J Mol Evol (1997) 44:625–631

JOURNAL OF MOLECULAR EVOLUTION

On the Informational Content of Overlapping Genes in Prokaryotic and Eukaryotic Viruses

Angelo Pavesi,[1] Bettina De Iaco,[2] Maria Ilde Granero,[3] Alfredo Porati[3]

[1]Department of Evolutionary Biology, University of Parma, Viale delle Scienze, I-43100, Parma, Italy
[2]Institute of Biomathematics, University of Parma, Viale delle Scienze, I-43100, Parma, Italy
[3]Department of Physics, University of Parma, Via del Taglio, I-43100, Parma, Italy

Received: 28 October 1996 / Accepted: 29 January 1997

Abstract. In genetic language a peculiar arrangement of biological information is provided by overlapping genes in which the same region of DNA can code for functionally unrelated messages. In this work, the informational content of overlapping genes belonging to prokaryotic and eukaryotic viruses was analyzed. Using information theory indices, we identified in the regions of overlap a first pattern, exhibiting a more uniform base composition and more severe constraints in base ordering with respect to the nonoverlapping regions. This pattern was found to be peculiar to coliphage, avian hepatitis B virus, human lentivirus, and plant luteovirus families. A second pattern, characterized by the occurrence of similar compositional constraints in both types of coding regions, was found to be limited to plant tymoviruses. At the level of codon usage, a low degree of correlation between overlapping and nonoverlapping coding regions characterized the first pattern, whereas a close link was found in tymoviruses, indicating a fine adaptation of the overlapping frame to the original codon choice of the virus. As a result of codon usage correlation analysis, deductions concerning the origin and evolution of several overlapping frames were also proposed. Comparison of amino acid composition revealed an increased frequency of amino acid residues with a high level of degeneracy (arginine, leucine, and serine) in the proteins encoded by overlapping genes; this peculiar feature of overlapping genes can be viewed as a way with which they may expand their coding ability and gain new, specialized functions.

Key words: Overlapping genes — Information Theory — Codon usage — Amino acid composition

Introduction

A particular issue in the statistical analysis of genomic DNA sequences concerns the characterization of codes and semantic patterns in the genetic language (Trifonov 1989; Smith 1989). In this language, overlapping genes represent an unusual pattern, as two, or exceptionally three, out-of-phase reading frames may lie in a single nucleotide sequence. Such an arrangement, called ‘‘overprinting,’’ is frequent in viruses, where it probably evolved to increase the density of genetic information (Lamb and Horvath 1991). The first genes of this type were identified by Barrell and co-workers (1976) in the genome of ϕX174, a single-stranded DNA phage, and similar overlapping regions were later detected in many other genes belonging to DNA or RNA viruses of both prokaryotes and eukaryotes (Normark et al. 1983; Samuel 1989 and references therein). Translation of the different reading frames has been shown to be mediated by ribosomal frameshifting, which requires an upstream site of ribosomal slippage and a downstream stem-loop structure known as a ‘‘RNA pseudoknot’’ (Jacks et al. 1988; Wilson et al. 1988; Brierley et al. 1989). On the other

Correspondence to: A. Pavesi

hand, translation of multiple reading frames can occur simply by internal de novo initiation in an alternative frame and does not require ribosomal frameshifting (Atkins et al. 1979; Chang et al. 1989).

Originally developed to maximize the efficiency of transmission of electronic signals, information theory (Shannon and Weaver 1949) was later utilized to evaluate the complexity of DNA sequences (Gatlin 1968, 1972). In the past years, several papers have dealt with the connection between information theory and the analysis of overlapping coding regions (Yockey 1979; Granero-Porati et al. 1980; Smith and Waterman 1981). Other studies have addressed the problem of the evolution of overlapping arrangement (Miyata and Yasunaga 1978; Soeda and Maruyama 1982; Keese and Gibbs 1992) and the restrictions imposed on proteins encoded by the overlaid genetic messages (Sander and Schulz 1979; Smith and Waterman 1980).

