ELSEVIER

BioSystems 45 (1998) 165–176

Bio Systems

The role of the genetic code in generating new coding sequences inside existing genes

Stanisław Cebrat [a,*], Paweł Mackiewicz [a], Mirosław R. Dudek [b]

[a] *Institute of Microbiology, Wrocław University, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland*
[b] *Institute of Theoretical Physics, Wrocław University, pl. Maxa Borna 9, 50–204 Wrocław, Poland*

Received 10 April 1997; received in revised form 13 October 1997; accepted 4 November 1997

Abstract

The genetic code has a very interesting property—it generates an open reading frame (ORF) inside a coding sequence, in a specific phase of the antisense strand with much higher probability than in the random DNA sequences. Furthermore, these antisense ORFs (A-ORFs) possess the same features as real genes—the asymmetry in the nucleotide composition at the first and second positions in codons. About two thirds of the 2997 overlapping ORFs in the yeast genome possess this feature. Thus, the question arises: has this feature of the genetic code been exploited in the evolution of genes? We have searched the FASTA data bases for homologies with the antisense translation products of a specific class of genes and we have found some sequences with relatively high homology. Many of them have scores which could be randomly found in the searched data bases with a probability lower than 10^{-6}. We conclude that some genes could arise by positioning a copy of the original gene under a promoter in the opposite direction in such a way that both, the original gene and its copy initially use the same nucleotides in the third, degenerated positions in codons.

Keywords: Antisense; Genetic code; Gene evolution; Overlapping reading frame; Yeast genome

1. Introduction

Analysing whole genomes, it is easy to note that long open reading frames (ORFs) located inside ORFs with known function or partially overlapping other ORFs, are more frequent than the predicted frequency of the generation of reading frames in the random DNA sequence (Yomo et al., 1992; Ikehara and Okazawa, 1993; Merino et al., 1994; Cebrat and Dudek, 1996). Yomo et al. (1992) considered even the generation of ORFs inside ORFs as a basic mechanism of gene evolution. Facchiano et al. (1993) and Facchiano (1995) found homology between hypothetical proteins coded in alternate frames ($+1$ and $+2$)

* Corresponding author.
E-mail: cebrat@angband.microb.uni.wroc.pl

0303-2647/98/$19.00
PII S0303-2647(97)00072-5

and some protein sequences deposited in the SWISS-PROT data bank. Boldogkoi et al. (1995) proposed a mechanism for the generation of reading frames inside genes. According to them, the accumulation of G and C bases in the silent positions of herpesvirus genes lowers the probability of stop codons generation in the frame shifted by two nucleotides because all stops start with T base. Cebrat and Dudek (1996) have shown that there is a specific bias in ORF generation inside genes. The most frequently generated ORFs are located in the antisense strand and overlap the gene in such a way that the third positions of gene codons overlap with the third positions of the overlapping ORF triplets (A3-ORF). Since this bias is a result of the occurrence of palindrome doublets in two stop codons, it could be predicted from the genetic code properties and a similar effect can be generated also in the random DNA sequence. Although the generation of overlapping ORFs inside genes is a property of the genetic code, it is not conclusive enough to assume that the mechanism has been exploited by evolution for gene generation.

As in eukaryotic genomes a large fraction of DNA does not code for proteins, in orderto resolve the problem, it is very important to distinguish between coding and noncoding sequences, as well as to indicate the phase in which the sequence is coding. Much software has been developed to recognise genes by computer analysis (Fickett, 1996). We have used the method utilising the strong asymmetry between sense and antisense strands in specific nucleotide positions in codons (Cebrat et al., 1997). This method has allowed us to show that the A3-ORFs share many features specific for coding sequences with genes.

In this paper we have analysed the putative coding properties of overlapping ORFs found in the yeast genome.

2. Data for analysis and methods

Sequences for analysis were downloaded on 23 September, 1996 from: genome-ftp.stanford.edu. Information on the gene function, ORF homology and/or their presumed functions was downloaded on 16 November, 1996 from: http://www.mips. biochem.mpg.de. We have analysed the set of all ORFs longer than 300 nucleotides (7440 ORFs), including all ORFs formerly discarded by SGD (*Saccharomyces* Genome Database programme).

The probability that a sequence has a protein coding function has been estimated using the method described previously (Cebrat et al., 1997). In the method we have analysed in the two dimension space, the displacement of a DNA walker which checked each position of codons separately. For the DNA walk we have used a modified method of Berthelsen et al. (1992). In fact, we have analysed the relations between the ratios (G–C)/(A–T) for the first and for the second codon positions for the sense strand of ORFs. To avoid the infinite values of the ratio (G–C)/(A–T), it was expressed in degrees as arcus tangent(G–C)/(A–T). The examples of the DNA walks performed for a yeast gene coding for neutral trehalase (YDR001c) are shown in Fig. 1. The three DNA walks performed for one sequence have been called a spider and the result of walk for one codon position—a spider leg.

