Theory Bioscienc. (1998) 117: 78–89

Sequence asymmetry as a parameter indicating coding sequences in Saccharomyces cerevisiae genome

Stanisław Cebrat[1], Mirosław R. Dudek[2] and Paweł Mackiewicz[3]

[1]Institute of Microbiology, Wrocław University, ul Przybyszewskiego 63/77, 54-148 Wrocław, Poland. email: cebrat@angband.microb.uni.wroc.pl
[2]Institute of Theoretical Physics, Wrocław University, pl. Maxa Borna 9, 50-204 Wrocław, Poland. email: mdudek@ift.uni.wroc.pol
[3]Institute of Microbiology, Wrocław University, ul Przybyszewskiego 63/77, 54-148 Wrocław, Poland. email: pamac@angband.microb.uni.wroc.pl

Address for corresponding: Stanisław Cebrat, Institute of Microbiology, Wrocław University, ul Przybyszewskiego 63/77, 54-148 Wrocław, Poland. email: cebrat@angband.microb.uni.wroc.pl

Received: March 5, 1997; accepted: February 10, 1998

Key words: DNA asymmetry, orphan, open reading frame, coding probability, yeast genome.

Abbreviations: ORF – Open Reading Frame; SGD – *Saccharomyces* Genome Database; CAI – Codon Adaptation Index.

Summary: We have compared a symmetry in purine and pyrimidine occurrence in different codon positions of coding, presumably coding and noncoding sequences of the whole genome of *S. cerevisiae*. We have shown that there is a very strong asymmetry in sense versus antisense strand in nucleotide occurrence in the first and second positions in codons. Since the observed asymmetry results from specific composition of the first two codon positions – the parameter is not correlated with Codon Adaptation Index (CAI) and this property could be used as an independent parameter discriminating Open Reading Frames (ORFs) as coding sequences. We have also estimated the number of presumably coding ORFs in the *S. cerevisiae* genome as 4718 (without interrupted genes). This approximation has been done for all ORFs longer than 100 codons identified in the yeast genome. The same method of approximation performed for ORFs published by SGD program (after selection made before publication of the data base) gave the total number of 4691 coding ORFs. That means: a – the previously suggested number of coding ORFs is overestimated; b – some ORFs discarded by the first selection could be coding (if we assume that there is any significant difference between the two results cited above); c – the method of estimation is, at least roughly, correct since it eliminates more than 2700 noncoding ORFs from our database and about 1400 ORFs from the published SGD, leaving discrepancy for only 27 ORFs and resulting in almost the same number of coding ORFs.

Introduction

The extensive genome sequencing programme have inverted the routine methods in genetics. The previous way of gene describing begun with determining its function, (sometimes even the protein sequence) followed by mapping it on a chromosome, DNA sequence isolating, sequencing and predicting the coded protein structure. Now, due to genome sequencing programmes, at the first step a DNA sequence is available and next a putative function for it is searched. That is why a lot of software identifying coding sequences have been developed. They use the sequence similarity to other, already known genes or motifs, statistical regularity in codons appearance or specific consensus controlling the DNA trancriptional activity (see Fickett 1996, for review).

We have assumed, that theoretical, noncoding sequence (random?) is changed by forcing it to code the genetic information. This has to introduce changes, which should be statistically measurable. The property which is conserved for random sequences is a symmetry in nucleotide composition of both complementary strands. This symmetry is almost perfectly preserved for whole yeast chromosomes, but we have shown, that there is a strong asymmetry between sense and antisense strands of coding sequences (Cebrat, Dudek 1996a, Cebrat, Dudek and Rogowska, 1997). Sense strands of coding sequences are richer in purines. Such asymmetry could be also expected in sequences coding for genetic information other than for aminoacid sequences, but in these cases the asymmetry is not expected to reflect the triplet composition of the sequence. In this paper we are going to show, using special, two-dimensional DNA walks, that there are very strong rules introducing asymmetry in each nucleotide position in codons.