Here, we present an analysis at different levels of complexity (divergence from randomness of mono- and dinucleotide composition, choice of synonymous codons, and frequency of occurrence of amino acid residues) of the informational content of overlapping genes. Using information theory indices and statistical methods of sequence analysis, the constraints acting on overlapping coding regions were quantitatively evaluated and compared to those occurring in the nonoverlapping regions belonging to the same viral genomes. Results obtained from the information theory approach and those derived from codon usage and amino-acid composition correlation analyses are discussed in terms of evolution of overlapping genes.

Methods

The nucleotide sequences of the complete genomes of three prokaryotic (ϕX174, G4, and α3 coliphages), five animal (two avian hepatitis B viruses and three different strains of lentivirus human immunodeficiency type 1), and four plant viruses (beet and barley luteoviruses, turnip and eggplant tymoviruses), all containing a large density of overlapping coding regions, were selected from the EMBL database (Rice et al. 1993). The genomic map of the five virus families is reported in Fig. 1.

Divergence from randomness at the level of mono- and dinucleotide composition was evaluated, respectively, by the informational indices D_1 and D_2 (Gatlin 1968):

$$D_1 = H_{\max} - H_1$$

where

$$H_1 = -\sum_{i=1}^{n} p_i \log_2 p_i$$

$$H_{\max} = \log_2 n$$

where n is equal to 4 (number of symbols in the genetic language) and p_i is the relative frequency of base ''i'' in a sequence under examination. Entropy H is measured in bits per symbol and its maximum value, $H_{\max}$, corresponds to a 25% frequency for each base equalling 2 bits/symbol. The D_1 index represents the divergence from maximum entropy due to constraints on mononucleotide composition.

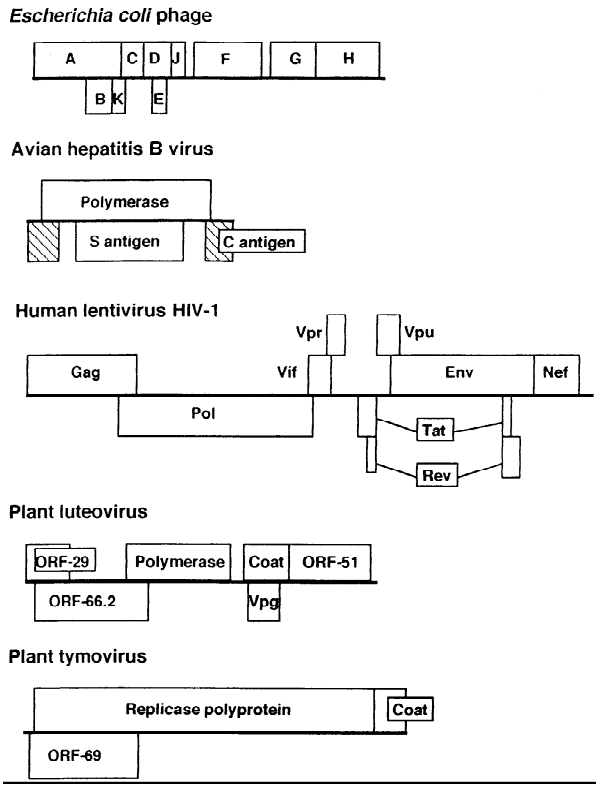

Fig. 1. Schematic organization of the genome in the five virus families under examination. The two *hatched boxes* represent the gene encoding for the C antigen, which, in the circular genome of avian hepatitis B virus, is continuous.

$$D_2 = H_2^{\text{ind}} - H_2^{\text{dep}}$$

where

$$H_2^{\text{ind}} = -\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \log_2 p_i p_j$$

$$H_2^{\text{dep}} = -\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} \log_2 p_{ij}$$

where p_{ij} is the relative frequency of dinucleotide ''ij'' in a sequence under examination. The absolute frequency of dinucleotides is calculated by moving along the sequence with steps corresponding to one nucleotide position. The D_2 index measures the divergence from an independent ordering of bases, thus accounting for the constraints acting on dinucleotide composition. Therefore, for a random sequence with no order at any level we would expect values of D_1 and D_2 indices nearly equal to zero.

