phase and the differences [A]-[T] and [G]-[C], with the correlation coefficients 0.95 and

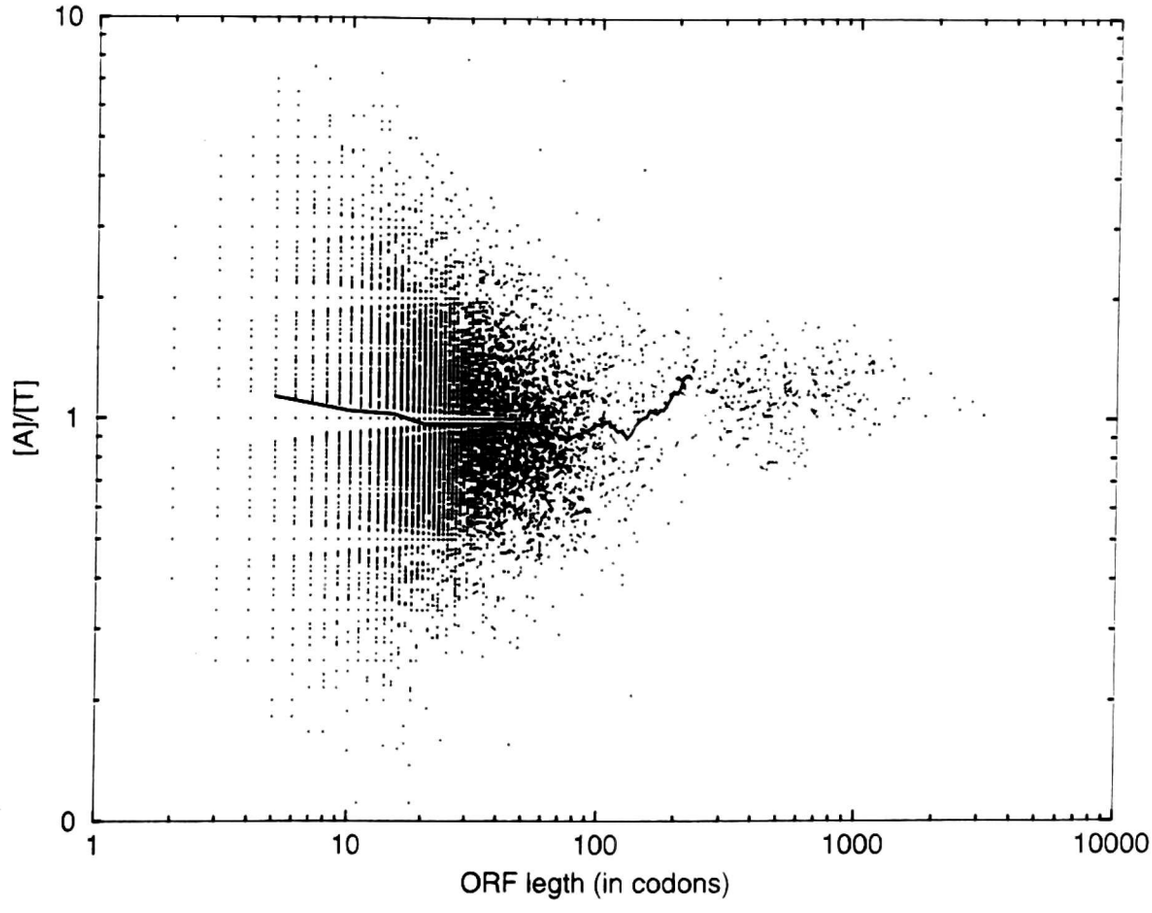

Fig. 1. Distribution of the A/T ratio with respect to ORF length for the yeast chromosome II. All sequences beginning from start codon end ending with stop codon have been considered as ORFs. Solid line represents averaged values of the A/T ratio.

0.72, respectively (the brackets [] mean the numbers of nucleotides). The results seem to be general, since the correlation was calculated for all phases of the chromosomes I, II, III, VI and XI.

