J Mol Evol (2000) 50:284–295
DOI: 10.1007/s002399910033

JOURNAL OF MOLECULAR EVOLUTION

Detection of Signature Sequences in Overlapping Genes and Prediction of a Novel Overlapping Gene in Hepatitis G Virus

Angelo Pavesi

Department of Evolutionary and Functional Biology, University of Parma, Parco Area delle Scienze 11/A, I-43100 Parma, Italy

Received: 21 July 1999 / Accepted: 26 October 1999

Abstract. In viruses an increased coding ability is provided by overlapping genes, in which two alternative open reading frames (ORFs) may be translated to yield two distinct proteins. The identification of signature sequences in overlapping genes is a topic of particular interest, since additional out-of-frame coding regions can be nested within known genes. In this work, a novel feature peculiar to overlapping coding regions is presented. It was detected by analysis of a sample set of 21 virus genomic sequences and consisted in the repeated occurrence of a cluster of basic amino acid residues, encoded by a frame, combined to a stretch of acidic residues, encoded by the corresponding overlapping frame. A computer scan of an additional set of virus sequences demonstrated that this feature is common to several other known overlapping ORFs and led to prediction of a novel overlapping gene in hepatitis G virus (HGV). The occurrence of a bifunctional coding region in HGV was also supported by its extremely lower rate of synonymous nucleotide substitutions compared to that observed in the other gene regions of the HGV genome. Analysis of the amino acid sequence that was deduced from the putative overlapping gene revealed a high content of basic residues and the presence of a nuclear targeting signal; these characteristics suggest that a core-like protein may be expressed by this novel ORF.

Key words: Overlapping genes — Amino acid composition — Codon usage — Constrained evolution — Synonymous substitution rate — Hepatitis G virus

Introduction

The genome of viruses often contains more coding information than a colinear relationship between nucleotide and protein sequences would suggest. An increased coding density is provided by overlapping genes, in which two alternative open reading frames (ORFs) on the same strand may be translated to yield two distinct polypeptide products. This pattern not only compacts the genetic information of a limited-size genome but also is a way to keep the expression of linked genes under the same regulatory control. The first overlapping-gene systems were identified in ϕX174 and G4 bacteriophages and consisted of the genes E, B, and K being entirely nested within the D, A, and C genes, respectively (Barrell et al. 1976; Fiddes and Godson 1979). Other overlapping-gene arrangements were then described in several families of both animal (Samuel 1989 and references therein) and plant viruses (Morch et al. 1988; Mayo et al. 1989; Beck et al. 1991).

Severe evolutionary constraints should be imposed on a nucleotide sequence encoding two functionally unrelated proteins, as all silent third codon-position substitutions in a frame cause amino acid changes in the sequence translated from the +1 overlapping frame. A clear example of constrained evolution is given by the pattern of nucleotide substitutions occurring in the polymerase/surface antigen overlapping genes of hepatitis B virus, as evidenced by an extensive phylogenetic analysis of 27 primate strains (Mizokami et al. 1997). An accurate elucidation of this kind of evolutionary pressures should also allow for the discovery of novel overlapping genes, hidden in viral genomic sequences. For example, a de-

Correspondence to: Angelo Pavesi; *e-mail:* maestri@unipr.it

tailed analysis of the rates of substitution at the first, second, and third codon positions within the phosphoprotein gene of 18 vesicular stomatitis virus isolates led to identification of a short region with an extreme reduction of third-base variability (Bilsel et al. 1990). The speculation that this anomaly may be explained by the presence of a functional, one position-shifted frame was supported by the finding that this overlapping ORF actually encodes a small arginine-rich protein (Spiropoulou and Nicol 1993), for which a stimulating effect both on the level and on the fidelity of mRNA synthesis was then demonstrated (Peluso et al. 1996). Based on a similar approach, a novel overlapping gene, essential for systemic virus spread in host plants, was shown to be highly conserved in cucumovirus, but absent in the other members of the Bromoviridae family (Ding et al. 1994, 1995).

We recently evaluated the information content of overlapping genes lying in the genomes of prokaryotic and eukaryotic viruses, by a series of comparisons with the corresponding nonoverlapping-gene counterpart. A pattern of evolutionary constraints acting on mono- and dinucleotide compositions, and thus affecting the use of synonymous codons, was found to be characteristic of bifunctional genes. A greater content of amino acid residues with the highest level of codon degeneracy (arginine, leucine, and serine) was found in proteins encoded by overlapping reading frames, compared to proteins expressed by the nonoverlapping ones (Pavesi et al. 1997).

