Nucleic Acids Research, 1994, Vol. 22, No. 10 ***1903–1908***

Antisense overlapping open reading frames in genes from bacteria to humans

Enrique Merino, Paulina Balbás, José Luis Puente and Francisco Bolívar*
Departamento de Biología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de Mexico, Apartado Postal 510-3, Cuernavaca, Mor. CP 62271, Mexico

Received December 1, 1993; Revised and Accepted April 13, 1994

ABSTRACT

Long Open Reading Frames (ORFs) in antisense DNA strands have been reported in the literature as being rare events. However, an extensive analysis of the GenBank database revealed that a substantial number of genes from several species contain an in-phase ORF in the antisense strand, that overlaps entirely the coding sequence of the sense strand, or even extends beyond. The findings described in this paper show that this is a frequent, non-random phenomenon, which is primarily dependent on codon usage, and to a lesser extent on gene size and GC content. Examination of the sequence database for several prokaryotic and eukaryotic organisms, demonstrates that coding sequences with in-phase, 100% overlapping antisense ORFs are present in every genome studied so far.

INTRODUCTION

Significantly long Antisense Overlapping Open Reading Frames (AO-ORFs) exist in the complementary DNA strand of genes from various organisms and, so far, they have been regarded as unusual events. Theoretical studies have suggested the possible coding capacity of such sequences (1–6) as well as the properties that should be exhibited by their corresponding putative proteins (2, 7–9). However, overlapping transcripts have been demonstrated for only a limited number of genes (10–16). These earlier studies had suggested that the AO-ORFs of protein-coding genes could have had a function other than serving as a conduit for semi-conservative DNA replication. However, these examples have not been analyzed in sufficient detail to allow the proposal of a cogent model. Therefore, we conducted a computer analysis of the nucleotide sequences reported in the GenBank database to determine whether long AO-ORFs are an intrinsic feature of all genomes.

EXPERIMENTAL PROCEDURES

Computer programming

All the programs developed for this study were written in Turbo Pascal language and ran on an Intel 386 based IBM-PC computer. The only source of information used in this study was the one recorded in Genbank release 74 (December 1992).

Calculation of the antisense overlapping percentage

We defined an AO-ORF of a gene as the largest reading frame between two in-phase termination codons TGA, TAG or TAA in the 5′→3′ direction of the antisense DNA strand. The sequences chosen were exclusively those that code for proteins, so extragenic regions and genes for tRNAs and rRNAs were not included in this study. When the antisense ORF extends beyond the initiation or termination triplets of the gene sequence, the length is limited by the coding region of the sense strand. A generic term called 'antisense overlapping percentage' was created in order to standarize the results plotted in the graphics. This term indicates the percent value of the largest AO-ORF for each gene, and it was calculated by the following equation:

Antisense overlapping percentage = (longest AO-ORF × 100)/gene-coding region

Selection of information from GenBank

For every analysis, gene entries with less than 300 nucleotides were eliminated because in most cases they corresponded to incomplete proteins (such as carboxy and amino termini, signal peptides, etc.). The information about *E.coli* genes was screened prior to the analysis in order to remove redundant nucleotide sequences. Organisms of the same genera were studied together in order to increase the size of the samples. The eukaryotic sequences studied were edited to remove sequences corresponding to introns.

Generation of hypothetical *E.coli* databases

The procedure followed was: (i) Each coding sequence in the GenBank was used to generate the corresponding amino acid sequence. (ii) This amino acid sequence was then used for the generation of seven hypothetical amino acid sequences on a gene-by-gene basis. (iii) The corresponding nucleotide sequences were derived from these putative proteins, identical in size but with different patterns in codon usage. (iv) The antisense strand for each of these hypothetical nucleotide sequences was deduced and analyzed. (v) The process was executed for every gene in the database, and this collection of putative genes is database number one. (vi) This cycle was repeated for one hundred times, and the average values obtained were plotted in the histogram as the number of genes against the antisense overlapping percentage.