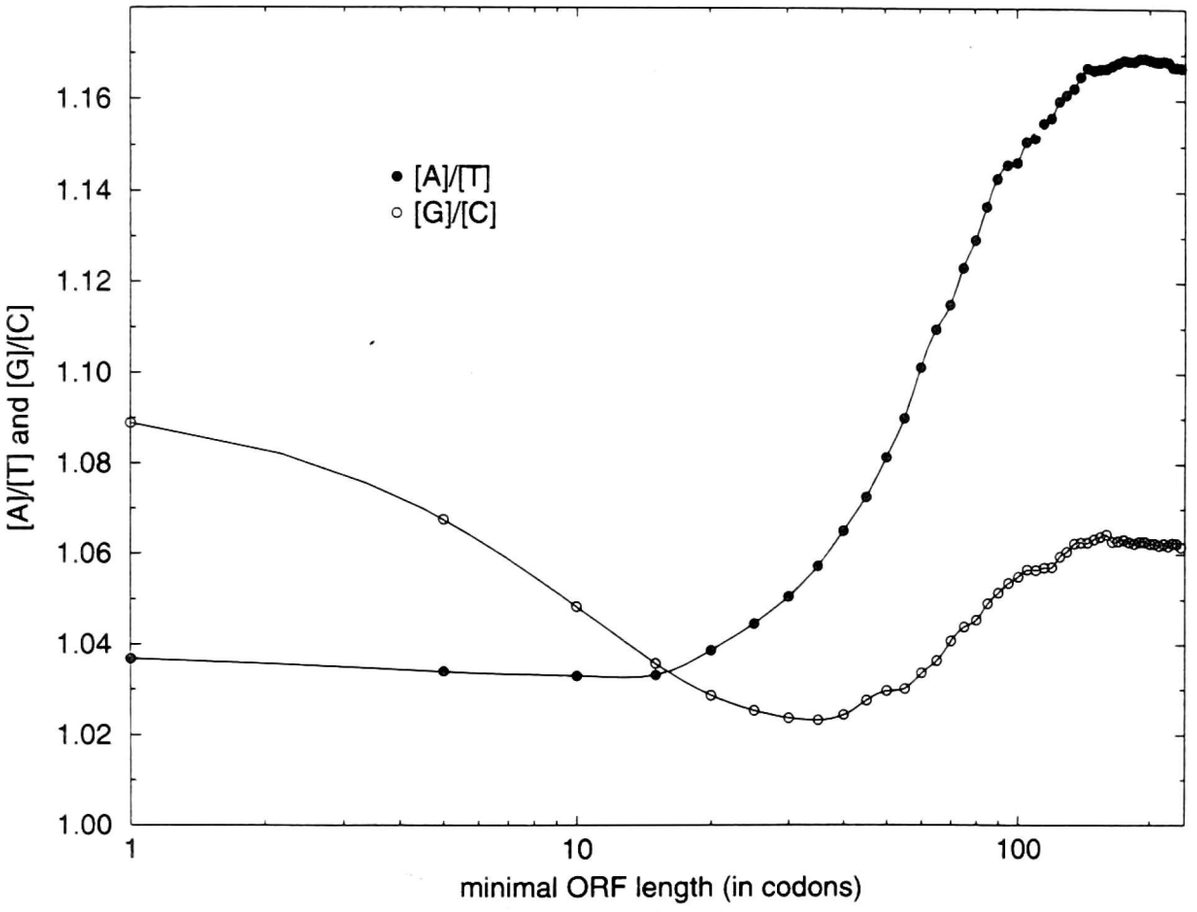

Fig. 2. [A]/[T] and [G]/[C] ratios calculated for all ORFs longer than the value of ORF length (in codons) shown on X-axis.