Since a very characteristic distribution of the slopes of spider legs had been observed for coding sequences (especially for the first and second positions), we have used these parameters to discriminate coding and noncoding sequences. We have plotted the values of the slope of leg 1 versus the values of the slope of leg 2 for each examined sequence. This allows us to present data in the finite surface—a specific projection of a torus, where each sequence is represented by co-ordinates x and y being the slopes in degrees plotted in scale $\pm 180°$. The results are shown in the Fig. 2 where two plots are presented: (a) for 2205 ORFs with known functions, (b) for all overlapping ORFs. Each point represents one ORF with co-ordinates equal to arcus tangent(G–C)/(A–T) for the first positions in codons (x) and for the second positions in codons (y). For easier referring, we have divided the plot into 16 parts each described by letters A–D and numbers 1–4.

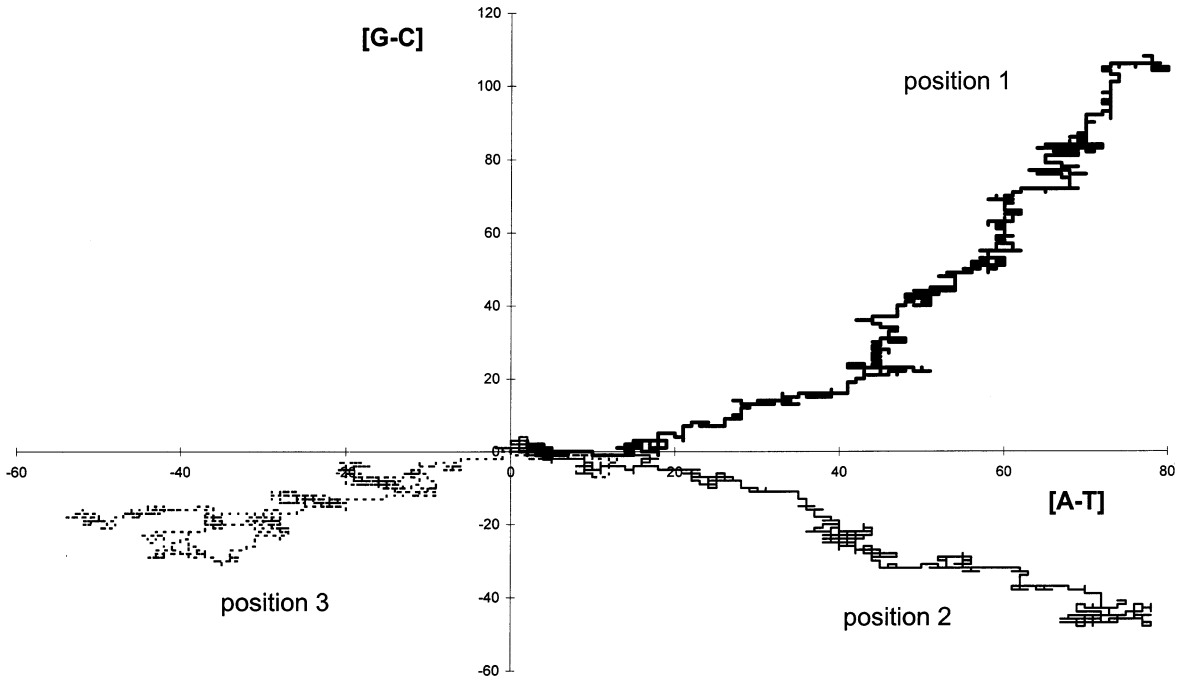


Fig. 1. DNA walk on a yeast gene coding for neutral trehalase (YDR001c); position 1, the walker visits every first nucleotide of codons and move a unit up if the nucleotide is G, down if it is C, right if it is A and left if it is T; position 2, the walker visits every second nucleotide of codons; position 3, the walker visits every third nucleotide of codons.

3. Statistics of overlapping ORFs

The total number of overlapping ORFs > 100 codons in the yeast genome is 2997. Some ORFs belong to clusters of more than two overlapping ORFs. More than 75% of genes coding for proteins are located in square C3 of Fig. 2(a). These typical genes are represented by an example presented in Fig. 1. There are also some ORFs with known function located in squares D3 and D4. Comparing both plots shown in Fig. 2, it is easy to note that there are relatively many ORFs located in squares D3 and D4 in the set of overlapping ORFs. To understand this phenomenon, ORFs should be divided into subclasses depending on the relations between two ORFs in the overlapping pair. In Table 1, we have presented the numbers of pairs of overlapping ORFs in the yeast genome and the numbers of triplets in overlaps, depending on the relation between these ORFs. About two thirds of all nucleotides belonging to more than one ORF has been found in ORFs from opposite strands, overlapping with the third positions of codons (we have called them A3-ORFs). This tenfold surplus of overlapping codons in one class of ORFs can be explained by the properties of the genetic code. Two stop codons (TAG and TAA) have the first two nucleotides in palindromes and if any pyrimidine is placed upstream of these codons—it generates a stop codon in the related phase of the opposite strand (for phase 1 in this relation the respective phase is the phase 6, and for phases 2 and 3 in the same relations the respective ones are phases 4 and 5). Since there are no stops inside an ORF (by definition), the frequency of stop codons in the antisense strand drops significantly. In addition, start codon—ATG—has the same property as amber and ochre codon but inversely to stop codons, there are many methionine codons generating starts in the overlapping A3-ORFs. These two properties of the genetic code are responsible for the generation of ORFs inside ORFs in these specifically related phases. In Table 2, the results

of scanning of all ORFs (1266) found in the first phases of all yeast chromosomes are presented. The scanning was performed on the five remaining phases. The most dramatic differences in the stop codons occurrences have been found for A3-ORFs. Furthermore, in this class of overlapped ORFs, 94 394 codons were found in overlapping regions of Mummy/Baby pairs, ('Mummy' is a larger ORF totally nesting a smaller one—a 'Baby' ORF) and only 1302 codons were found inside Mummy/Baby overlapped with positions 1/3 (+ 2 phase) in the same strand. The ORFs overlapping with positions 1/3 are expected to be an overwhelming class according to the hypothesis of Boldogkoi et al. (1995). Analysing the results presented in Tables 1 and 2, it could be concluded that the bias in the frequency of different classes of overlapping ORFs is fully explained by the properties of coding sequences and the genetic code;