By extending this kind of comparative analysis to a larger set of data, I describe additional features peculiar to overlapping genes. They consist mainly in the repeated occurrence of a cluster of basic amino acid residues encoded by a frame, combined to a stretch of acidic residues encoded by the corresponding overlapping frame. Such an arrangement was used as a probe in the scan of viral genomic sequences, to detect, nested within known genes, novel potentially functional overlapping ORFs.

Methods

The nucleotide sequence data for 21 fully sequenced viral genomes, all containing a relevant amount of overlapping coding regions, were collected from the EMBL database. Prokaryotic viruses were represented by single-stranded (SS) circular-DNA *Escherichia coli* bacteriophages, corresponding to ϕX174, G4, and α3 species. Animal viruses included double-stranded (DS) circular-DNA hepatitis B virus (HBV) and both types of SS-RNA human immunodeficiency lentivirus. In particular, hepadnaviruses included three strains of avian HBV (two duck and one heron isolates) and the three major serological subtypes (adw4, ayw, and adr) of human HBV. Three HIV-1 strains (two Zaire and one Cambridge isolates) and three HIV-2 strains (two Rhodesia and one Mali isolates) were chosen as representatives of the lentivirus group of retroviruses. Plant viruses were given by six SS-RNA genomes, corresponding to members of luteovirus (barley, beet, and potato leafroll viruses) and tymovirus (eggplant, turnip, and ononis mosaic viruses) families. The entry names and the accession numbers of the 21 selected genomic sequences are as follows: PHIX174; M10867, MIG4XX; V00657, BACHALPA; X60322, HBDGA; M21953, HBDGENM; M95589, HBHCG; M22056, HBVADW4A; X69798, HBVAYWCI; X65258, HPBCG; D12980, HIVBRUCG; K02013, HIVNDK; M27323, HIVCAM1; D10112, REHIV2NI; J03654, HIV2RODZ; M15390, HIV2BEN; M30502, BYDCG; L25299, BWYVFL1; X13063, PLLGRNA; X14600, MTYRPVP; J04374, TYMVCG; X16378, MTYCG; J04375.

Referring to position coordinates featured in the database, the nucleotide sequences of overlapping and nonoverlapping coding regions were extracted from each of the examined genomes as separate entities, and the corresponding amino acid sequences were obtained. A k-tuple statistical analysis was carried out by the chi-square contigency-table test (Snedecor and Cochran 1967), to assess for significant differences between the compositions of overlapping and nonoverlapping protein sequences at individual amino acid ($k = 1$), dipeptide ($k = 2$), and tripeptide ($k = 3$) levels. Bias in the use of synonymous codons was tested by estimating the “effective number of codons” (N_c index) for a given amino acid residue (Wright 1990). The multiple alignment of nucleotide sequences was carried out using the computer program CLUSTAL W (Thompson et al. 1994). The evolutionary distance between two protein coding sequences was evaluated by the method of Nei and Gojobori (1986): the fraction of synonymous differences was first estimated, and the number of synonymous substitutions per site between the two compared sequences was then obtained using the K_s index. The statistical comparison among the K_s index values obtained from different sets of homologous sequences was carried out using the nonparametric test of Wilcoxon (Siegel 1956). Management and retrieval of virus nucleotide sequences stored in the genetic database were carried out using facilities included in the software package PC/GENE Release 18.0 (Intelligenetics, Mountain View, CA).

Results

Description of the Sample Set of Virus Genomic Sequences

The 21 viruses under examination were subdivided into seven groups, in accordance to the host specificity or for belonging to different families or types (bacteriophage, avian HBV, human HBV, human HIV-1 lentivirus, human HIV-2 lentivirus, plant luteovirus, and plant tymovirus). The genomic organization peculiar to each group is reported in Fig. 1.

The amount of the genetic information encoded by all of the overlapping reading frames was evaluated for each of the seven virus groups. The highest coding fraction provided by overlapping genes (a 67% of the total coding capacity) was found in human HBV, whose compact genomic organization consists of the surface, core, X, and polymerase genes; the surface gene is completely contained within the polymerase gene, and the core and X genes overlap the polymerase gene with a fourth and a half of their sequence lengths, respectively (Fig. 1). In the tymovirus group, the coding fraction provided by overlapping genes showed an intermediate content (48% of the total). The genome of tymoviruses contains a set of only three genes (Fig. 1); its notable level of overlap is due mainly to a gene which is entirely embedded within the replicase gene and encodes a cell-to-cell movement protein necessary for systemic spread of the virus in the plant host (Bozarth et al. 1992). The lowest content of