This finding means that in all examined yeast chromosomes the ORFs longer than 130 codons tend to contain more purines than pyrimidines in the sense strand. We conclude that there should be a specific relation between ORF length and its nucleotide composition. It is evident from Fig. 1 where the values

Table 1. List of genes and ORFs with homology to known genes which are outside the ellipse of concentration in the Fig. 4

Systematic ORF nomenclature	C.A.I.	Start	Gene name	Function or homology	Length in codons
YBL051c	0.187	116 760		Homology to RNP-binding proteins	668
YBR067c	0.449	366 107	TIP1	Temp. shock-inducible protein precursor SRP1/TIP1	210
YBR112c	0.161	456 877	SSN6	Transcription regulatory protein	966
YBR215w	0.150	647 451	HPC2	Cell-cycle regulatory protein, Regulator of S-phase transcription of histone genes	623
YBR266c	0.104	733 836		Probable membrane protein	187
YBR267w	0.168	734 143		Probable Zn-finger protein (C2H2 type)	295

Table 2. Absolute and relative nucleotide frequencies at the first, the second and the third positions in codons of all ORFs of yeast chromosome II

Position in codon	Number of nucleotides			Relations between nucleotide frequencies at different position in codon				
	A	T	G	C	A/T	G/C	A+T/G+C	A+G/C+T
1	69514	47926	58601	33784	1.51	1.84	1.36	1.57
2	72164	60260	30383	47018	1.17	0.74	1.71	0.95
3	61150	69059	39103	40513	0.94	1.15	1.60	0.97

of the [A]/[T] versus ORF length have been plotted in log-log scale. The solid line in Fig. 1 represents averaged values of the [A]/[T] ratio.

In Fig. 2 we show how the ratios [A]/[T] and [G]/[C] calculated from the sum of all ORFs (in six phases of chromosome II) exceeding the assumed minimal length depend on it. If the minimal length of ORF is close to 1 then the ratios [A]/[T] and [G]/[C] are relatively high, due to the number of the shortest ORFs dominating the results. Notice, that the shortest ORF consists only