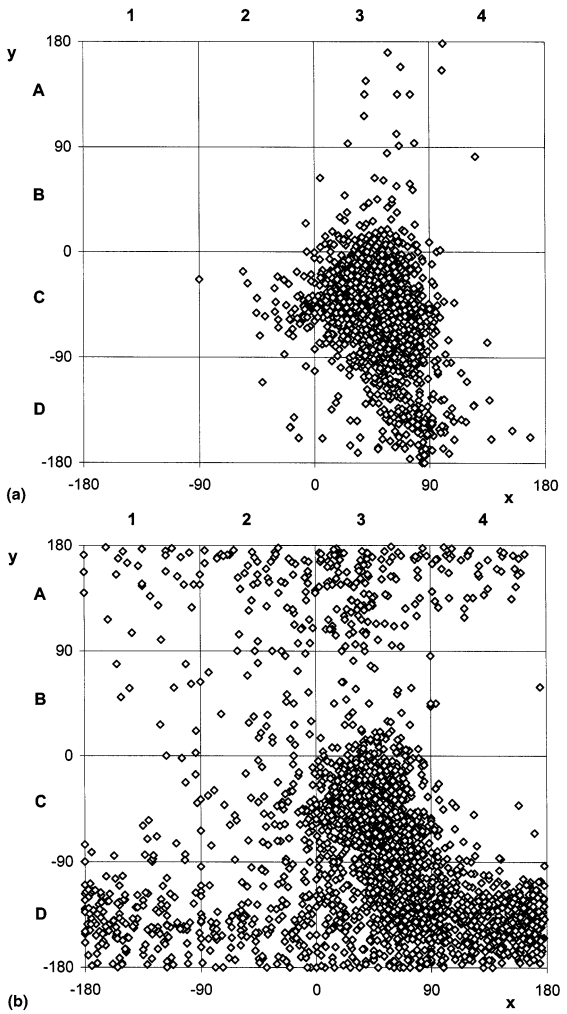


Fig. 2. The distribution of ORFs on the torus surface. (a) ORFs with known function, (b) all overlapping ORFs in yeast genome. Each point corresponds to one ORF. Co-ordinates: x = the slope of the vector for a DNA walk performed for the first positions of ORF codons and y = the slope of the vector for a DNA walk performed for the second positions.

- the frequency of generating the amber and ochre stop codons inside ORFs in the reading frame from the opposite strand, overlapping with the third positions of codons drops to the level below half of that frequency for any other reading frame. The average distance between stops is almost twice as big as for other reading frames, and 55% longer than the average distance in the whole yeast genome. Then, the average length of ORFs inside ORFs in such a phase relation is longer, which has also been found for random DNA sequences generated by computer (unpublished data).
- the most frequent nucleotides in the first positions of coding sequences are A and G, in the second position, A and C and in the third positions, T. Then, when a reading frame is shifted by two nucleotides in the same strand, the frequency of starts drops significantly and the frequency of stops grows to a very high level, resulting in an average distance of 11 codons between stops. That is why there is a very small number of ORFs of this class, and that is why they usually begin outside the overlapped ORFs and the overlapping fragments are relatively short.

4. Properties of overlapping ORFs

According to our estimations, the total number of coding ORFs in the set of all overlapping ORFs is about 1380. If we assume that usually only one ORF of the pair is coding, then in about 90% of pairs, one ORF is coding and the other one reflects its nucleotide composition. That depends on the phase relation between coding and

Table 1
Numbers of pairs of overlapping reading frames in the yeast genome according to the relations between the phases in which the ORFs occur

		Strand relations				
		The opposite strands			The same strand	
		Stop to stop → ←	Start to start → ←	Mummy/baby → ←	Start to stop → ←	Mummy/baby → ←
Overlapping Codon positions	Sum	155 *14417*	240 *22260*	939 *122128*	115 *7244*	120 *14860*
3/3	993 *121840*	84 *9228*	180 *18218*	729 *94394*	—	—
2/2	223 *24662*	43 *3378*	39 *2765*	141 *18519*	—	—
1/1	118 *12303*	28 *1811*	21 *1277*	69 *9215*	—	—
1/2	157 *15943*	—	—	—	49 *2385*	108 *13558*
1/3	78 *6261*	—	—	—	66 *4959*	12 *1302*

noncoding ORFs. Since 75% of coding ORFs are located in square C3 and about 2/3 of all overlapping ORFs overlap with the third positions, the plot seen in Fig. 2(b) should reflect these quantity relations.