The additional step of our analysis takes into account the comparison, at the level of both codon usage and amino-acid composition, between overlapping and nonoverlapping coding regions. Correlation analysis of the codon choice was carried out using both the Relative Synonymous Codon Usage (RSCU) index (Sharp and Li 1987) and the

Table 1. Values of information theory D_1 and D_2 indices (see Methods) in overlapping and nonoverlapping regions of the 12 viral genomes under examination

EMBL entry name	Description	D_1 Overlapping regions	D_1 Nonoverlapping regions	D_2 Overlapping regions	D_2 Nonoverlapping regions
BACALPHA	α3 coliphage	0.0042	0.0166	0.0347	0.0170
MIG4XX	G4 coliphage	0.0103	0.0122	0.0604	0.0288
PHIX174	ϕX174 coliphage	0.0080	0.0192	0.0465	0.0151
HBDGENM	Duck hepatitis B virus	0.0103	0.0325	0.0420	0.0252
HBHCG	Heron hepatitis B virus	0.0098	0.0224	0.0420	0.0185
HIVBRUCG	HIV-1 African isolate	0.0237	0.0596	0.0830	0.0488
HIVCAM1	HIV-1 European isolate	0.0252	0.0627	0.0682	0.0500
HIVNDK	HIV-1 African isolate	0.0234	0.0650	0.0740	0.0474
BYDCG	Barley luteovirus	0.0056	0.0062	0.0235	0.0174
BWYVFL1	Beet luteovirus	0.0041	0.0033	0.0418	0.0174
MTYRPVP	Eggplant tymovirus	0.0999	0.0715	0.0418	0.0421
TYMVCG	Turnip tymovirus	0.0861	0.0748	0.0212	0.0310

Pearson correlation coefficient *r*. The RSCU value for each of the 59 degenerate codons was calculated as follows:

$$\mathrm{RSCU} = (N_{\mathrm{codon}}/N_{\mathrm{aminoacid}})D$$

where N_{codon} is the total number of times a given codon is used in a given coding region, $N_{\mathrm{aminoacid}}$ is the absolute frequency for the amino acid specified by that codon and its synonyms, and D is the degeneracy of that amino acid (when all synonyms are used with equal frequencies, a RSCU value of 1 for each codon is expected). A set of RSCU values obtained from a given overlapping gene was then compared with that of the nonoverlapping regions of the same viral genome by means of the Pearson correlation coefficient (r), whose values, ranging from −1 to 1, reflect a completely discordant or concordant degree in the usage of synonymous codons, respectively. At the level of composition in amino acid residues, the degree of similarity between proteins encoded by overlapping and nonoverlapping genes was carried out by the chi-square test (Snedecor and Cochran 1967). Data were arranged in a 2 × 2 contingency table to identify amino acid residues whose frequency of occurrence in proteins encoded by overlapping genes is significantly higher than that observed in the nonoverlapping counterpart.

Results

From each of the 12 viral genomes under examination, two sets of data, including overlapping and nonoverlapping genes, respectively, were obtained and the constraints acting on base composition and base ordering were evaluated, respectively, by the D_1 and D_2 indices, whose values are reported in Table 1. In eight cases, including α3, G4, and ϕX174 coliphages (BACALPHA, MIG4XX, PHIX174), duck and heron hepatitis B viruses (HBDGENM, HBHCG), and strains of HIV-1 lentivirus family (HIVBRUCG, HIVCAM1, HIVNDK), the D_1 value of overlapping regions was found to be smaller than the D_1 value of the nonoverlapping part. In two cases, corresponding to barley and beet luteoviruses (BYDCG, BWYVFL1), the D_1 values appeared to be similar and near to zero. The exception was represented by the family of eggplant and turnip tymoviruses (TYMVCG, MTYRPVP), with a D_1 value of the overlapping sequences higher than the D_1 value of the nonoverlapping ones. Moreover, the D_1 value considerably different from zero obtained from both types of coding regions in plant tymoviruses reflects the highest divergence from a random base composition in the set of sequences considered in our analysis.