*To whom correspondence should be addressed

Calculation of the codon usage, amino acid composition and GC content

The codon usage and aa composition for *E.coli* genes was calculated directly from the previously selected information recorded in the database. The GC content of genes was also calculated from the DNA sequences, therefore the values may differ from those reported experimentally.

Calculation of the RNY coefficient

The RNY index evaluation was previously described (17). A coefficient indicating the preferential usage of RNY codons versus non-RNY codons whenever a choice is possible (triplets for ile, thr, ser, gly, val, ala) was obtained. The bias due to either the absence of stop codons in the reading frame, or to the average aa composition of proteins was eliminated. Absence of an RNY pattern results in a value of 1; therefore, deviations from this figure indicate bias by codon preference.

RESULTS

We first examined the *E.coli* genome. A statistical analysis performed on its protein-coding nucleotide sequences confirmed previous observations (18) that the length of certain AO-ORFs was not random. The value of the overlapping percentage (see EXPERIMENTAL PROCEDURES) of the largest AO-ORF for every gene larger than 300 nucleotides was calculated and the results are shown in figure 1A. The observed trend corresponds to a Poisson distribution but, unexpectedly, 5% of the genes whose sequences are available possess a 100% AO-ORF. In many cases, the antisense ORF extends beyond the initiation or termination codon of the sense strand.

Control analysis were designed to verify if the results obtained were exclusively an effect of amino acid position within the peptide sequence, preferential codon usage, and/or amino acid composition. The amino acid sequence deduced from each *E.coli* gene in the GenBank was used to randomly generate seven different hypothetical nucleotide sequence databases. Permutation of the position of the original amino acids with their original triplets generated a Poisson distribution of the AO-ORFs (Fig. 1B) with a high resemblance to the data obtained from the GenBank and present in figure 1A. The number of genes with totally AO-ORFs remains constant as required. None of the other databases, either with codon usage bias (Figs. 1C, 1E and 1G) or without it (Figs. 1D, 1F and 1H) , equalled the frequency of 100% AO-ORFs encountered in *E.coli*, even though they were equal in size. Figure 1H represents a random universe, without codon usage bias and without amino acid composition bias; the only feature conserved was the length of each qene. As it can be seen, codon usage was found primarily responsible for the length of the AO-ORFs (Fig. 1B), whereas gene size and amino acid composition played a less important role. The statistical significance of these results was evaluated using the G-test (19) and the results obtained indicated that the distribution of the AO-ORFs was in fact a non-random phenomenon not due to a numerical artifact (data not shown).

It was also found that the AO-ORF pattern was associated with the reading frame of the sense strand. Comparison of the reading frames on both DNA strands confirmed that the 100% AO-ORFs are prevalently in-phase with the coding frame (Fig. 2A). Figures 2B and 2C show that the Poisson distribution obtained for the non-coding reading frames are similar to the pattern obtained for