The diagram in Fig. 3 shows the relations between codon positions in ORFs overlapping with the third positions. The most characteristic feature for coding ORFs is $G/C > 1$ for the first codon positions and $G/C < 1$ for the second codon positions. More than 90% of all known yeast genes conform to these rules (universal not only for yeast genes—our unpublished data). A3-ORFs also conform to these rules(!)—the first codon positions are rich in guanine and the second ones are rich in cytosine. In Fig. 4(a) we have presented a spider made for a sequence complementary to the gene shown in Fig. 1. Note that the shape of this spider looks like a spider made for a gene located in square D4 (Fig. 4(b)). The differences in the third positions do not seem to be important. Since this position is degenerated and because the changes in this position are neutral or 'close to neutral' (Kimura, 1986), they can be modified by the mutation pressure (Sueoka, 1988). Both spiders, the one made for a gene laying in D3 sector (Fig. 4(c)) and the other one, made for its antisense sequence overlapping with the third codon positions (Fig. 4(d)), lay in the same part of the plot in Fig. 2, both confirming the same coding rules.

The phenomenon of overlapping ORFs has three striking features:

- The occurrence of A3-ORFs is about ten times higher than any other class of overlapping ORFs.
- A3-ORFs share the features of nucleotide composition of codons with a specific group of genes (located in the squares D3 and D4).
- The number of ORFs > 100 codons of this type in the yeast genome is much higher than the number of ORFs randomly generated in the stochastic DNA sequence of the size of the yeast genome.

All these features are the implication of the genetic code properties.

Table 2
The occurrence of start and stop codons 'not in frame' inside the ORFs found in the first phase of all yeast chromosomes

Codons	No. codons in all ORFs of phase 1 scanned in phase						Whole phase 1[a]
	1	2	3	4	5	6	
ATG	11 255	17 373	3832	7254	9063	10 116	74 113
TAG	283	9395	7244	7061	11 371	3511	51 905
TAA	566	11 750	17 358	13 586	11 779	5745	88 138
TGA	417	13 042	21 588	10 576	11 731	9451	80 945
Sum of stops	1266	34 187	46 190	31 223	34 881	18 707	220 988
Number of codons per one stop[b]	416	15	11	17	15	28	18

[a] All triplets in the first phases of 16 yeast chromsomes.
[b] For the first phase this number is equal to the average length of ORFs; for the rest of phases, it is the total number of codons in the ORFs of the first phase divided by the number of stop codons in the scanned phase.

5. Homology for hypothetical proteins coded by antisense

If these properties of genetic code had ever been exploited in the generation of new coding sequences, it should be possible to find the traces of such a process now.

The generation of a new coding sequence from the existing one could be possible after duplicating the gene and locating it under a promoter in the inverse direction. The sequence of aminoacids in the new peptide could be quite different than that of the original gene. Let us assume that this mechanism has generated some genes. If the gene located in the square C3 (75% of all identified genes in yeast) had undergone such a process, the resulting coding sequence should be located somewhere in square D4. However, if the gene has been located in square D3, the resulting coding sequence should stay in the same square. If it is true, it should be possible to find the homology between theoretical peptide sequences coded by antisense strands of genes located in squares D3 and D4 and other known genes. To find such homology, we have translated the antisense strands of about 200 genes located in these regions (D3, D4) in the related phase and have performed a FASTA search versus full data bases for homology with these theoretical peptide sequences (Pearson and Lipman, 1988). The eventually generated stop codons were translated to tyrosine because stop codons with the first two nucleotides in palindromes need only one substitution in the third position to be changed into tyrosine codon. To evaluate the returned homologies, we have performed the same searches for 'antisense peptide' read in the other two reading frames, (overlapping with the first and second codon positions—A1-ORFs and A2-ORFs, respectively). The results of searches are shown in Fig. 5(a) and (b). In Fig. 5(a) we have presented 30 best 'optimised scores' returned by FASTA, for sequences translated from each phase of antisense sequence. Putative proteins coded by ORFs inside duplicated sequences belong to the first group, with the highest homology. For example sequences laying in genes FLO1, FLO5 and FLO9. Since FLO genes (responsible for flocculation) belong to one family and have the same origin, it is obvious that the overlapping ORFs found inside both genes are homologous. These sequences (overlapping ORFs) have been referred to in the *Saccharomyces* Genome Database project as the 'hypothetical proteins'. According to our estimation, the probability that these sequences are coding is very low. That is why we have excluded these sequences from any further consideration.

Proteins with very low, inconclusive homology to their antisense relatives, belong to the second, the most numerous group (not shown). But when we have discarded from the plot 5a all ORFs with homology to the so called 'hypothetical proteins', we have still got a difference between the level of homology for proteins coded by antisense overlapping with the third positions and the remaining two phases.

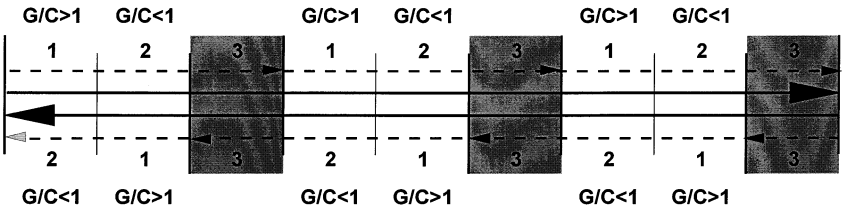


Fig. 3. Diagramm representing the relation between the sense and antisense strands overlapping with the third positions (A3-ORFs). Both, the first and the second positions of codons in both strands conform to the rule of G/C asymmetry.