In 1996, the *Saccharomyces cerevisiae* genome sequence has been completed (for review, the special issue of Nature, 29 May 1997). The simple, standard analysis of this sequence has identified all Open Reading Frames (ORFs) in the genome. It has been assumed that an ORF can be considered as coding if its function has been already found, if it is longer than 100 codons or shows homology to known genes. Dujon et al. (1994) introduced also a definition of questionable ORFs as ORFs shorter than 150 codons and with Codon Adaptation Index (CAI) < 0.11. As to overlapping ORFs, usually the longer one is considered as to be coding, the shorter one usually is not represented in published *Saccharomyces* Genome Database (SGD). This rule concerns especially the ORFs whose sequences are entirely contained within other ORFs. The criterion of length is arbitrary and even in the *S. cerevisiae* genome genes coding for peptides shorter than 100 aminoacids are known and some cases when the shorter ORF of two overlapping ones is coding are also known. On the other hand, in the sequence of above 12 million of nucleotides, a lot of

ORFs longer than 100 could be randomly generated. Furthermore, long coding sequences generate other ORFs with much higher probability than random sequence. This is due to specific genetic code properties – two out of three stop codons possess two nucleotide palindromes and can generate stop codons in the related phase of the opposite strand (Cebrat, Dudek, 1996b). The start codon has the same property. It can generate another start in the opposite strand. These two features cause that inside the ORF identified in one strand the frequency of stop codons in the related phase of opposite strand is lower and frequency of starts is higher than in random sequence (Cebrat, Mackiewicz, Dudek, 1997). Nevertheless, there is no reason to conclude definitely that the longer ORF is coding and the shorter one is noncoding.

Material and methods

Sequences for analysis were downloaded September 23 1996 from *genome-ftp-stanford.edu* Information on gene function, ORF homology and their presumed functions was downloaded 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 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 ORFs and intergenic regions is lower than the total length of genome).

All software used in the work has been written by one of authors (M.R.D.)

To represent graphically the asymmetry in nucleotide composition of three different codon positions we have used a modification of Berthelsen walk (Berthelsen et al., 1992) – 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 next 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 the walker is associated with a unit shift in two-dimensional space with the bases (A,T) and (G,C) depending on the type of nucleotide visited. The shifts of our DNA walker are: one unit up for G, right for A, down for C and left for T. Hence, each DNA walk represents a "history" of nucleotide composition of the first, the second and the third position of codons along a DNA sequence. The three walks together have been called a spider and a single walk has been called a spider leg.

We have estimated:
- the co-ordinates (x, y) of the ends of spider legs; x = [A–T] and y = [G–C]; the brackets denote number of nucleotides.
- the length of the vectors (L_1, L_2, L_3) starting the origin and ending at the end of spider legs representing the first, the second and the third positions in codons respectively,
- the angle between each vector and the X-axis (A_1, A_2, A_3 for respective leg).

Results and discussion

In Table 1, the relative frequencies of nucleotide in the three positions of codons for all ORFs longer than 100 codons are presented. In the same table the resulting asymmetry in sense/antisense strands of coding sequences has been shown. In the first codon position G is used more frequently than C and again A more frequently than T. There is no such a distinct difference in the third positions. The asymmetry in the first two positions is much more distinct than accumulation of GC pairs in coding sequences, which was observed several times for yeast genome (Bernardi 1993, Dujon 1994). The effect of the asymmetry in occupation of codon positions by different nucleotides is very well seen in Fig. 1 where we have presented the results of a DNA walk in two-dimensional space with the bases (A,T) on the opposite ends of one axis and (G,C) on the other to represent DNA sequences. The example in Fig. 1 corresponds to a gene 841 codons long with known protein coding function.
The spider in Fig. 1 is typical according to the averaged data shown in Tab. 1 and 2. Spiders representing coding ORFs have the first legs in the first quarter of the plot, the second legs in the fourth quarter and the third legs

Table 1. Relations between frequencies of nucleotide occurrence in three positions in triplets of different classes of sequences.