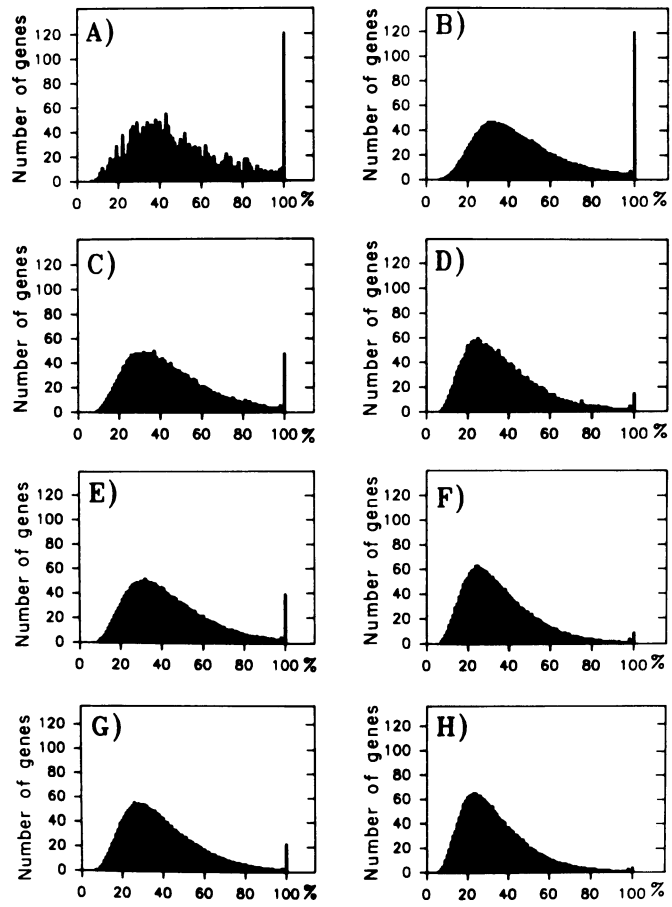

Figure 1. Analysis of the AO-ORFs in the *E.coli* genome. **A**, distribution of the AO-ORFs in genes from the GenBank database. The distributions B to H correspond to computer generated hypothetical databases; **B**, amino acids identical to those present in the actual proteins, with their corresponding actual triplets, in random positions. **C**, proteins with the same amino acid sequences as the real ones, but with codon choice changed according to the codon preference values calculated for *E.coli*. **D**, protein sequences identical to the real ones, but with random codon choice. **E**, amino acid choice according to the average amino acid composition in random positions, but codon choice according to codon preference. **F**, amino acid choice according to the average amino acid composition in random positions, but with a random codon choice. **G**, random amino acid composition, random amino acid position, using the *E.coli* preferential codon usage. **H**, random amino acid position, amino acid composition and codon usage.

sequences with random amino acid and codon composition. These results reflect the fact that fewer stop codons are always found in the in-phase antisense strand (20).

Computer programs were designed in order to scan for patterns in the nucleotidesequences that might influence the prevalence of AO-ORF. Results presented in figure 3A show that there is a direct correlation between the overlapping percentage and the GC content of the genome. This might be expected since in the TTA, CTA and TCA codons, which are the triplets that generate stop codons in the antisense strand, only two of the nine bases involved are G or C. Furthermore, the positional correlation between purines and pyrimidines was determined to assess whether triplets with the pattern RNY (21−25) correlated with the overlapping percentage. The algorithm has been described in detail elsewhere (17), and the results show that the RNY index is positively correlated to the overlapping percentage (Fig. 3B).

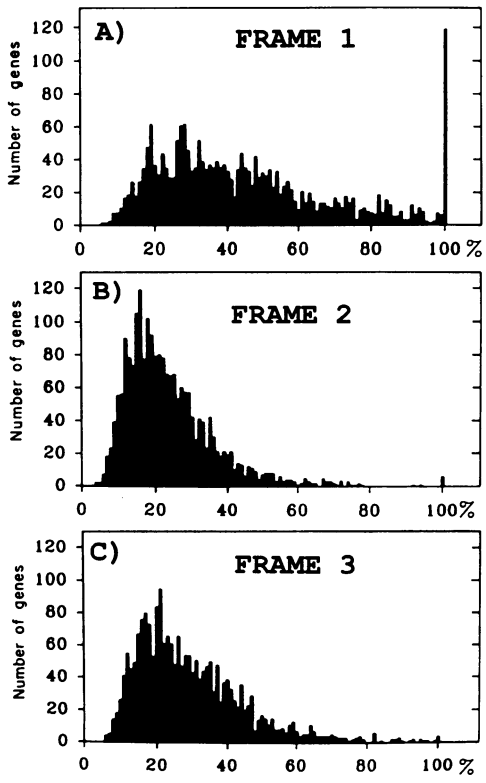

Figure 2. Analysis of the reading frames of the longest AO-ORFs in *E.coli* genes. **A**, in-phase, coding frame; **B** and **C**, non-coding reading frames.