We have analysed a few antisense sequences, for which the highest homogies have been found. One of these sequences is antisense for gene TIR2 (YDR010c) coding for the cold shock induced protein. Protein sequence coded by the antisense of this gene shows homology to *Paracoccus denitrificans* cytochrome bc1 with the optimised score 198 and 42% of aminoacid identity in a 150 overlap. The inverted search based on the sequence of the protein generated by the antisense of bc1 gene found the TIR2 protein with the optimised score 245. The nucleotide sequences were found with even higher scores and the expected probability of finding a sequence with a better score in the searched databases was of the order 10^{-6} (FASTA, Pearson and Lipman, 1988). Other examples are genes coding for H^+-transporting P-type ATPases (PMA1 and PMA2). Antisense proteins for both genes found Hansenula MrakII k9 killer toxin—resistance protein coded by the yeast HKR1 gene (YDR420w). The last protein is coded by three distinct regions (see Fig. 6). The internal region of the coding sequence looks like a coding sequence from square D4. The antisense protein for this region found four sequences with optimised homology scores above 200. In this case homologies between proteins were higher than between the nucleotide sequences. It could be concluded that the positive selection for the protein sequence was stronger than for the nucleotide sequence. It is even possible that there was a selection favouring fast divergence of the nucleotide sequences because of the sense/antisense hybridisation effect of the transcripts. The effect of the sense/antisense hybridisation forces the fast divergence of nucleotide sequences while the protein function found for the antisense was conserved. This effect should be expected when both sequences evolved in the same organism. That could be also an explanation why so many antisense sequences are found for genes in different organisms. Finding homologies with virus sequences suggests even that viruses could be the vectors for horizontal dissemination of the 'antisense sense'. In one case, a very conservative for many organisms DNA-directed RNA polymerase, we have even found the specific sense/antisense symmetry inside a protein, which could be a trace of duplication with inversion and a generation of a palindrome.

It seems that the differences in the results of searching for homology to the translation products of hypothetical antisense sequencess read in different phases, are significant. For antisense read in frames overlapping with the first or second positions, we have also found some homologous proteins but these proteins usually were monotonous, with long stretches of the same aminoacids or short repetitions like in a sex-determining *y* protein, keratin, fibrynogen or dynein. These cases need futher studies. Nevertheless, we would like to note that it could be difficult to prove that there are statistically significant differences in using different frames for the generation of new coding sequences.

The genetic code has another property—GCN box, CGN box UAY semibox and partially AUN code in both, sense and antisense strands (overlapping with the third positions), the same aminoacids resulting the same distances between them in the peptide sequence. These could also generate some correlations and homologies in the aminoacid sequences, especially in proteins with repetitions and even the palindromic peptides.