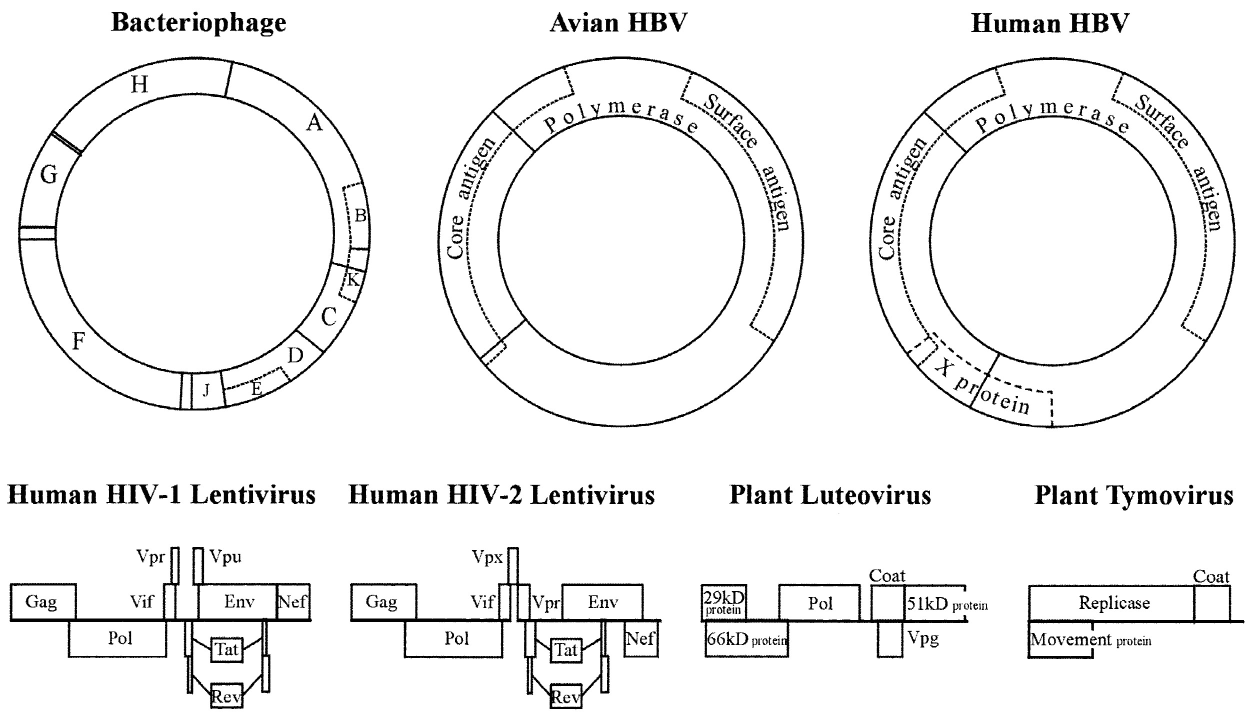

Fig. 1. Genomic organization in the seven virus groups under examination. The average genome size was 5.7 kilobases (kb) in bacteriophages, 3.0 kb in avian hepatitis B viruses, 3.2 kb in human hepatitis B viruses, 9.3 kb in HIV-1 lentiviruses, 9.8 kb in HIV-2 lentiviruses, 5.8 kb in luteoviruses, and 6.3 kb in tymoviruses. The percentage content of amino acids encoded by overlapping genes was 26.2 in bacteriophage, 59.9 in avian HBV, 67.1 in human HBV, 17.3 in HIV-1 lentivirus, 24.5 in HIV-2 lentivirus, 48.4 in luteovirus, and 47.6 in tymovirus.

amino acid residues encoded by bifunctional genes (17% of the total) was found in human HIV-1 lentiviruses, which show, however, the highest degree of genomic complexity. In addition to the retroviral ancestral genes (Gag, Pol, and Env), HIV-1 contains the Tat and Rev regulatory genes, as well as the short accessory Nef, Vpu, Vif, and Vpr genes (Fig. 1). Beside several bicistronic regions, the most complex arrangement is given by the 3′ terminal coding exons of the genes Tat and Rev; they overlap with one other and with the Env gene, leading to all three reading frames of this region being simultaneously expressed.

Identification of Discriminant Features in Proteins Encoded by Overlapping Genes

An increased frequency of amino acids with a high level of codon degeneracy was described previously for proteins encoded by overlapping frames compared to the nonoverlapping ones (Pavesi et al. 1997). Here, such comparative study was extended, using a larger set of data and performing a *k*-tuple analysis. The aim of this analysis was the identification of the amino acids showing, in the overlapping set, a higher frequency of occurrence not only at the individual residue level ($k = 1$), but also at the dipeptide ($k = 2$) and tripeptide ($k = 3$) levels. In other words, the tendency of amino acids to originate short clusters of identical residues was tested as a feature peculiar to proteins encoded by overlapping genes.