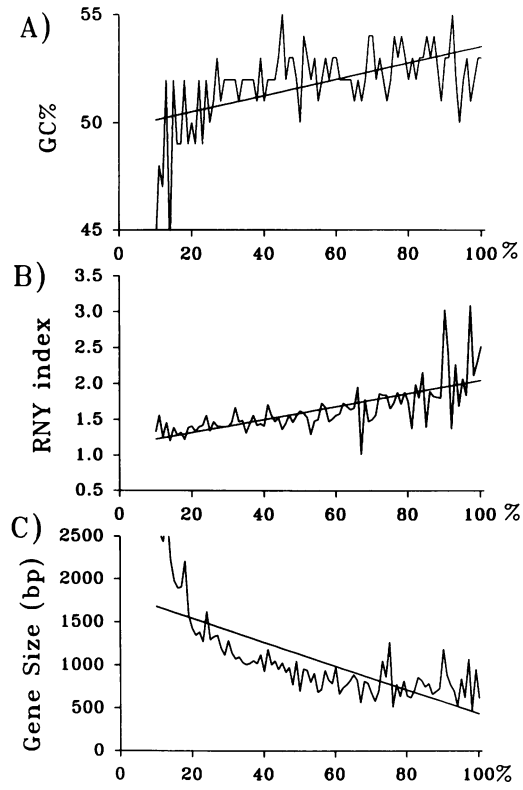

Figure 3. Correlation of the overlapping percentage for *E.coli* genes and: **A**, the average GC content; **B**, the average RNY content, and; **C** the average gene size.

When this analysis was performed on a gene by gene basis in the *E.coli* database, the average coefficient obtained was 1.6, indicating the existence of a predilection for codons with the RNY form. However, when the genes with totally AO-ORF were tested in identical conditions, a higher coefficient of 2.3 was obtained. On the other hand, the overlapping percentage shows an inverse correlation with the average size of the genes suggesting that the length of genes with 100% AO-ORFs might be biased towards smaller sizes. However, some long genes (the longest having 2856 bp) exhibited 100% AO-ORFs, indicating that this is not exclusively a size-dependent phenomenon (Fig. 3C).

This analysis was extended to the bacterial genera having 50 or more gene entries in the database. The results are plotted in figure 4; Table 1 shows the numerical values from this analysis. Figure 4A reveals that the GC content correlates positively with the overall overlapping percentages. When an individual species is studied, as shown for *E.coli* in figure 3A, this phenomenon may not be clearly noticed because no significant variation is observed in the GC content on a gene-by-gene basis. However, when comparing different genera, the importance of this parameter becomes evident. Organisms with genomes containing high GC content tend to have a large number of genes with 100% AO-ORFs (i.e., *Streptomyces* sp). The analysis of the RNY pattern in figure 4B, also shows a positive correlation with the overlapping percentage. As it can be seen, species such as *Azotobacter* have an atypically high RNY content, which correlates positively with a relatively high overlapping percentage. Nevertheless, some exceptions were found. Also, it is important to remark that all the organisms studied, except *Clostridium* sp, prefer the use of RNY triplets. As it can be seen in figure 4C, the average gene size shows a negative correlation to the average overlapping percentage, as seen with *E.coli*. The bias inflicted by these parameters has made it difficult to find an association between genes with 100 % AO-ORFs and specific interspecies gene families.

The protein-coding nucleotide sequences of several eukaryotic organisms were also studied. Eight organisms having a significant number of gene entries in the database were chosen. The results in figure 5 reveal that a significant number of genes with a 100% AO-ORF pattern are readily identifiable, although the overlapping percentages vary to some extent. Furthermore, the association found in prokaryotic genes between gene size, RNY index and GC content and overlapping percentage, is also present in eukaryotic genes, as it can be seen in Table 1.