	GENES	ORFs	INTERGENIC sequences
number	2341	7440	5137
total number of nucleotides	4011453	9305301	4953183
A1/T1	1.527	1.464	1.007
A2/T2	1.280	1.198	1.005
A3/T3	0.888	0.904	1.008
G1/C1	1.822	1.721	1.012
G2/C2	0.616	0.638	1.008
G3/C3	0.944	0.981	1.005
(G1+C1)-fraction	0.449	0.442	0.366
(G2+C2)-fraction	0.367	0.368	0.368
(G3+C3)-fraction	0.376	0.381	0.369
(G+C) *total*-fraction	0.397	0.397	0.368

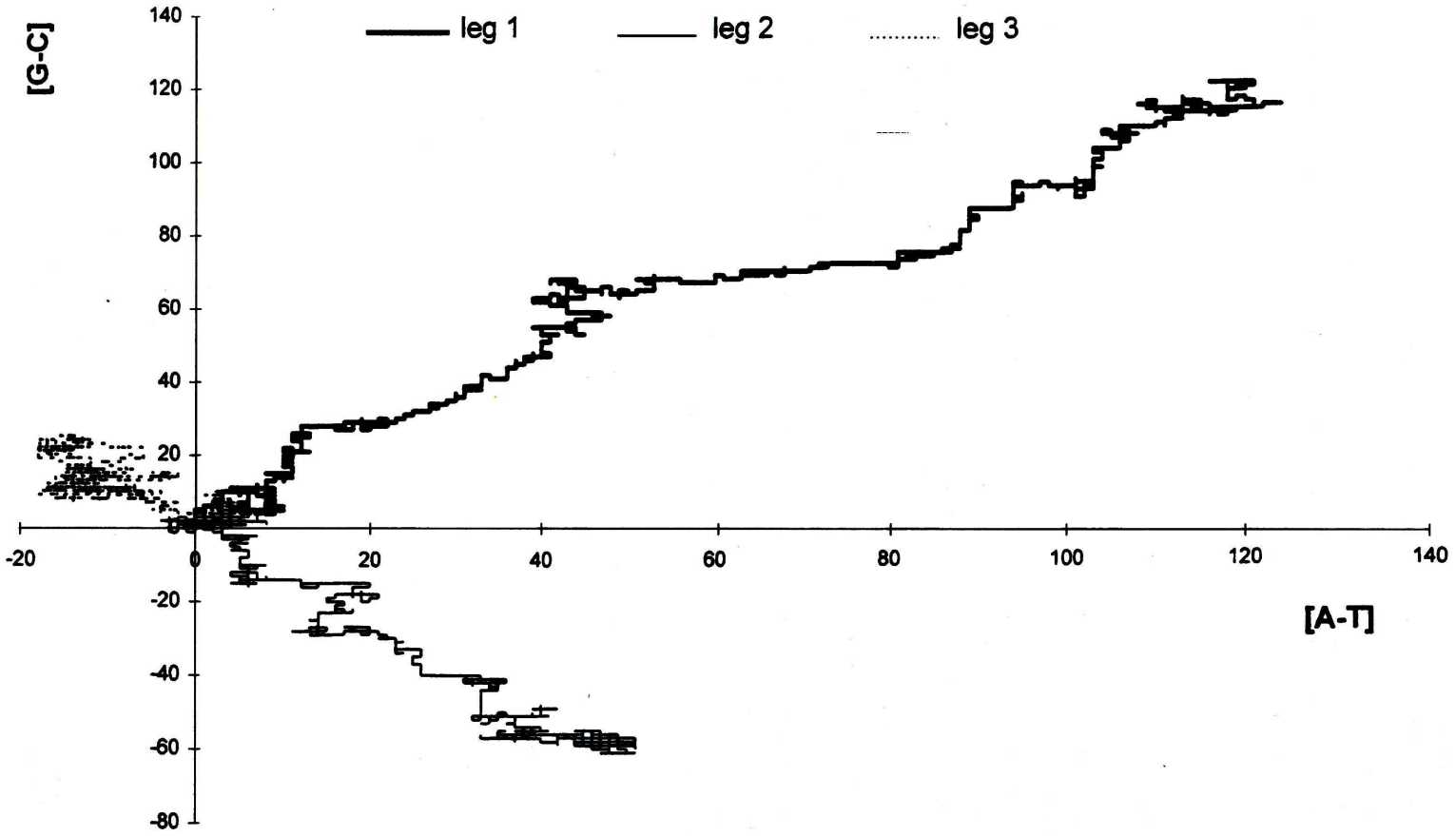

Fig. 1. The example of spider for coding sequence of 841 codons long. Numbers 1, 2, 3 denotes DNA walks for the first, the second and the third positions of codons respectively.