When analyzed with respect to the divergence from an independent ordering of bases (Table 1), all the overlapping sequences exhibited, with the exception of tymoviruses, a higher D_2 index value, as compared with the nonoverlapping counterpart. The graphical representation (Fig. 2) of the average values of the D_1 and D_2 indices, calculated by grouping the 12 viruses under examination in the five corresponding families (coliphage, hepatitis B virus, HIV-1 lentivirus, luteovirus, and tymovirus), led to the identification of two different informational patterns in the viral coding sequences. The first pattern is characterized by a clear tendency to possess, in the regions of overlap, a more uniform nucleotide composition (a lower D_1 value) and more severe constraints in base ordering (a higher D_2 value), with respect to the nonoverlapping regions lying in the same genome. It includes four of the five families considered in this study (coliphage, hepatitis B virus, HIV-1 lentivirus, and luteovirus). In the second pattern, which appears to be limited to the family of tymoviruses, both regions show, instead, similar compositional constraints, as evidenced by a slight variation of the corresponding D_2 values.

The additional step of our analysis concerned the relationship between the frequencies of synonymous codons in overlapping or nonoverlapping genes. For each of the five virus families, the nonoverlapping regions belonging to the corresponding members were combined into a single entity, while the two frames of each overlapping gene arrangement were considered as two sepa-

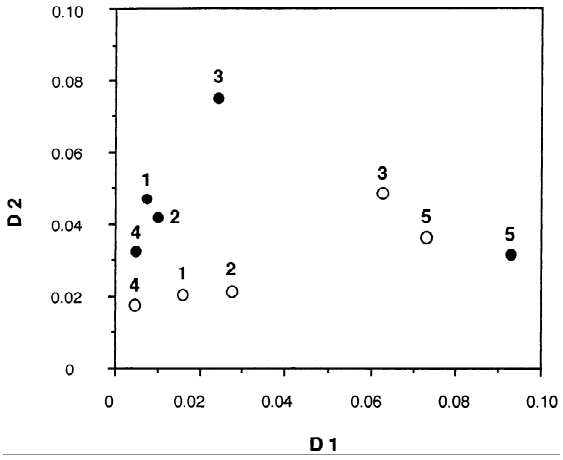

Fig. 2. Graphical representation of the average values of the D_1 and D_2 values (see Methods) in the overlapping and nonoverlapping regions belonging to the five virus families under examination. *Black circles* correspond to overlapping sequences, *empty circles* to nonoverlapping sequences. *1.* COLIPHAGE *2.* HEPATITIS B VIRUS *3.* HIV-1 LENTIVIRUS *4.* LUTEOVIRUS *5.* TYMOVIRUS.

rate entities. A total of 24 distinct overlapping reading frames (Table 2) were then characterized, thus increasing the statistical relevance of our analysis. For example, the nonoverlapping set of tymoviruses (see the genomic map of tymoviruses in Fig. 1) includes the coat gene and the nonoverlapping fraction of replicase gene of both eggplant and turnip virus. The overlapping regions of tymovirus family were considered, instead, as two distinct sets of data, the one including the overlapping region of replicase gene, the other the 69-kD protein gene.

The subsequent correlation analysis (Table 2) evidenced that six overlapping genes exhibited a choice of synonymous codons highly different from that occurring in the corresponding nonoverlapping genes. They include the A, C, E, and K genes of coliphage family, the Tat gene of lentivirus, and the Vpg gene of luteovirus, all exhibiting an *r* value near to zero. The highest degree of relationship was found in the overlapping genes encoding the replicase and the 69-kD protein of tymoviruses, as evidenced by an *r* value of 0.90 and 0.74, respectively. More generally, when the *r* mean values of the virus families were considered (Table 2) the overlapping regions related to the first informational pattern (coliphage, hepatitis B virus, HIV-1 lentivirus, and luteovirus families) exhibited a very low correlation with the usage of synonyms in the nonoverlapping counterpart, as documented by a range of variation from 0.24 in coliphage to 0.38 in hepatitis B virus. In contrast, a much higher relationship between overlapping and nonoverlapping regions (an *r* mean value of 0.82) was found in the family of tymoviruses, representing the alternative informational pattern.