DISCUSSION

It is a well established phenomenon that a nonrandom coding capacity exists in the complementary DNA strand of some protein encoding genes (1, 4, 5, 18, 20). This capacity is given by large AO-ORFs, and their frequency in *E.coli* is depicted in figure 1A. The data presented thereafter extended these observations by, (a) calculating the overlapping percentage of these ORFs, the effect of amino acid position, amino acid composition and codon usage in the appearence of this pattern, and (b) by analyzing some patterns exhibited by the antisense strand of the genes with totally AO-ORFs. Our analysis has demonstrated that this feature is a wide spread pattern encountered from bacteria to humans.

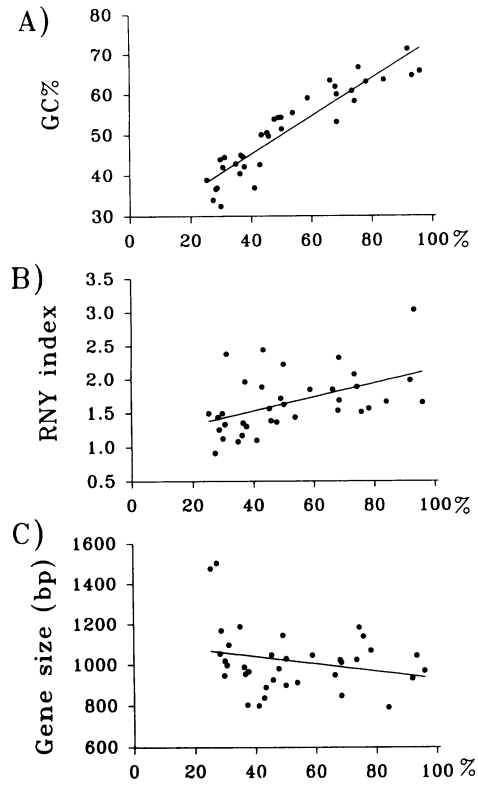

Figure 4. Correlation of the overlapping percentage of genes from different bacterial genus and: **A**, the average GC content; **B**, the average RNY content; and, **C**, the average gene size.

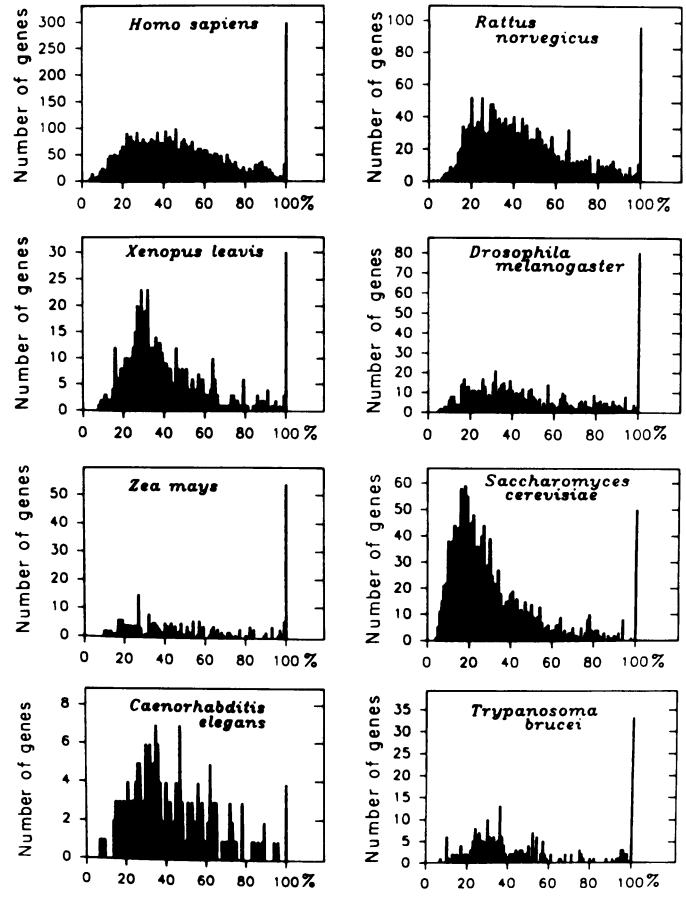

Figure 5. Analysis of the AO-ORFs in some eukaryotic species. The number of genes with AO-ORFs is plotted against the overlapping percentage.