As shown in Table 1, two sets of data were compared for each of the virus groups under examination; they corresponded to the amino acid composition of proteins encoded by all of the overlapping frames and that obtained from proteins expressed by the respective nonoverlapping counterpart. The statistical comparison was carried out at the mono-, di-, and tripeptide levels, using the chi-square contingency-table test. At a chi-square value of 3.84 ($P < 0.05$), very few amino acids matched all of the three imposed conditions, that is, a higher frequency of occurrence, in the overlapping set, at the individual, dipeptide, and tripeptide levels (Table 1). They corresponded to the sixfold degenerate leucine residue in bacteriophages and to the equally degenerate arginine residue in all of the other groups, with the exception of tymoviruses. A detailed analysis pointed out that the clusters of leucines peculiar to bacteriophages are specifically encoded by the overlapping genes E and K. In avian and human HBVs, repeated stretches of three or four consecutive arginines were localized in the overlapping fraction of both core-antigen and polymerase genes. In HIV-1 and HIV-2 lentiviruses, highly conserved polyarginine motifs were found in the Tat and Rev regulatory genes. In luteoviruses, the arginine-rich regions that are encoded by overlapping frames were localized within the 29-kD, 66-kD, polymerase, and coat proteins;

Table 1. Comparison between the amino acid compositions of overlapping and nonoverlapping protein sequences: residues with a higher frequency of occurrence in the overlapping set at the individual ($k = 1$), dipeptide ($k = 2$), and tripeptide ($k = 3$) levels are reported

	$k = 1$			$k = 2$			$k = 3$			
Amino acid	Overlap	Nonoverlap	$\chi^{2a,c}$	Overlap	Nonoverlap	$\chi^{2b,c}$	Overlap	Nonoverlap	$\chi^{2b,c}$	Virus group
Leucine	195 (1366)	334 (4048)	33.66***	41 (195)	21 (334)	19.82***	14 (195)	1 (334)	19.66***	Bacteriophage
Arginine	222 (2354)	82 (1644)	23.54***	45 (222)	3 (82)	9.72**	12 (222)	0 (82)	4.37*	Avian HBV
Arginine	233 (2994)	85 (1493)	5.77*	48 (233)	4 (85)	8.87**	18 (233)	0 (85)	6.44*	Human HBV
Arginine	175 (1463)	406 (7398)	70.44***	26 (175)	20 (406)	13.64***	8 (175)	1 (406)	14.31***	HIV-1
Arginine	267 (2155)	393 (7035)	96.03***	30 (267)	26 (393)	3.86*	16 (267)	1 (393)	19.62***	HIV-2
Arginine	261 (3126)	190 (3410)	17.04***	52 (261)	3 (190)	27.95***	23 (261)	0 (190)	16.17***	Luteovirus

[a] The χ^2 test at $k = 1$ was performed using a 2 × 2 contingency table, whose values correspond to the content of a given amino acid residue in the overlapping and nonoverlapping sets and to the total amount of the other residues (parenthesized values) in the same two sets of data.

[b] In the χ^2 test at $k = 2$ or $k = 3$, the frequencies of dipeptide or tripeptide sequences in overlapping and nonoverlapping sets were included in the respective 2 × 2 table; these values were compared with the frequencies of the corresponding amino acid residue (parenthesized values) in the same two sets of data.

[c] Asterisks denote the level of statistical significance at 1 degree of freedom: $*P < 0.05$; $**P < 0.005$; $***P < 0.0005$.

in particular, a cluster of seven consecutive arginines was found in the coat protein of potato leafroll virus.

The natural question raised by these results was what restrictions are imposed on the use of synonymous codons in regions encoding such polyleucine or polyarginine sequences. If a frame is highly biased in the choice of the six Leu or Arg synonyms, we would expect that the amino acid sequence encoded by the alternative frame should be specifically affected. Using the N_c index of Wright (1990), no bias was found in bacteriophages, where the E and K overlapping frames use five of the six Leu synonyms to express the corresponding leucine-rich regions (Table 2). On the contrary, a relevant bias was found in the overlapping regions that encode the arginine clusters of HBV and lentivirus proteins; it was ascribed to a strong preference for AGA codon, with respect to the AGG codon, and to a low content of most of the CGX codons, due mainly to a selective pressure against the CpG dinucleotide in the entire genomic sequence (see last column in Table 2). Although to a lesser extent, similar restrictions in the use of Arg synonyms were also observed in luteoviruses, where, instead, no CpG suppression was observed.

By also taking into account the N_c values of Leu and Arg in nonoverlapping genes, the data reported in Table 2 should stress the following points: (*i*) the use of Leu synonyms in overlapping frames encoding polyleucine genes is similar to that determined for leucine residues encoded by the nonoverlapping set; (*ii*) the use of Arg synonyms in overlapping regions encoding polyarginine sequences is, instead, more biased than that occurring in nonoverlapping ORFs, with the exception of HIV-1 and HIV-2 lentiviruses, where a strong CpG suppression occurs; and (*iii*) the content of AGA and CGA codons in overlapping frames expressing Arg clusters is always higher than that of the respective nonoverlapping set.