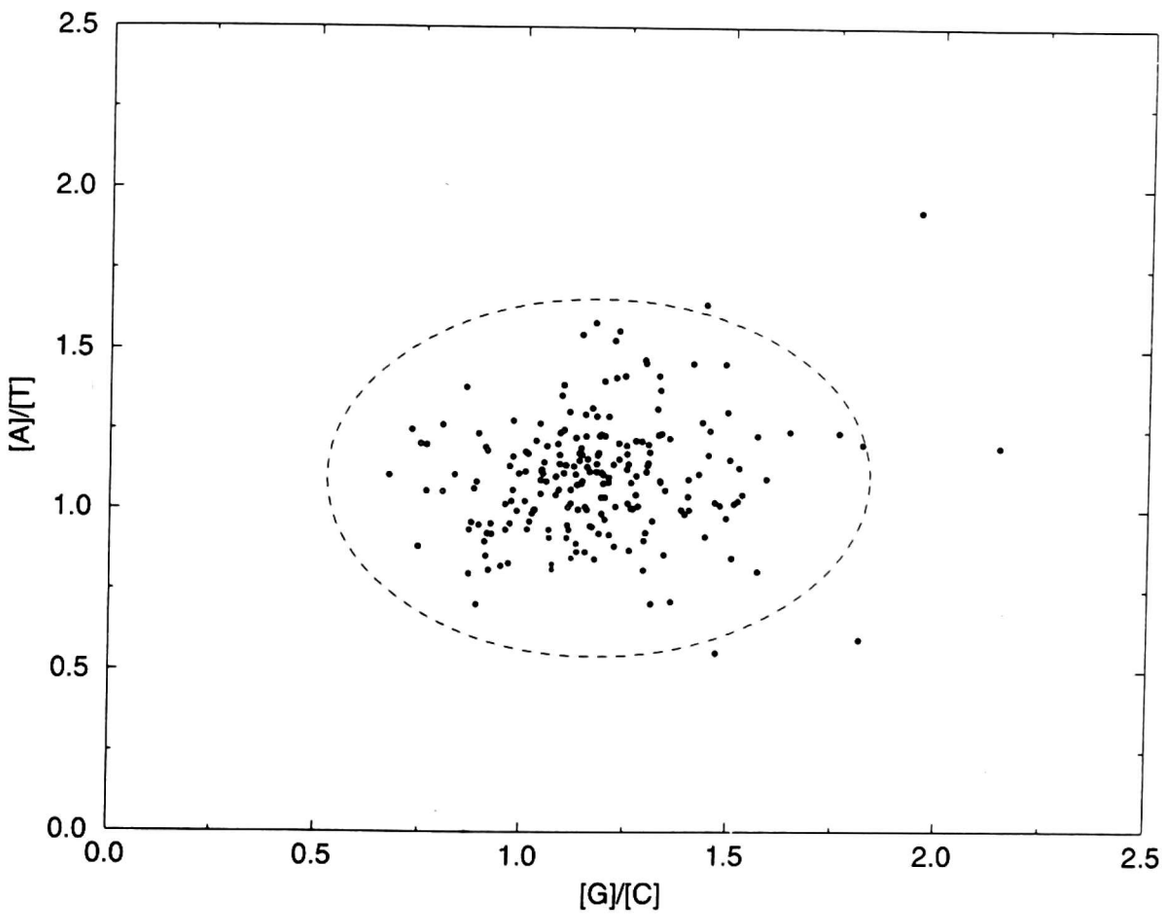

Fig. 3. Distribution od [A]/[T] ratio versus [G]/[C] ratio for 210 genes randomly chosen from LISTA. The centre of ellipse has been determined by the mean values of [A]/[T] and [G]/[C] and the half-axes are equal to their standard deviations multiplied by factor 3 (3 SD).

of codon start and codon stop being rich in purines, especially there is no cytosine in them. The influence of these codons on the total ORF composition diminishes with the elimination of the shortest frames, i.e., for higher value of the minimal ORF length. Once the minimal ORF length is increasing, the number of ORFs is decreasing and the ratios [A]/[T] and [G]/[C] start to reflect the features of coding sequences. The results presented in Fig. 2 suggest that sense strands of long ORFs become rich in purines.

To test this hypothesis we randomly drowned 210 sequences with known functions from LISTA (MOSSE et al. 1993) and plotted the values [A]/[T] against [G]/[C]. The results are presented in Fig. 3. Next, we have superimposed on the plot the ellipse of concentration, with the centre determined by the average values of [A]/[T] and [G]/[C] and the half-axes being equal to 3SD where SD is the standard deviation (we have assumed a normal distribution of [A]/[T] and [G]/[C] values). We have also chosen the ellipse axes to be parallel to the plot axes, once the correlation coefficient between data for [A]/[T] and [G]/[C] is negligible (0.058). As we had expected about 98% of ORFs were closed inside the ellipse of concentration.