None of the databases generated in this study, either with (Fig. 1C, 1E and 1G) or without (Figs. 1D, 1F and 1H) codon usage bias, equalled the frequency of 100% AO-ORFs encountered in *E.coli*, eventhough they were equal in size, amino acid composition and codon usage. Codon usage was found to be primarilly responsible for the AO-ORFs pattern, whereas, size and amino acid composition played a less important role. Corroboration that the largest AO-ORF in these genes were in-phase with the coding strand (Fig. 2), suggested that this pattern might be conditioned by the coding capacity of the sense strand. These results were consistent with those reported (20) which demonstrated that fewer stop codons may be found in the in-phase antisense strand. From our results, it is not possible to conclude whether these patterns have been maintained since an earlier evolutionary stage, or whether these AO-ORFs are the result of a selective process.

Examination of the GC content, RNY index and size of the *E.coli* genes in the database showed three important facts. First, that there was a direct correlation between the GC content and the overlapping percentage of the genes, although the variability of the GC content is only ± 10% (Fig. 3A). The correlation between GC content and frequency of genes with totally AO-ORFs increased dramatically when genes from individual genera were studied (figure 4A). This might be expected since in codons TTA, CTA and TCA, which are the triplets that generate stop codons in the antisense strand opposite to the stop codons in the 5′→3′ direction, only two of the 9 bases involved in the stop codons are GC.

Second, the existence of a preferential usage of the RNY codons for those amino acids where a choice between RNY and non-RNY triplets is possible (for the amino acids: ile, thr, ser, gly, val, ala). There is theoretical work supporting the view that RNY codons might have been preferentially used in a previous evolutionary stage, thus providing a primitive message in which translatable codons could mainly be found in one of the three possible reading frames. This is the basic statement of the Comma-less theory (17, 20–26), which has been debated both in conceptual and statistical terms. However, no consensus about its validity has been ascertained so far. If this hypothesis is correct, it is possible to speculate that in some stage in evolution, this RNY code was imprinted into double-stranded RNA, so these molecules would exhibit the same overlapped, in-phase RNY code pattern in the 5′→3′ direction. This fact could indicate that both strands could have directed the synthesis of different proteins, providing a clear advantage in the adaptative response to the environment. The discovery of present-day genes with the capacity to code for mRNAs in overlapping strands supports this view (10, 11, 14–16, 23). The direct correlation of RNY and the overlapping percentage in bacteria was presented in figure 4B and the numerical values for the RNY coefficients in Table 1 showed that there is important variation of this value among species. *Clostridium* is the only genera studied so far, that showed