resembling Brownian walk. In case of a intergenic sequence, all the three legs resemble the trace of Brownian motion (not shown).
In Tab. 2 and Fig. 2 we have presented data characterising sets of specific classes of sequences. In Fig. 2 the vectors representing legs of 300 randomly chosen genes with known function and 300 randomly chosen intergenic sequences are plotted. The length of vectors has been normalised by dividing it by the square root of the length (in triplets) of the corresponding sequence, assuming that the average displacement of a Brownian walker after N steps is proportional to the square root of N. Therefore we could expect that the average length of normalised vectors should be close to 1 if there are no correlations (no dependence on "history") in nucleotide appearance at specific codon positions.
Both data from Table 2 and plots in Fig. 2 show the differences in properties of spider legs for genes and intergenic sequences. The average lengths of vectors for 2205 analysed genes are much longer than 1 which means that appearance of nucleotide in particular positions in codons is highly correlated. It can be also seen in Fig. 2. On average, vectors for intergenic sequences are shorter than for genes which means that there are no such strong roles in nucleotide composition. Furthermore, no preferences in slope of vectors are observed in intergenic sequences, which is obvious, since there is no sense in considering a reading frame in these sequences. Nevertheless, we have found a lot of sequences in intergenic space which show strong rules in triplets composition which could be a remainder of duplicated genes.

Estimation of the total number of coding ORFs in the yeast genome

If any method of recognising the coding sequences in the genome can indicate all coding sequences, then the question arises: how many noncoding sequences are there in the genome which are wrongly indicated as coding? To answer the question we have used our method for estimation of the total number of coding ORFs in the yeast genome. We have represented

Table 2. The average length of vectors (L) of three "legs" for genes, questionable ORFs and intergenic sequences.

Position in codon	Genes		Questionable ORFs		Intergenic sequences	
	L	SDEV	L	SDEV	L	SDEV
1	3.41	1.75	1.65	1.15	1.59	1.11
2	2.69	1.78	2.06	1.24	1.60	1.14
3	1.82	1.20	1.86	2.24	1.60	1.15