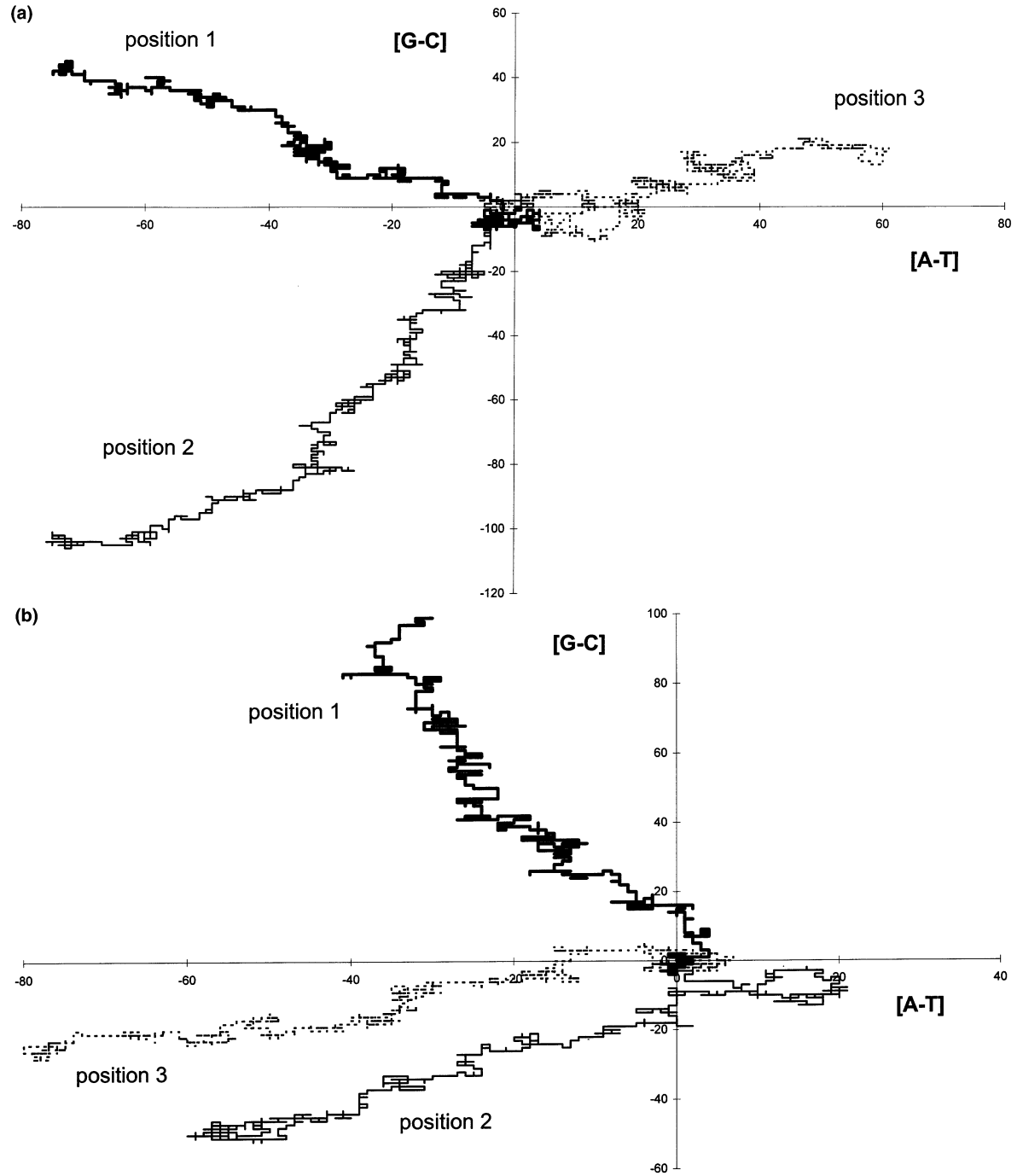
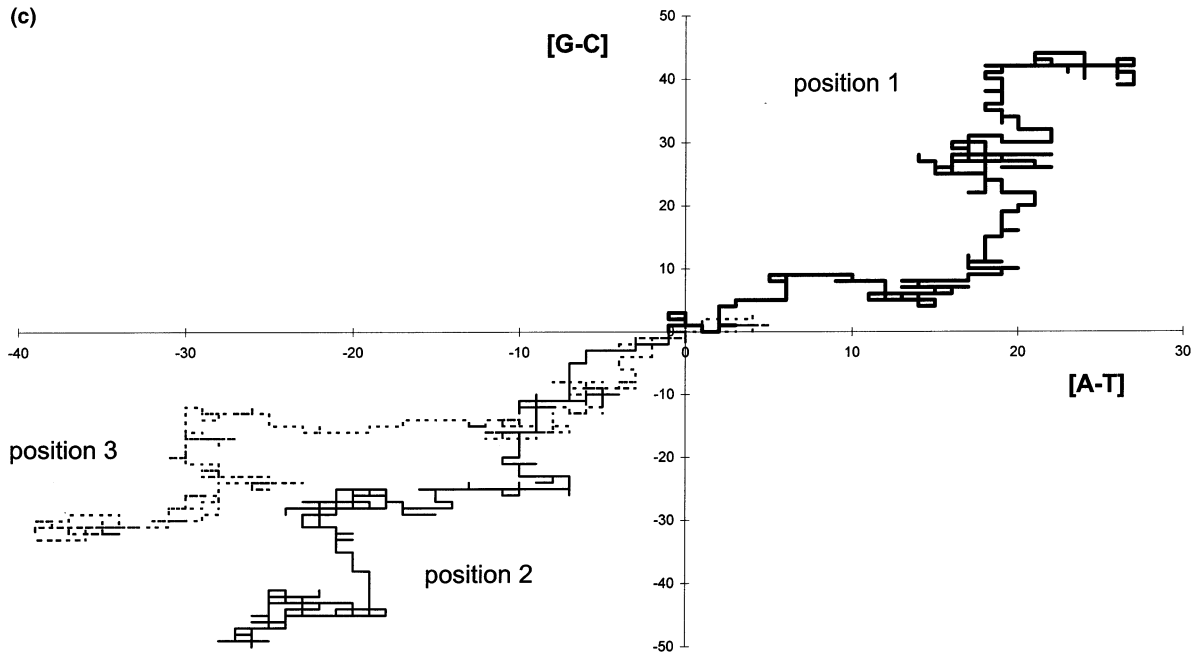


Fig. 4. Spiders for sense and antisense sequences. All antisense spiders are made for sequences overlapping with the third codon positions; (a) antisense for the gene coding for neutral trehalase (spider for the gene is presented in Fig. 1); (b) spider for the sense strand of the yeast uracil permease (YBR021w); (c) spider for the sporulation-specific wall maturation protein (YHR139c); (d) spider for the antisense of the gene presented in the plot c.