Considering, for example, a cluster of five arginine residues encoded by the AGA.AGA.AGA.AGA.AGA sequence, a one-base frameshift would yield a stretch of four glutamic acid residues $(GAA)_4$. Alternatively, the same arginine cluster may be expressed by the AGA. CGA.AGA.CGA.AGA sequence, with the alternative frame giving a stretch of four acidic residues (Asp–Glu–Asp–Glu). Referring to the observed bias toward AGA and CGA arginine codons, ranging from 57% in luteovirus to 83% in HIV-1 lentivirus (Table 2), the coding potential of the alternative +1 frame was found to be really affected, leading to the expression of clusters with a prevalent content of acidic residues. This feature was a common characteristic of avian and human HBVs, of both HIV-1 and HIV-2 lentiviruses, as well as of plant luteoviruses. Altogether, against a total of 171 Arg residues, encoded by a reading frame as blocks of three or more consecutive residues, 115 Asp or Glu residues were found expressed by the corresponding, one position-shifted, overlapping frame.

A Computer Algorithm Searching for Signature Sequences Peculiar to Overlapping Genes

A computer algorithm searching for stretches of at least three consecutive Arg residues in one frame with acidic amino acids in the +1 overlapping frame was developed. The filters included in the algorithm were the following: (i) a ratio equal to or higher than 0.5 between the number of Arg residues in a frame and the number of acidic residues in the alternative frame; and (ii) a minimal

Table 2. Frequencies of the six Leu or Arg synonymous codons in the viral overlapping regions which encode blocks of at least three consecutive Leu or Arg residues

	Leu synonymous codon						
	TTA	TTG	CTA	CTT	CTG	CTC	N_c(Leu)[a,b]
Bacteriophage	4	8	2	3	11	4	5.0 (4.3)

	Arg synonymous codon								
	AGA	AGG	CGA	CGT	CGG	CGC	N_c(Arg)[a,b]	(AGA+CGA) %[c]	Obs/Exp CpG[d]
Avian HBV	20	5	2	3	0	0	2.1 (3.7)	73 (55)	0.58
Human HBV	21	2	9	3	0	7	3.2 (4.9)	71 (30)	0.53
HIV-1 lentivirus	10	3	5	0	0	0	2.6 (2.0)	83 (69)	0.21
HIV-2 lentivirus	16	4	5	2	3	0	3.1 (2.4)	70 (61)	0.31
Luteovirus	17	5	12	1	3	13	4.4 (5.8)	57 (39)	0.87

[a] The effective number of codons (N_c) for Leu or Arg was calculated as follows. The frequencies of the six synonymous codons ($p_1, \ldots, p_6$) were obtained by dividing the respective usages ($n_1, \ldots, n_6$) by n, the total number of Leu or Arg synonyms (for example, $p(\mathrm{TTA})^{\mathrm{Leu}} = 4/32$). The sum of the six squared frequencies (S) was then calculated, and the homozygosity (F) was obtained as $F = (nS - 1)/(n - 1)$. The N_c value was evaluated as $N_c = 1/F$.

[b] Parenthesized values correspond to the effective number of codons (N_c) for Leu or Arg determined in nonoverlapping coding regions.

[c] Parenthesized values correspond to the percentage frequencies of AGA and CGA codons in nonoverlapping coding regions.

[d] The ratio of observed/expected CpG was calculated by dividing the frequency of dinucleotide CG for the product between the frequencies of mononucleotides C and G and then multiplying by N, the total number of nucleotides in the sequence being analyzed.

length of 73 nt for a potential overlapping sequence, a limit coinciding with the shortest overlapping-gene arrangement containing the proposed discriminant motif (the 5′ coding exons of the genes Tat and Rev of HIV-2 lentiviruses). The accuracy of the algorithm was tested on both strands of the 21 genomic sequences belonging to the sample set (a total of 259 kb). Under the above rules, 39 of 46 motifs lying in overlapping frames were detected by the algorithm; the motifs which were missed are sequences with a ratio Arg/(Asp + Glu) lower than 0.5. The main drawback of the method was a relatively high number of false-positive cases, since 14 motifs were erroneously identified within nonoverlapping coding regions.

Such computer analysis was extended to a total of 422 complete genomic sequences extracted from the EMBL database and belonging to taxonomic groups that are different from those of the sample set. They were chosen on the basis of a limited genomic size, ranging from 1239 bp in satellite tobacco necrosis virus to 27.6 kb in avian infectious bronchitis coronavirus. As additional subroutine, the search for acidic residues, paired with Arg residues, was carried out not only in the +1 frame, but also in the +2 frame. This option was motivated by the fact that a 2-base frameshift in the AGG.AGG.AGG.AGG.AGG sequence, for example, would yield five arginine residues by the AGG codon against four glutamic acid residues by the GAG codon.