Length of the vector $L = \sqrt{(A-T)^2 + (G-C)^2}$

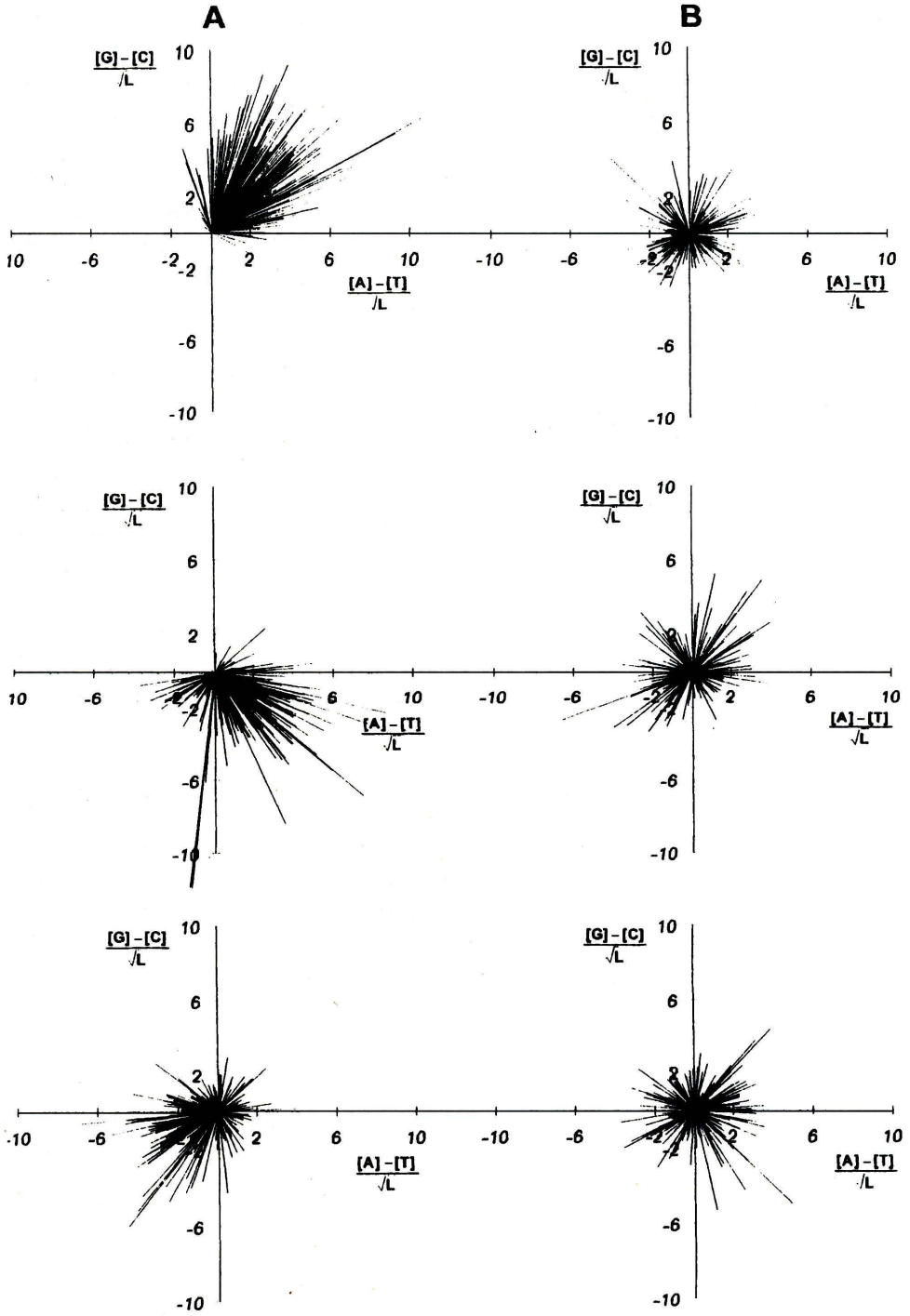

Fig. 2. The normalised vectors representing sets of; A – 300 coding sequences and B – 300 intergenic sequences, randomly chosen.