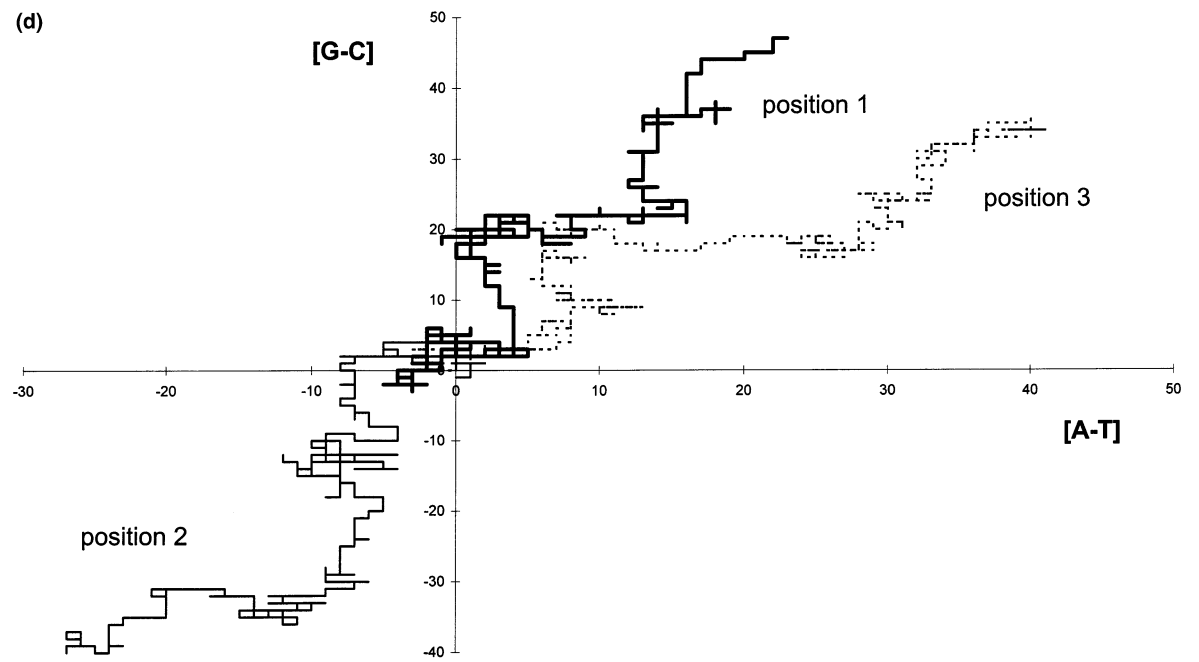


Fig. 4. (*Continued*)

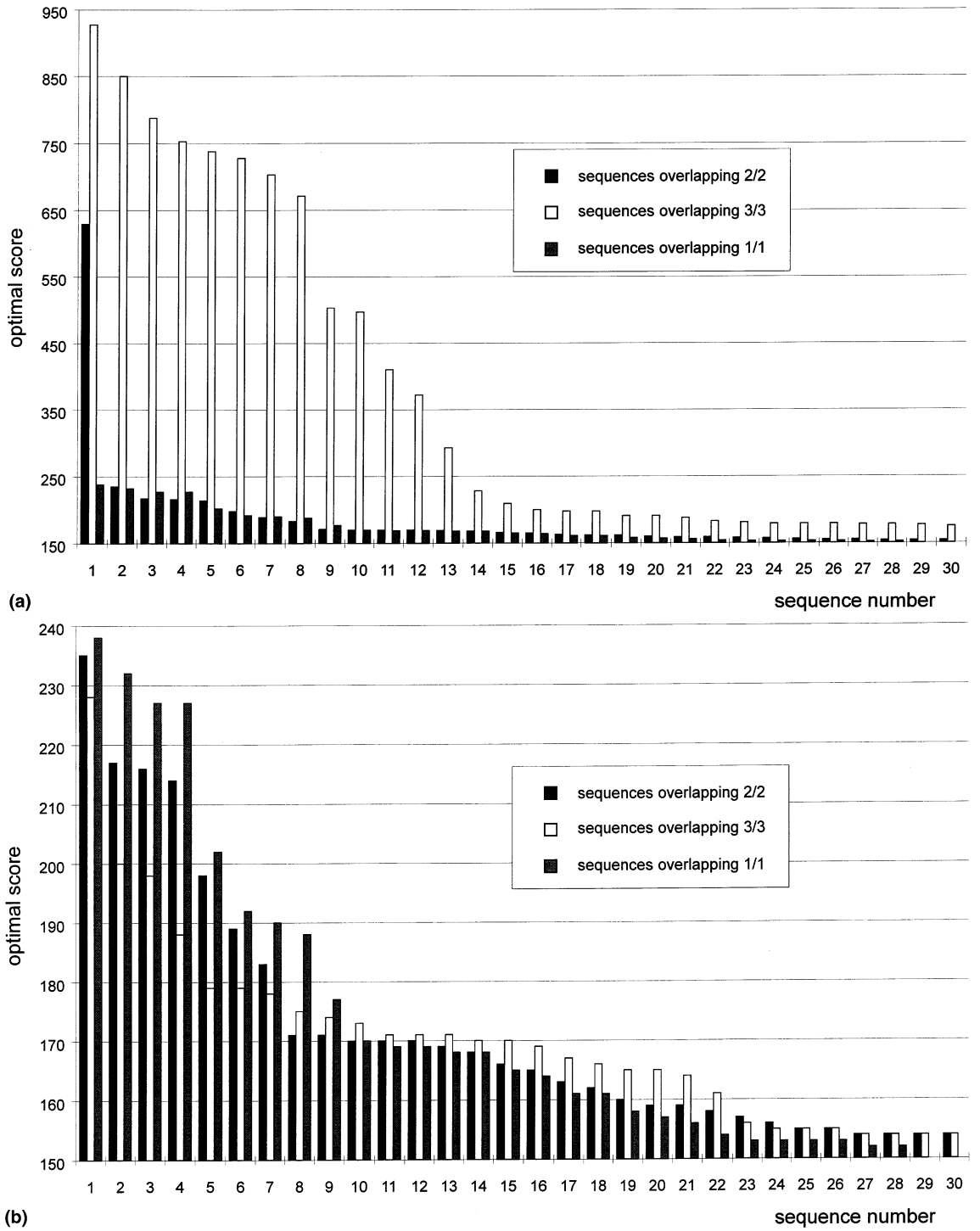


Fig. 5. Optimised scores returned by FASTA for homologies found for hypothetical proteins coded by antisense in different phases; (a) the best 30 scores in the set of 200 sequences sent to data bases; (b) the best thirty scores in the set of 200 sequences after discarding all the sequences referred in SGD as 'hypothetical proteins'. All scores in this diagram represent homologies to known genes.