The statistical analysis testing a difference in the composition of amino-acid residues between each of the 24 overlapping frame encoded proteins and the corresponding nonoverlapping counterpart was performed by the

Table 2. Correlation between the codon usage patterns of overlapping and nonoverlapping coding regions (asterisks denote the level of statistical significance)

	% overlap	Correlation coefficient (*r*)	Mean value (*r*) and s.d.
Coliphage			
A protein	29	0.13	
B protein	100	0.55 (***)	
C protein	33	0.01	
K protein	95	0.03	
D protein	55	0.68 (***)	
E protein	99	0.02	
			0.24 (0.29)
Hepatitis B virus			
Core antigen	28	0.39 (**)	
Surface antigen	100	0.33 (*)	
Polymerase	53	0.41 (**)	
			0.38 (0.04)
HIV-1 lentivirus			
Gag	15	0.49 (***)	
Pol	9	0.59 (***)	
Vif	21	0.29 (*)	
Vpr	29	0.38 (**)	
Vpu	35	0.46 (***)	
Tat	50	0.07	
Rev	100	0.37 (**)	
Env	14	0.31 (*)	
			0.37 (0.16)
Luteovirus			
29-kD protein	81	0.39 (**)	
66-kD protein	64	0.35 (*)	
Polymerase	30	0.55 (***)	
Coat protein	81	0.42 (**)	
VPG protein	100	−0.09	
			0.32 (0.24)
Tymovirus			
Replicase	35	0.90 (***)	
69-kD protein	99	0.74 (***)	
			0.82 (0.11)

* $P < 0.05$
** $P < 0.005$
*** $P < 0.0005$ for 57 degrees of freedom

chi-square contingency-table test. Data were arranged in a 2 × 2 table whose *a*, *b*, *c*, *d* values correspond to the content of a given amino-acid residue in a given overlapping frame (*a*), in the nonoverlapping frames (*b*), and to the total amount of the other amino acid residues in the same overlapping frame (*c*) and in the nonoverlapping frames (*d*). The counting of the chi-square values above the 3.8 cutoff ($P < 0.05$ for 1 degree of freedom), expressing a significantly higher content of amino-acid residues in the overlapping genes, led to the general representation shown in Fig. 3. It appears that the amino-acid composition bias within overlapping genes can be mainly ascribed to amino-acid residues with the highest level of codon degeneracy (e.g., arginine, serine, and leucine residues are expressed by six synonymous codons each and proline residue by four synonyms). This findings was also corroborated when considering the highest compositional differences. As summarized in

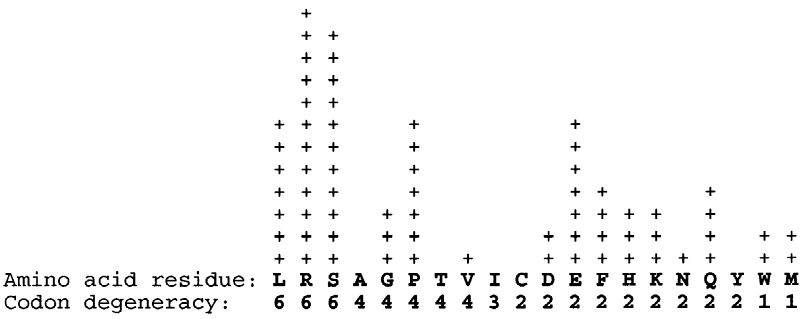

Fig. 3. Frequency of occurrence of chi-square values above the 3.8 cutoff value which reflect a significantly higher content of individual amino acid residues in proteins encoded by overlapping genes.

Table 3, out of a total of ten chi-square values higher than 30.0 ($P < 0.00001$), four were ascribed to arginine, three to leucine, two to proline, and one to methionine residues.