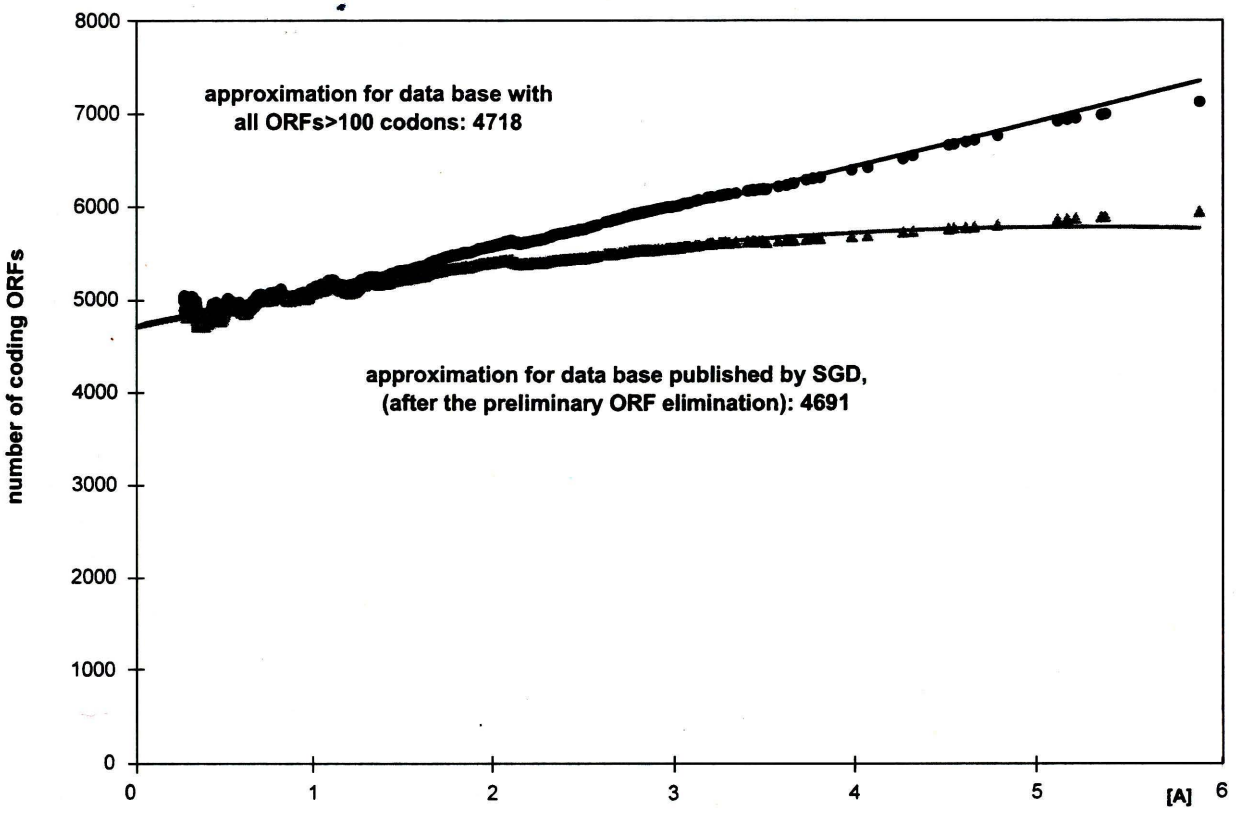

Fig. 3. Approximation of the total number of coding ORFs in yeast genome, in the set of all 7440 ORFs > 100 codons and in the set of 6095 ORFs published in SGD.

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

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

where:

- A_1 and A_2 are the values of angles in degrees of the first and second legs, respectively;
- $\bar{A}_1$ and $\bar{A}_2$ are the average values of slopes of the first and second legs, respectively for all ORFs with known function, and;
- SDEV is a standard deviation for angles of first and second legs for all ORFs with known function. Next we have counted for each ORF values:

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

where A_i represents the distance between the point representing a given sequence and the centre of gene distribution. To estimate the number of coding ORFs we have counted the number of genes inside the space determined by A_i, for different values of A_i and we have estimated the ratio of the number of genes inside the determined space to the number of genes outside the ellipse. Next we have counted the number of all examined ORFs inside the same surface and assuming that they are presumably coding, we have counted, from the ratio for genes, how many presumably coding ORFs should be expected outside the ellipse. We have got a plot shown in Fig. 4. The extrapolated line crosses the y-axis at 4718 for our data base and 4691 for the data base published by SGD. The obtained values are the numbers of ORFs which are expected to code for proteins, present in the analysed data bases. This estimation should be true if we assume that already known genes are a statistically representative sample of all genes. If unknown genes were different in nucleotide composition, the estimation would not be correct. Nevertheless, we should obtain comparable results using the ORFs referred in SGD. The number of these ORFs precisely located in our database sequences is 6085. The obtained number of coding sequences for SGD is only 27 lower than in our database. It seems to be in a very good agreement with data obtained for our database. The difference, if it is of any statistical significance, could mean that criteria used for preliminary elimination of ORFs in SGD had discarded some coding ORFs. We suppose that there are some shorter ORFs in pairs of overlapping ones which have been incorrectly eliminated. The estimated number of coding ORFs could be higher if we assume, that there are some coding ORFs shorter than 100 codons. Note that this number does not include most genes with introns (some should be exons are represented by ORFs laying inside them). These results seem to prove that there is no "mystery of orphans" described by Dujon (1996) and by Casari et al. (1996).