The scan of both strands of the selected genomic sequences (a total of six megabases) led to the identification of 48 motifs within known overlapping-gene systems. The location of these motifs, most of which were highly conserved in different virus strains, is given in Table 3. In human papillomavirus type 4, two contiguous stretches of five and three Arg residues, both encoded by the E2 frame, were found coupled with two clusters of acidic residues (Asp–Glu–Asp–Asp–Glu and Glu–Glu–Glu), both expressed by the E4 overlapping frame. A similar motif was detected in other overlapping genes, such as the large and middle tumor antigens of human polyoma virus (Arg–Arg–Arg/Glu–Asp–Glu), the 14- and 24-kD proteins of Borna disease virus (Arg–Arg–Arg/Glu–Glu–Asp), and the RNA polymerase/movement protein of peanut stunt cucumovirus (Arg–Arg–Arg/Asp–Glu–Asp). Similar signal sequences were found highly conserved in six isolates of chicken anemia virus: two distinct Arg–Arg–Arg sequences, both encoded by the major capsid 52-kD protein gene, overlap the Asp–Glu–Asp and the Val–Asp–Glu sequences, both encoded by the 24-kD structural protein gene. Three other overlapping-gene systems exhibited a similar arrangement: the V3/V2 genes of beet curly top geminivirus (Arg–Arg–Arg–Arg/Glu–Glu–Gly–Glu), the P21/P20 genes of cucumber necrosis tombusvirus (Arg–Arg–Arg/Glu–Asp–Ala), and the NS1/NP1 noncapsid protein genes of bovine parvovirus. In the two last examples in Table 3,

Table 3. Paired blocks of arginine (R) or acidic (D, E) amino acid residues detected within known overlapping-gene systems by the computer search algoorithm

Virus	Accession No.	Overlapping-gene system	Overlapping coding region	Nucleotide position coordinates
Human papillomavirus type 4	X70827	E2 protein E4 protein	R R R R R C G A C G A A G A C G A C G A G D E D D E	3319–3334
			R R R A G A A G A A G A G E E E	3376–3385
Murine polyoma virus	J02289	Large tumor antigen Middle tumor antigen	R R R A G A A G A C G A A E D E	1072–1081
Borna disease virus	U04608	P14 protein P24 protein	R R R C G A A G A A G A T E E D	1313–1322
Peanut stunt cucumovirus	U15729	RNA polymerase Movement protein	R R R C G A C G A A G A T D E D	2574–2583
Chicken anemia virus	U66304	52-kD protein 24-kD protein	R R R C G A C G A A G A T D E D	958–967
			R R R C G T A G A C G A G V D E	1000–1009
Beet curly top virus	X97203	V3 protein V2 protein	R R R R A G A A G A A G G C G A G E E G E	474–486
Cucumber necrosis virus	M25270	P21 protein P20 protein	R R R A G A A G A C G C C E D A	4303–4312
Bovine parvovirus	M14363	NS1 protein NP1 protein	R R R C G A C G C A G A A D A D	2717–2726
Rous sarcoma virus	J02342	Transcriptional activator Gag–Pol–PR180 polyprotein	R R R C G C C G G A G G A A P E E	1031–1041
Apple chlorotic leaf spot virus	X99752	28-kD protein 50-kD protein	R R R R A G A A G A A G G A G G A T K K E D	6778–6791

corresponding to the Rous sarcoma virus and to a plant trichovirus, the motif was found expressed by a two position-shifted overlapping-gene arrangement.

Again, the scan of this set of data evidenced a low selective power of the algorithm: on a total of six megabases of examined sequences, 69 motifs were localized in genomic regions that have anything to do with the overlapping-gene arrangement.