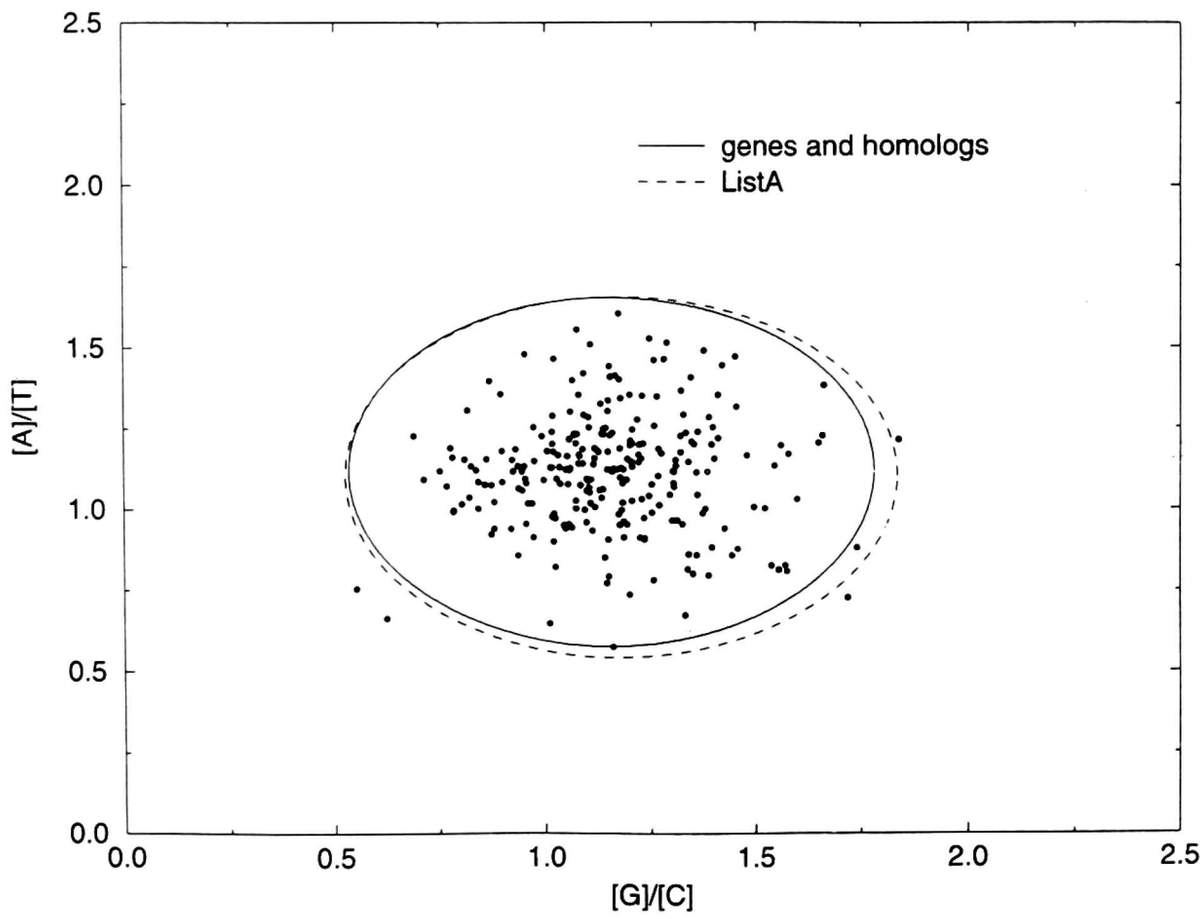

Fig. 4. Distribution of [A]/[T] ratio versus [G]/[C] ratio for all ORFs from yeast chromosome II with known functions and/or high homology to genes with known functions. The ellipses of concentration for these ORFs and for genes from LISTA represented in Fig. 3 were superimposed on the plot.