Table 3. Overlapping genes exhibiting the highest difference in the frequency of individual amino-acid residues with respect to the nonoverlapping genes

Overlapping gene	Amino acid residue	Frequency (%)	Frequency nonoverlap genes (%)	Virus family
K protein	Leucine	26.1	7.6	Coliphage
E protein	Leucine	23.5	7.6	Coliphage
Core antigen	Arginine	22.7	4.8	Hepatitis B virus
Gag	Proline	15.1	4.9	Lentivirus
Vif	Methionine	10.3	1.8	Lentivirus
Rev	Arginine	12.9	2.5	Lentivirus
Env	Leucine	17.2	8.3	Lentivirus
Env	Arginine	13.0	5.2	Lentivirus
69-kD protein	Proline	18.0	10.0	Tymovirus
69-kD protein	Arginine	9.2	4.1	Tymovirus

Discussion

It has recently been proposed (Keese and Gibbs 1992) that overlapping gene arrangements may arise de novo, thus encoding novel specialized proteins; it has also been hypothesized that a new gene arisen in this way will have an unusual codon usage and will encode a protein with biased physicochemical properties. Taking into account these observations, we have analyzed the informational content of viral overlapping genes at different levels of complexity.

The use of information theory indices shows that viral sequences, albeit deriving from different sources, can be referred to two distinct patterns. Considering that (Luo et al. 1988) ''the smallness of D_1 represents the abundance of vocabulary and the largeness of D_2 represents the clarity of grammatical rules,'' the informational measures of overlapping sequences related to the first pattern (a low D_1 value, a high D_2 value, see Fig. 2) suggest a level of genetic information storage closely resembling natural languages. The occurrence of these constraints in coliphage, hepatitis B virus, HIV-1 lentivirus, and luteovirus also reflects a peculiar pattern in the usage of synonymous codons for most of the corresponding overlapping frames (Table 2).

Since the most striking difference in the choice of synonyms concerns the family of coliphages, some speculations on the origin of its overlapping genes (see the genomic map shown in Fig. 1) can be proposed. For example, the codon usage pattern of overlapping frames encoding the structural ''scaffolding'' B and D proteins is well correlated (an *r* value of 0.55 and 0.68, respectively) with that of the nonoverlapping genes, which encode the similarly structural J, F, G, and H proteins. This relatively high degree of correlation suggests an ancient origin for the B and D frames. On the other hand, the highly peculiar choice of synonyms in the genes E and K, which are entirely embedded within the D and A/C genes and exhibit an *r* value of 0.02 and 0.03, respectively, supports the idea of a more recent acquisition. Since a low expression of the gene E is necessary and sufficient to induce lysis of the host cell (Blasi et al. 1990), its peculiar codon usage pattern could represent a mechanism to regulate the rate of translation, thus preventing premature lysis of the host. A low expression during infectious cycle could also be required for the gene K, whose regulative role consists in increasing the burst size of phage production (Gillam et al. 1985). The codon usage pattern of the region of gene A which overlaps both B and K genes (an *r* value of 0.13 with the nonoverlapping frames) contrasts with that found in the nonoverlapping part of the gene A (an *r* value of 0.91). This observation suggests that a shorter gene A originally terminated in close proximity to a preexisting gene B and that the present overlapping arrangement evolved by a new termination codon of the gene A beyond the gene B. In a similar way, the codon usage of the overlapping fraction of gene C (an *r* value of 0.01 with the regions of nonoverlap) markedly differs from that occurring in the nonoverlapping region (an *r* value of 0.41). Considering that the gene C of ϕX174, G4, and α3 coliphages contains a second in-phase ATG codon localized in the nonoverlapping region, we predict an originally shorter gene C which evolved using as initiator codon an upstream ATG localized at the end of the gene A.

The alternative pattern revealed by the information

theory approach, characterized by similar constraints in overlapping and nonoverlapping sequences, concerns the family of plant tymoviruses. In this case, the strong bias occurring in the base composition of the regions of nonoverlap (A = 22%, T = 23%, G = 17%, C = 38%) is preserved in the overlapping regions (A = 21%, T = 23%, G = 16%, C = 40%), and this excess of C residues tends to be clustered in the third base position of codons. In fact, a 49% content of C residues occurs in the third base position of nonoverlapping regions, a 44% content in the replicase protein (RP) overlapping frame, and a 36% content in the 69-kD protein overlapping frame. The high degree of relationship in the codon usage (an *r* mean value of 0.82) likely suggests that, at variance with the case of ϕX174, the infectious cycle of tymoviruses may require, in all coding regions, a more uniform adaptation to the translationary machinery of the host. Tymoviruses infect various members of the Cruciferae (e.g., *Brassica rapa*, *Arabidopsis thaliana*) and they accumulate in leaves (Bozarth et al. 1992), where the highly expressed genes code for the small subunit of ribulose 1,5-bisphosphate carboxylase and for the chlorophyll a/b-binding protein (Murray et al. 1989). Interestingly, the third base positions of the coding regions of these latter genes show a frequency of C residues (40%) similar to that occurring in all different frames of tymoviruses. Since the base frequency at third degenerate position in nonoverlapping regions (A = 16%, T = 21%, C = 49%, G = 14%) is closely similar to RP overlapping frame (A = 19%, T = 18%, C = 44%, G = 19%) and contrasts with 69-kD protein frame (A = 24%, T = 29%, C = 36%, G = 11%), we can also predict that this latter overlapping gene arose later by superimposition on a preexisting RP gene.