Identification of a New Potential Overlapping Gene in Hepatitis G Virus

Besides the repeated occurrence of the proposed motif in known overlapping genes, the computer analysis described above led to the detection of a new potential overlapping gene in hepatitis G virus (HGV). The first clue was the identification of an Arg–Arg–Arg/Asn–Glu–Asp motif in all of the 37 worldwide geographic strains of HGV, for which the complete genomic sequence (9 kb) was available. The Asn–Glu–Asp sequence was found expressed by the region that encodes the nonstructural protein NS5A. Unexpectedly, the Arg–Arg–Arg sequence was deduced from an uncharacterized ORF, one position shifted with respect to the NS5A frame. By delimiting the boundaries of this ORF (an initiator ATG codon and a TAA stop codon), a 255-nt potentially coding region entirely nested within the NS5A gene was characterized. As shown in Fig. 2, translation of this overlapping ORF would potentially give rise to a conserved arginine-rich protein, for which a bipartite nuclear targeting signal was predicted by a PROSITE database search (Hofmann et al. 1999). The notion of a highly basic protein was supported by a relevant pooled content of Arg, His, and Lys basic residues (average frequency of 29.2%). When the consensus sequence of this putative protein was compared to the SWISS-PROT database, no significant similarity to known proteins was evidenced. By examining the aligned set of sequences (Fig. 2), the evolutionary conservation of this additional ORF was apparent, in spite of

```
 1. MSSWRTAVHPLFVVVAERCLCGEKTSPALHRQHLSRLLRAAQMRRPRRCPLRRRIPR------------PLTHSRSSKSPRQPRGRKVSSTWLFPYK...*   (AF081782  6855-7109   CHI)
 2. MSSWRTAVHPLFVVVAERCLCGEKTSPAHHRLHLSRLRRAAQMRRPRRRPLRRRIPR------------PRTHLKSSKSLILLRVRIASSTWLFPYP...*   (AF006500  6755-7009   CHI)
 3. MSSWRTAVHPLFVVVAERCLCGEKTSPALHRQHLSRLLRAAQMRRPRRCPLRRRIPR------------PLTHSRSSKSPRQPRGRKVSSTWLFPY*       (D87255    6857-7111   JAP)
 4. MSSWRTAVHPLFVVVAERCLCGEKTYPALHRLHLSRLLRAAQMRRPRRCPPRRRIPR------------PQTHLKSSKSLILLRVRKASSTWLFPY*       (D87713    6875-7129   JAP)
 5. MSSWRTAVHPLSVVVAERCLYGEKTSPAHHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PQTHLKSSKSLILLKVRKASSTWLFPY*       (U94695    6839-7093   CHI)
 6. MSSWRIAVHPLSVVVAERCLCGEKTSPALHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PQTHSKSSKSQRQLRGRTVSSTWLFLY*       (U63715    6846-7100   EAF)
 7. MSSWRTAVHPLSVEVAERCLCGEKTYPALHRLHLSRLRRAAQMRRPRRCLPRRRIPR------------PQTHLKSSKSLILLRVRKASSTWLFPY*       (D90601    6874-7128   JAP)
 8. MSSWRTAVHPLFVGVAERCLCGEKTSPALHRQHLSRLPRAAQMRRPRRCPLRRRIPR------------PQTHSRSSKSPRQLKGRKASSTWLFPY*       (D90600    6877-7131   JAP)
 9. MSSWRTAVHPLSVGVAERCLYGEKTYPVLHRLHLSRLRRAAQMRRPRRCLPRRRIPR------------PRTHLKSSKSLILLRVRKASSTWLFPY*       (D87263    6875-7129   JAP)
10. MSSWRTAVHPLSVGVAERCLYGEKTYPVLHRLHLSRLRRAAQMRRPRRCLPRRRIPR------------PRTHLKSSKSLILLRVRKASSTWLFPY*       (D87262    6875-7129   JAP)
11. MSSWRIAVHPLFVVVAERCLCGEKTSPALHRQHLSRLPRAAQMRRPRRCPPRRRIPR------------PLTHSKSSKSLRQLKERKVSSTWLFPY*       (U45966    6802-7056   USA)
12. MSSWRTAVHPLFVVVAERCLYGEKTSPVLHRQHLSRLLRAAQMRRPRRCPPRRRIPR------------PLTHSRSSKSPRQPKGRKVSSTWLFPY*       (U44402    6874-7128   USA)
13. MSSWRIAVHPLSVEVAERCLCGEKTYPALHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PRTHLKSSKSLILLRVRKASSTWLFPY*       (D87712    6875-7129   JAP)