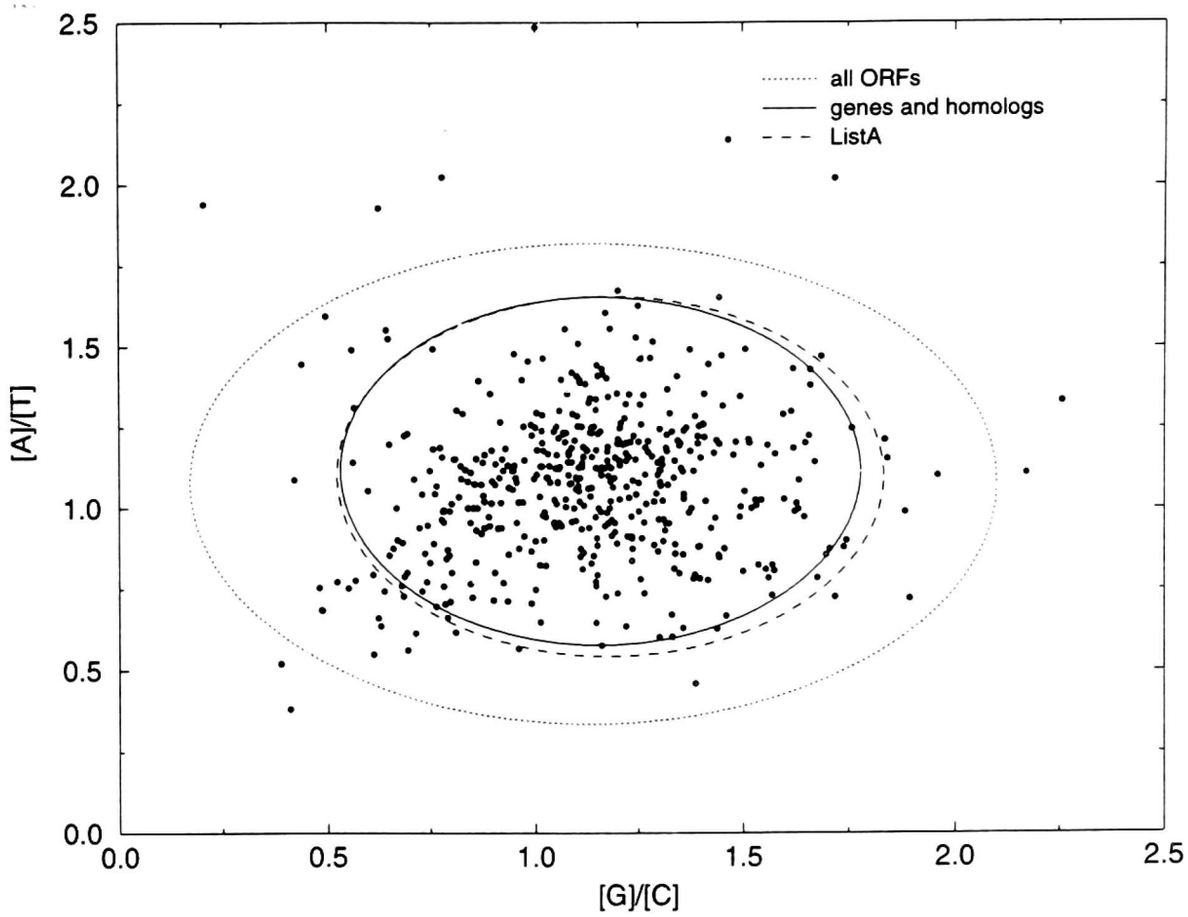

Fig. 5. Distribution of [A]/[T] ratio versus [G]/[C] ratio for all ORFs from yeast chromosome II longer than 100 codons. The ellipses of concentration for these ORFs, for ORFs of known functions and/or homology to known genes and for genes from LISTA were superimposed on the plot.

Ellipses of concentration for stochastic sequences are not presented because they are theoretically predictable. They should possess identical length of axes (to be circles in fact) and the co-ordinates of the centre of circle should be (1, 1). The only deviation ratio could be generated by start and stop codons, what should be negligible for longer ORFs.

Similarly, we have plotted [A]/[T] and [G]/[C] co-ordinates for data representing only these ORFs from the yeast chromosome II which have known functions or high homology to known genes (Fig. 4). Since there are no special listing of ORFs of yeast chromosome II with homology to known genes and without known function in the last edition of SacchDB (September 1996, genome-ftp.stanford.edu), we have taken the information about homology from the edition available in September 1995. Inside the ellipse of concentration we have found again 98% of ORFs and only five ORFs outside it. In Table 1 we present all these ORFs with their co-ordinates and CAI. In addition, in Fig. 4 the ellipse of concentration for 210 genes from LISTA has been superimposed onto the plot. It is evident that it almost ideally overlaps the ellipse specific