The statistical analysis of the amino acid composition evidenced that the peculiar amino-acid usage occurring in overlapping genes can be mainly ascribed to a significantly higher frequency of amino acid residues having the highest level of codon degeneracy (Fig. 3). For example, the high content of leucine and arginine residues in the overlapping region of lentivirus Env gene (Table 3) is related to a very peculiar choice of synonyms (Leu/CTA 4.9%; Leu/TTA 16.9%; Leu/CTC 29.0%, Arg/AGA 31.9%, Arg/CGC 17.0%), when compared with that occurring in the lentivirus nonoverlapping regions (Leu/CTA 18.6%; Leu/TTA 34.2%; Leu/CTC 6.5%; Arg/AGA 65.0%; Arg/CGC 0.7%). Therefore, a localized high frequency of both leucine and arginine residues combined with a strongly different strategy of codon usage within the ancestral Env gene frame can be hypothesized as a basic event to originate the Tat, Rev, and Vpu overlapping frames.

Some overlapping genes exhibiting a strongly preferred occurrence of leucine or arginine residues (Table 3) have previously been demonstrated to perform a crucial function in the viral life cycle. For example, the high content of leucines in the overlapping E protein of coliphages lie within a transmembrane domain that is required to determine *Escherichia coli* cell lysis (Buckley and Hayashi 1986). The high frequency of arginines in the overlapping fraction of the core antigen of hepatitis B virus corresponds to a carboxyl-terminal signal that is involved in nuclear targeting of the protein (Eckhardt et al. 1991). A similar function has been ascribed to the polyarginine motifs of the Rev protein of the HIV-1 lentiviruses (Kubota et al. 1989). It has also been demonstrated (Zapp et al. 1991) that the run of arginines in the center of Rev protein is involved in the recognition, and nucleocytoplasmic transport, of unspliced viral mRNAs.

These data support the notion that the high frequency of amino-acid residues with a high level of codon degeneracy, which appears to be a peculiar feature of overlapping genes, can be viewed as a valuable tool with which to achieve a more flexible strategy in the choice of synonymous codons and/or to gain new specialized functions in the viral life cycle.

Acknowledgments. We are grateful to Professor Franco Conterio for support and encouragement. We also appreciate critical readings of the manuscript by Simone Ottonello and Elena Maestri. This work was supported by the National Research Council of Italy and by the Ministry of University and Scientific and Technological Research.

References

Atkins JF, Steitz JA, Anderson CW, Model P (1979) Binding of mammalian ribosomes to MS2 phage RNA reveals an overlapping gene encoding a lysis function. Cell 18:247–256

Barrell BG, Air GM, Hutchinson CA III (1976) Overlapping genes in bacteriophage ϕX174. Nature 264:34–40

Blasi U, Nam K, Lubitz W, Young Ry (1990) Translational efficiency of ϕX174 lysis gene E is unaffected by upstream translation of the overlapping gene D reading frame. J Bacteriol 172:5617–5623

Bozarth CS, Weiland JJ, Dreher TW (1992) Expression of ORF-69 of turnip yellow mosaic virus is necessary for viral spread in plants. Virology 187:124–130

Brierley I, Digard P, Inglis SC (1989) Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. Cell 57:537–547

Buckley KJ, Hayashi M (1986) Lytic activity localized to membrane-spanning region of ϕX174 E protein. Mol Gen Genet 204:120–125