Table 1. Properties of AO−ORFs in genes from bacteria to humans

Source	Number of reported genes	Average of the Overlapping percentage	Percentage of genes totally overlapped	Average gene size	Average GC content	Average RNY index
Prokaryotes:						
Rhodobacter	182	96.00	85.16	972	65.77	1.65
Azotobacter	92	93.31	76.09	1044	64.66	3.03
Streptomyces	318	91.92	75.79	934	71.32	1.98
Halobacterium	99	84.01	54.55	794	63.59	1.66
Alcaligenes	60	78.14	41.67	1071	63.17	1.56
Thermus	67	75.68	40.30	1140	66.74	1.51
Serratia	54	74.28	35.19	1185	58.35	1.88
Pseudomonas	534	73.46	40.64	1024	60.88	2.07
Neisseria	185	68.40	27.57	849	53.19	2.32
Rhyzobium	182	68.34	23.08	1012	60.05	1.68
Bradirhyzobium	55	67.94	23.64	1024	61.95	1.53
Mycobacterium	71	66.23	23.94	951	63.50	1.84
Klebsiella	167	58.73	9.58	1047	59.12	1.84
Agrobacterium	120	53.77	13.33	914	55.45	1.43
Escherichia	2423	50.09	5.30	1031	51.46	1.62
Synechococcus	115	49.96	8.70	901	54.32	2.22
Erwinia	93	49.05	3.23	1146	54.24	1.71
Salmonella	419	45.25	3.82	1049	50.57	1.56
Synechocystis	71	43.38	4.23	897	49.99	2.44
Anabaena	87	37.27	1.15	805	44.58	1.96
Yersinia	86	36.59	1.16	956	45.00	1.35
Chlamydia	107	36.29	3.74	989	40.38	1.17
Bacillus	969	34.84	1.65	1189	42.82	1.08
Staphylococcus	172	29.94	0.00	1020	32.44	1.12
Hemophilus	95	28.72	2.11	1169	36.84	1.25
Lactococcus	95	28.26	1.05	1055	36.61	1.43
Clostridium	155	27.35	1.94	1505	33.94	0.91
Streptococcus	227	25.25	0.44	1476	38.79	1.49
Eukaryotes:						
Zea mays	268	56.19	20.15	974	55.79	1.47
Caenorhabiditis elegans	76	52.06	3.95	1045	48.14	1.27
Drosophila melanogaster	724	50.59	11.33	1621	55.49	1.65
Trypanosoma brucei	237	49.57	13.92	1369	51.88	1.21
Homo sapiens	4776	48.90	6.18	1189	53.86	1.43
Rattus norvegicyus	1947	45.37	4.93	1283	52.51	1.47
Xenopus leavis	528	43.36	5.11	996	47.73	1.32
Sacharomyces cerevisiae	1473	31.52	3.33	1541	41.21	2.04

preference for non-RNY triplets. It has been discussed that this repeating RNY pattern might be either the residue of an antique gene structure or the outcome of natural selection (27). Although natural selection seemed to be the driving force that accounted for this phenomenon, our results about the high prevalence of in-phase, totally AO-ORFs indicate that natural selection may be acting not only on the coding strand, but also on the antisense strand of a considerable number of genes of every organism.

Third, as the overlapping percentage increased, gene size was biased towards smaller sizes (Fig. 3C). This result was expected because the probability of generating a stop codon in the antisense DNA strand increases as the gene size does. However, the 100% AO-ORF pattern was found not to be exclusive of small genes (data not shown). This inverse correlation was also noted when bacterial genera were analized in figure 4C.

Eukaryotic genes without their introns, also present this pattern, as it was shown in figure 5. The overlapping percentages vary to some extent, partly as a consequence of the genes that have been sequenced up to date, but other factors, not described up to date, may account for these differences. In spite of these variations, the results sustain our conclusion that this feature is inherent of coding sequences.

Our confirmation that the reading frame of the sense strand with the largest possible AO-ORF were identical, strengthens the hypothesis about the possible capacity of both strands of DNA (or RNA) to code for proteins in an earlier stage of life. Two independent lines of evidence support this view. The first is the theory of the coevolution of the genetic code and amino acid biosynthesis (28, 29). According to the author, not all the 20 aminoacids were originally present in the primaeval genetic code, but were accomodated in a series of stages until the actual genetic code was established. Remarkably, the earliest amino acids were those which nowadays retain prevalently the RNY form, and none of them produces a stop codon in the antisense strand. In accordance with this notion is the fact that the RNY message is particularly well preserved in genes which have been required to be heavily expressed throughout their evolutionary history. The ribosomal proteins and the translation-related protein factors

studied in detail reveal that, in fact, the RNY pattern in these proteins is highly preserved. The second supporting line of evidence is offered by a molecular recognition theory, which claims that AO-ORFs have the capacity to code for putative proteins with complementary hydropathy and structure. Therefore, these proteins that might have been originated from a single DNA molecule might also have the capacity of interaction, such as some eukaryotic hormone-receptor systems (2, 3, 7–9, 23). Further development and experimental verification of this theory are necessary before its validity can be ascertained, but if this mechanism is indeed present in such an evolved protein interaction system, it is suggestive that it might have been present throughout evolution.