for the yeast chromosome II coding frames. That means that the ellipse is universal for yeast coding sequences. Nevertheless, if one considers all ORFs from the yeast chromosome II with the length longer than 100 codons, then there are 58 ORFs laying outside the ellipse of concentration for genes of known function (Fig. 5). Assuming the normal distribution of [A]/[T] and [G]/[C] parameters we can estimate the probability that these ORFs are coding for about 0.02. Notice, that the ellipse of concentration for all ORFs of the chromosome II in Fig. 5 has much longer axes than the ellipse encircling coding frames in Fig. 4 or in Fig. 3. what proves that the distribution parameters for coding ORFs are different than those for noncoding ones.

In Table 2 we present some data about the occurrence of nucleotides in the first, the second and the third positions in codons. It is easy to conclude that it is the first position which is mainly responsible for the observed asymmetry. The third position, influencing the CAI seems to be well balanced. Data presented in Table 2 explains clearly why correlation coefficients between CAI and [A]/[T] and [G]/[C] values in the sense strand are close to zero (0.08 and 0.01, respectively).

Conclusions

Two very simple parameters – [A]/[T] and [G]/[C] inside the sense strand can discriminate some ORFs as coding sequences with very high, statistically described level of significance. The ellipses of concentration, introduced by us, close the coding ORFs inside, leaving a lot of noncoding ORFs outside the ellipses. Still some noncoding ORFs can appear inside the ellipse of concentration. However, even this simple approximation which we have applied seriously reduces the effort during discrimination of ORFs. Parameters used in the described method are particularly useful in discrimination between two overlapping frames. Since the method is based on the asymmetry existing between sense and antisense strands of coding DNA sequence, it can be used for choosing one of two overlapping ORFs as coding sequence or for indication to which ORF the overlapped fragment belongs. There are two ways for the extension of our method to the more general. One is based on multidimensional space of discrimination where more ORF's characterising parameters are taken into consideration and the second – based on the finding that the observed asymmetry in sense and antisense strands is even more evident for nucleotides in different codon positions (manuscript in preparation). As we have shown in Table 2, the first position – in general – is rich in both purines

and the second in adenine only. The described method is highly independent of CBI and CAI indices. The indices show (in general) which codon from box is chosen for coding the aminoacid, what is highly correlated not only with the overall coding function but also with the level of translation. That means that the indices depends very strongly on nucleotide usage in the third codon position (especially CAI). Correlations in nucleotide appearance in the first two positions of codons, what we have presented, cannot be directly connected with the feature of genetic code degeneration, since this feature of genetic code depends almost totally on degeneration of the third codon position.

REFERENCES

BENNETZEN J.H., HALL B.D. (1982). Codon selection in yeast. J. Biol. Chem. 257: 3026-3031.

CEBRAT S., DUDEK M.R. (1995). Coding rhythm of DNA strands, preprint IFT Uwr August 893/95, submitted to Phys. Rev. Letts.

CEBRAT S., DUDEK M.R. (1996). Generation of overlapping open reading frames. Trends Genet. 12: 12.

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

FICKETT J.W. (1996). Finding genes by computer: the state of the art. Trends Genet. 12: 316-320.

MOSSE M-O., LINDER P., LAZOWSKA J., SLONIMSKI P. (1993). A comprehensive compilation of 1001 nucleotide sequences coding for proteins from the yeast, *Saccharomyces cerevisiae* (=LISTA2) Curr. Genet. 23: 66-91.

PENG C.-K., BULDYREV S.V., GOLDBERGER A.L., HAVLIN S., SCIORTINO F., SIMONS M., STANLEY H.E. (1992). Long-range correlations in nucleotide sequences. Nature 356: 168-170.

SHARP P.M., 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.

VOSS R. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences Phys. Rev. Letts. 68: 3805-3808.

VOSS R. (1993). Reply to the comment. Phys. Rev. Letts. 71: 1777.