Chang LJ, Pryciak P, Ganem D, Varmus HE (1989) Biosynthesis of the reverse transcriptase of hepatitis B viruses involved de novo translational initiation not ribosomal frameshifting. Nature 337:364–368

Eckhardt SG, Milich DR, McLachlan A (1991) Hepatitis B virus core antigen has two nuclear localization sequences in the arginine-rich carboxyl terminus. J Virol 65:575–582

Gatlin LL (1968) The information content of DNA. J Theor Biol 18: 181–194

Gatlin LL (1972) Information theory and the living system. Columbia University Press, New York

Gillam S, Atkinson T, Markham A, Smith M (1985) Gene K of bacteriophage ϕX174 codes for a protein which affects the burst size of phage production. J Virol 53:708–709

Granero-Porati MI, Porati A, Zani L (1980) Informational parameters of an exact DNA base sequence. J Theor Biol 86:401–403

Jacks T, Madhani HD, Masiarz FR, Varmus HE (1988) Signals for

ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell 55:447–458

Keese PK, Gibbs A (1992) Origins of genes: ''Big bang'' or continuous creation? Proc Natl Acad Sci USA 89:9489–9493

Kubota S, Siomi H, Satoh T, Endo S, Maki M, Hatanaka M (1989) Functional similarity of HIV-I rev and HTLV-I rex proteins: identification of a new nucleolar-targeting signal in rev protein. Biochem Biophys Res Commun 162:963–970

Lamb RA, Horvath CM (1991) Diversity of coding strategies in influenza viruses. Trends Genet 7:261–266

Luo FL, Tsai L, Zhou YM (1988) Informational parameters of nucleic acid and molecular evolution. J Theor Biol 130:351–361

Miyata T, Yasunaga T (1978) Evolution of overlapping genes. Nature 272:532–535

Murray EE, Lotzer J, Eberle M (1989) Codon usage in plant genes. Nucleic Acids Res 17:477–493

Normark S, Bergstrom S, Edlund T, Grundstrom T, Bengtake J, Lindberg FP, Olsson O (1983) Overlapping genes. Annu Rev Genet 17:499–525

Rice CM, Fuchs R, Higgins DG, Stoehr PJ, Cameron GN (1993) The EMBL data library. Nucleic Acids Res 21:2963–2966

Samuel CE (1989) Polycistronic animal virus mRNAs. Prog Nucleic Acid Res Mol Biol 37:127–153

Sander C, Schulz GE (1979) Degeneracy of the information contained in amino acid sequences: evidence from overlaid genes. J Mol Evol 13:245–252

Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana

Sharp PM, Li WH (1987) The codon adaptation index, a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Smith TF (1989) Semantic and syntactic patterns in the genetic language. In: Colwell RR (ed) Biomolecular data. A resource in transition. Oxford University Press, Oxford, p 211

Smith TF, Waterman MS (1980) Protein constraints induced by multiframe encoding. Math Biosci 49:17–26

Smith TF, Waterman MS (1981) Overlapping genes and information theory. J Theor Biol 91:379–380

Snedecor GW, Cochran WG (1967) Statistical methods. Iowa State University Press, Ames, p 228

Soeda E, Maruyama T (1982) Molecular evolution in papova viruses and bacteriophages. Adv Biophys 15:1–17

Trifonov EN (1989) Searching for codes in the sequences. In: Colwell RR (ed) Biomolecular data. A resource in transition. Oxford University Press, Oxford, p 199

Wilson W, Braddock M, Adam SE, Rathjen PD, Kingsman SM, Kingsman AJ (1988) HIV expression strategies: ribosomal frameshifting is directed by a short sequence in both mammalian and yeast systems. Cell 55:1159–1169

Yockey HP (1979) Do overlapping genes violate molecular biology and the theory of evolution? J Theor Biol 80:21–26

Zapp ML, Hope TJ, Parslow TG, Green MR (1991) Oligomerization and RNA binding domains of the type I human immunodeficiency virus Rev protein: a dual function for an arginine-rich binding motif. Proc Natl Acad Sci USA 88:7734–7738