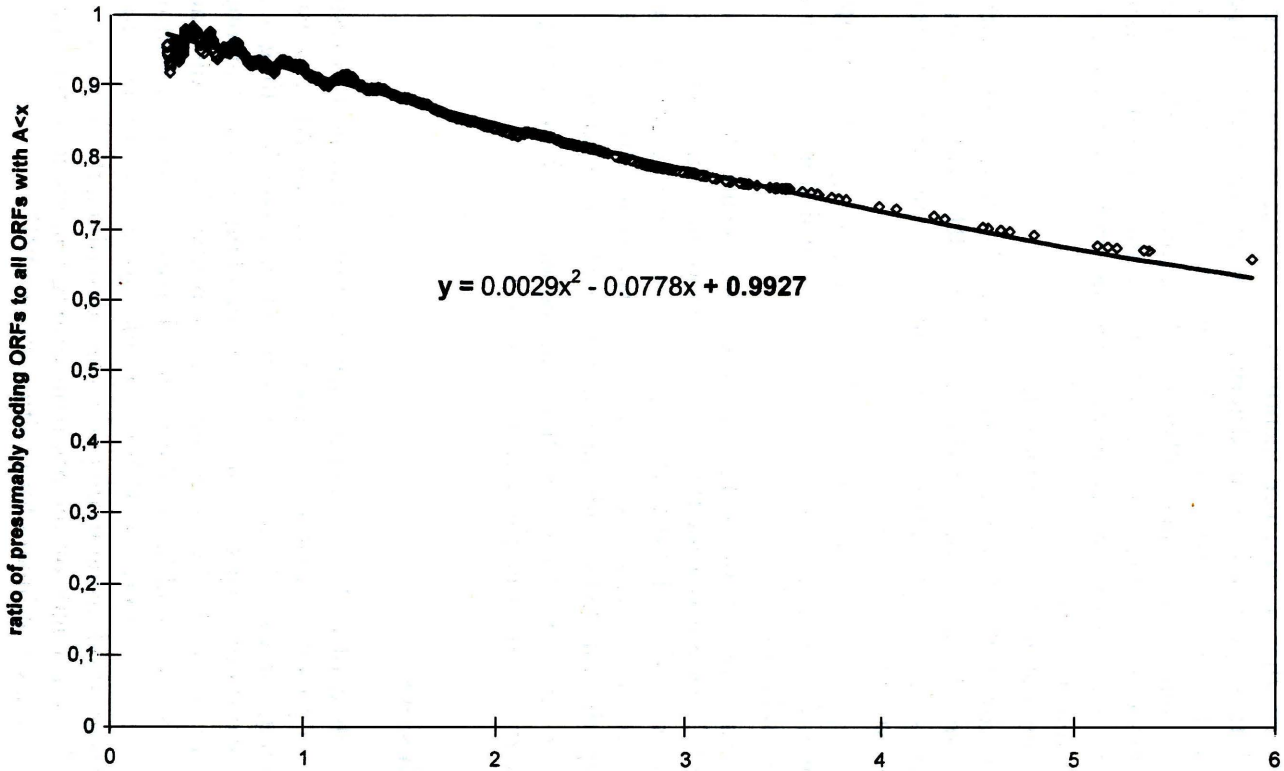

Fig. 4. The relation between *A* and the expected fraction of coding ORFs among all ORFs with distance to the centre of gene distribution smaller than *A*.

Discrimination of the noncoding ORFs

Our approximation of the number of coding ORFs enable us to evaluate the method as a tool for eliminating the noncoding ORFs from the set of presumably coding sequences. There are two questions which should be posed in evaluating the method of gene identifying: 1 – what is the probability that a coding sequence is discarded as noncoding, and 2 – what is the probability, that a sequence identified as coding is not coding. In our method, these probabilities are determined by the value *A* for a particular ORF. In Fig. 4 we have plotted *A* values versus the ratio between the expected number of coding ORFs and the number of ORFs found inside the distance *A* from the centre of gene distribution. When *A* is going to 0, the ratio is going to 1 which means that the probability that the ORF is coding approximates 1.
Evaluating the method we can say that:

- There are 338 ORFs in the space outside the "last" ORF with known function. That means, that there are 338 ORFs with a distance to the centre of gene distribution larger than the largest *A* found for any of 2205 known yeast genes.
- About 1300 ORFs are outside the space closing 99% of ORFs with known function. That means that coding probability for any of these ORFs is below 0.01.
- There are about 5700 ORFs inside the space closing 96% of genes. About 1700 ORFs stay outside this space. About 100 coding ORFs are expected in-between these 1700 ORFs. That means: if we accept the number of 5800 coding ORFs in the yeast genome (as assumed by SGD), we have to accept that all ORFs inside the space closing 96% of genes are coding(!).
- About 70% of all coding ORFs can be indicated as coding with the probability >0.9 that the decision is correct.

Conclusions

Any method of coding sequence identification could be too restrictive – there is a possibility that a lot of coding sequences would be considered as noncoding, or too liberal if it tries to encompass all coding sequences. Then, one should expect a fraction of noncoding sequences among coding ORFs. In our method we can estimate the probability of coding according to the *A* value. In addition, the method leaves the possibility to use other independent methods – CAI for example. We are also preparing a more complex method, using more parameters describing the sequence asymmetry such as a length of vectors representing spiders legs and correlations between legs co-ordinates for position inside ORF.

The method predicts the total number of coding ORFs significantly lower (~1100 ORFs less) than previously predicted. That seems to explain the so called "mystery of orphans" (Dujon, 1996, Casari et al., 1996). It is obvious, that a growing number of identified genes (ORFs with known function) should diminish the fraction of ORFs without known homology. For yeast genome quite an opposite phenomenon was observed – the fraction of ORFs without known function or homology has grown together with the number of sequenced DNA stretches. This effect could be expected only in two cases:

1. The set of already known genes is not a statistically representative sample of the total set of coding sequences of yeast genome, and/or;
2. There are much more noncoding ORFs in the set of presumably coding sequences than it was predicted.

Our approximations support this second possibility. Evaluating the method we have shown that decisions of discarding the ORFs from published SGD were usually correct. Only a small fraction of discarded ORFs (2%) should be considered as coding by our method. This loss of some coding sequences by preliminary elimination of ORFs before the publication of SGD probably is due to the arbitrary assumption that the shorter ORF of two overlapping ones should be discarded.

Acknowledgements

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

References

1. Bernardi, G: (1993) The isochore organization of the human genome and its evolutionary history – a review. Gene **135** (**1–2**): 57–66.
2. 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.
3. Casari, G.; de Druvar, A.; Sander, C.; Shneider, R. (1996) Bioinformatics and the discovery of gene function. Trends Genet. **12** (7): 244–255.
4. Cebrat, S.; Dudek, M. R. (1996a) Symmetry in chromosome fractal organization and DNA domain structure. In: Borchards, P.; Bubak, M.; Maksymowicz, A. (eds) Proceedings of the 8th Joint EPS-APS Int. Conference on Physics Computing '96. Academic Compuater Center, CYFRONET-KRAKÓW, pp 371–374.
5. Cebrat, S.; Dudek, M. R. (1996b) Generation of overlapping open reading frames. Trends Genet. **12** (**1**): 12.
6. Cebrat, S.; Dudek, M. R.; Rogowska, A. (1997) Asymmetry in nucleotide composition of sense and antisense strands as a parameter for discriminating open reading frames as protein coding sequences. J. Appl. Genet. 38 (**1**): 1–9.
7. Cebrat, S.; Mackiewicz, P.; Dudek, M. R. (1997) The role of the genetic code in generating new coding sequences inside existing genes. BioSystems, in press.
8. Dujon, B. and 106 coauthors (1994) Complete DNA sequence of yeast chromosome XI. Nature **369**: 371–378.
9. Dujon, B. (1996) The yeast genome project, what did we learn. Trends Genet. **12** (7): 263–270.
10. Fickett, J. W. (1996) Finding genes by computer: the state of the art. Trends Genet. **12** (**8**): 316–320.
11. 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.