The unexpectedly high number of genes containing an in-phase 100% AO-ORF, in organisms from bacteria to humans, may suggest that this is a non-fortuitous pattern intrinsic to coding sequences. From our data it cannot be established whether the AO-ORFs are related to an ancient genetic translation system, or whether they are the result of the action of a selective process. However, the previous belief that long AO-ORFs are oddities found in particular genes must be reevaluated, since they seem to be a widespread attribute of cell genomes.

ACKNOWLEDGEMENTS

We are indebted to Drs. C. Arias, D.L. Brutlag, E. Calva, G. Gosset, P. Lizardi, J. Merino, F. Recillas, R. S nchez, X. Sobern, V. Walbot and C. Yanofsky for comments on the manuscript.

REFERENCES

1. Alff-Steinberger,C. (1984) *Nucleic Acids Res.* **12**, 2235–2241.
2. Zull,J.E.and Smith,S.K. (1990) *Trends Biochem. Sci.* **15**, 57–2613.
3. Brentani, R. (1990) *Trends Biochem. Sci.* **15**, 463.
4. Casino,A., Cipollaro,M., Guerrini,A.M., Mastrocinque,G., Spena,A. and Scarlato,V. (1981) *Nucleic Acids Res.* **9**, 1499–1518.
5. Sharp,P.M. (1985) *Nucleic Acids Res.* **13**, 1389–1397
6. Yomo,T., Urabe,I. and Okada,H. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 3780–3784.
7. Blalock,J.E. (1990) *Trends Biotechnol.* **8**, 140–144
8. Goldstein,A. and Brutlag,D.L. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 42–45.
9. Sloostra,J. and Roubos,E. (1990) *Trends Biotechnol.* **8**, 279–281
10. Adelman,J.P., Bond,C.T., Douglass,J. and Herbert,E. (1987) *Science* **235**, 1514–1517.
11. Henikoff, S.,Keene,M.A., Fechtel,K. and Fristrom,J.W. (1986) *Cell* **44**, 33–42.
12. Kumamoto,C.A. and Nault,A.K. (1989) *Gene* **75**, 167–175
13. Miyajima, N., Horiuchi,R., Shibuya,Y., Fukushige,S., Matsubara,K., Toyoshima,K. and Yamamoto,T. (1990) *Cell* **57**:, 1–39.
14. Rak,B., Lusky,M. and Hable,M. (1982) *Nature* **297**, 124–128.
15. Simons,R.W. (1988) *Gene* **72**, 35–44.
16. Spencer,C.A., Geitz,R.D. and Hodgetts,R.B. (1986) *Nature* **322**, 279–281.
17. Merino,E. Balbás,P. and Bolivar,F. (1992) *Origins Life* **21**, 51–254.
18. Tramontano,A., Scarlato,V., Barni,N., Cipollaro,M., Franze,A., Macchiato,M.F. and Cascino,A. (1984) *Nucleic Acids Res.* **12**, 5049–5090.
19. Sokal,F.F. and Rohlf,F.J. *BIOMETRY,* H.W.Freeman and Co., San Francisco, 1981.
20. Staden,R. (1984) *Nucleic Acids Res.* **12**, 551–567.
21. Crick,F.H.C., Griffith,J.S. and Orgel,L.E. (1957) *Proc. Natl. Acad. Sci. USA* **43**, 416–421.
22. Crick,F.H.C., Brenner,S., Klug,A. and Pieczenik,G. (1976) *Origins Life* **7**, 389–397.
23. Keese,P.K. and Gibbs,A. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 9489–9493.
24. Shepherd,J.C.W. (1984) *Symp. Quant. Biol. CSH* **52**, 1099–1108.
25. Shepherd,J.C.W. (1984) *Trends Biochem.* Sci. **9**, 8–10.
26. Tyagi,S. (1981) *Origins Life* **11**, 343–351.
27. Wong,T.F. and Cedergren,R. (1986) *Eur. J. Biochem.* **159**, 175–180
28. Wong,T.F. (1981) *Trends Biochem. Sci.* **6**, 33–36.
29. Wong,T.F. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 1909–1912.