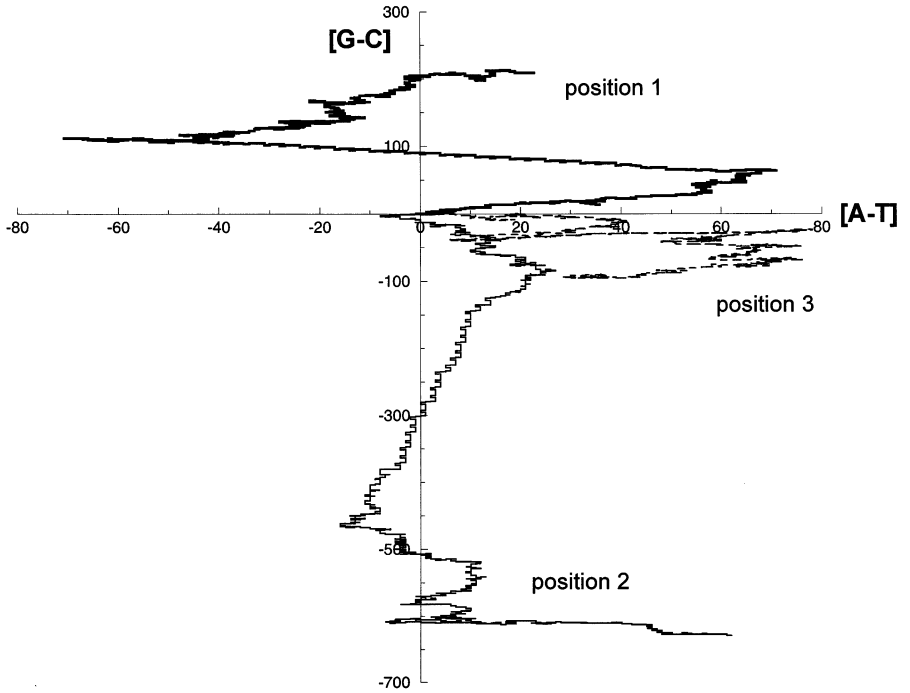


Fig. 6. A DNA walk on a yeast gene coding for Hansenula MrakII killer toxin-resistance protein. For explanation see description for Fig. 1.

6. Conclusions

The fact that two out of three stop codons possess a palindromic doublet in the first two positions enables the generation of long ORFs in one of three phases of the antisense strand of a coding sequence. This very class of ORFs shares some properties with coding sequences—the specific asymmetry in G/C contents of the first and second nucleotide positions of codons. It seems reasonable to assume that these ORFs are tested by evolution mechanisms for their usefulness in the fight for surviving. The homologies which we have found for antisense proteins suggest that this property of the genetic code has been exploited in gene evolution. Thus, the genetic code itself seems to be a potent evolution tool for the generation of new genes.

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.

Boldogkoi, Z., Murvai, J., Fodor, I., 1995. G and C accumulation at silent positions of codons produces additional ORFs. Trends Genet. 11, 125–126.

Cebrat, S., Dudek, M., 1996. Generation of overlapping open reading frames. Trends Genet. 12, 12.

Cebrat, S., Dudek, M.R., Mackiewicz, P., Kowalczuk, M., Fita, M., 1997. Asymmetry of coding versus noncoding strand in coding sequences of different genomes. Microb. Comp. Genomics. 2, 259–267.

Facchiano, A., Facchiano, F., van Renswoude, J., 1993. Divergent evolution may link human immunodefficiency virus GP41 to human CD4. J Mol. Evol. 36, 448–457.

Facchiano, A., 1995. Investigating hypothetical products from noncoding frames (HyPNoFs). J Mol. Evol. 40, 570–577.

Fickett, J.W., 1996. Finding genes by computer: the state of the art. Trends Genet. 12, 316–320.

Ikehara, K., Okazawa, E., 1993. Unusually biased nucleotide sequences on sense strands of Flavobacterium sp. genes

produce nonstop frames on the corresponding antisense strands. Nucleic Acid Res. 21, 2193–2199.

Kimura, M., 1986. DNA and the neutral theory. In: The evolution of DNA Sequences. The Royal Society, London.

Merino, E., Balbas, P., Puente, J.L., Bolivar, F., 1994. Antisense overlapping open reading frames in genes from bacteria to humans. Nucleic Acid Res. 22, 1903–1908.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.

Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85, 2653–2657.

Yomo, T., Urabe, I., Okada, H., 1992. No stop codons in the antisense strands of the genes for nylon oligomer degradation. Proc. Natl. Acad. Sci. USA 89, 3780–3784.