14. MSSWRTAVHPLSVGVAERCPCGEKTYPARHRLHLSRLQRAAQMRRPRRCPPRRRIPR------------PRTHLKSSRSLILLRVRKASSTWLFPY*       (D87710    6875-7129   JAP)
15. MSSWRTAVHPLSVEVAERCLYGEKTYPAHHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PQTHLKSSKSLILLRVRKGSSTWLFPY*       (D87708    6875-7129   JAP)
16. MSSWRTAVHPLFVEVAERCLCGEKTSPALHRQHLSRLPRAAQMRRPRRCPPRRRIPR------------PLTHSRPFLSPRQLKGRKVSSTWLFPY*       (AF104403  6822-7076   FRA)
17. MSSWRTVVHPLFVVVAERCLCGEKTSPALHRQHLSRLLRAAQMRRPRRCPLRRRIPR------------PLTHSKSSKSLRQPKGRTTSSTWPFPY*       (AF031829  6854-7108   USA)
18. MSSWRTVVHPLFVVVAERCLCGEKTSPALHRQHLSRLLRAAQMRRPRRCPLRRRIPR------------PLTHSKSSKSLRQPKGRTTSSTWPFPY*       (AF031828  6854-7108   USA)
19. MSSWRTVVHPLFVVVAERCLCGEKTSPALHRQHLSRLLRAAQMRRPRRCPLRRRIPR------------PLTHSKSSKSLRQPKGRTTSSTWPFPY*       (AF031827  6854-7108   USA)
20. MSSWRTVIHPLFVVVAERCLCGEKTYPALHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PQTHLKSSKSLIQLRVRKASSTWLFPY*       (AB021287  6712-6966   EAS)
21. MSSWRTAVHPLSVVVAERCLCGEKTYPALHRLHLSRLRRAAQMRRPRRCLPRRRIPR------------PRTHLKSSKSLIQLRVRKASSTWLFPY*       (AB018667  6710-6964   EAS)
22. MSSWRTAVHPLSVVVAERCLCGEKTYPALHRLHLSRLRRAAQESLPRRCPPRRRIPR------------LRTLLKSSRSLIQLKVRKTSSTWLFLF*       (AB013501  6711-6965   BOL)
23. MSSWRTVVHPLSVEVAERCPCGEKTYPARHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PRTHLKSSKSLILLRVRKASSTWLFPY*       (AB003293  6723-6977   JAP)
24. MSSWRIAVHPLFVVVAERCLCGEKTSPALHRQHLSRLLRAAQMRRPRRCPLRRRTPR------------PLTHSRSSKSPRQPRGRKVSSTWLFPY*       (AB003289  6746-7000   JAP)
25. MSSWRTAVHPLSVEVAERCLYGEKTSPAHHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PQTHLKSSKSLILLRVRKASSTWLFPY*       (D87714    6875-7129   JAP)
26. MSSWRTAVHPLSVEVAERCLCGEKTSPAHHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PQTHLKSSKSLILLRVRKASSTWLFPY*       (D87711    6875-7129   JAP)
27. MSSWRTAVHPLSVEVAERCLYGEKTYPAHHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PQTHLKSSKSLILLRVRKASSTWLFPY*       (D87709    6875-7129   JAP)
28. MSSWRTAVHPLSVEVAERCLCGKKTYPALHRLHLSRLRRAAQMRRPRRRPPRRRIPR------------PQTHLKSSKSLILLRVRKASSTWLFPY*       (AB008342  6874-7128   JAP)
29. MSSWRTAVHPLSVEVAERCLYGEKTSPAHHRLHLSRLRRAAQMRRPRRCPPRRRIPR------------PQTHLKSSKSLILLRVRKASSTWLFPY*       (AB008335  6875-7129   JAP)
30. MSSWRTAVHPLSVEVAERCLYGGKTYPVLHRLHLSRLRRAAQMRRPRPCPPRRRIRR------------PRTHLKSSKSLIQLKVRTTSSTWLFLY*       (AB003292  6723-6977   JAP)
31. MSSWRTAVHPLSVEVAERCLYGEKTSPAHHRLHLSRLRRAAQMRRPRRCLPRRRIPR------------PQTHLKSSKSLKLLRVMKASSTRLFRF*       (AB003290  6745-6999   JAP)
32. MSSWRTVVHPLSVEVAERCLCGEKTYPALHRLHLSRLRRAAQMRRPRRCLPRRRIPR------------PRTHLKSSKSLILLRVRKASSTWLFPY*       (AB003288  6723-6977   JAP)
33. MSSWRTAVHPLFVVVAERCLYGEKTSPALHRQHLSRLPRAAQMRRPCRCPPRRRTPR------------PLTHWKSSKSPRQPKRRKASSTWLFPY*       (U75356    6783-7037   CHI)
34. MSSWRTAVHPLSVEVAERCLYGEKTYPAHHRLH*SRLRRAAQMRRPLRCPPRRRIPR------------PQTHLKSSKSLILLRVRRASSTWLFPY*       (D87715    6875-7129   JAP)
35. MSSWRTAVHPLSVEVAERCLCGEKTYPALHRPH*SRSQRAAQMRRPRPCPPRRRIRR------------PRTHLKSSKSLIQLNQRKASSTWLFPY*       (AB013500  6713-6967   GHA)
36. MSSWRIAVHPLSVVVAERCLCGEKTYPALHRLHLSRLRRAAQMRRPCR*PPRRRTPR------------PQTHLKSSKSLILLNQRKASSTWLFPY*       (U36380    6856-7110   WAF)
37. MSSWRIAVHPLSVEVAERCLCGEKTSPALHRLHLSRLRRAAQMRNHCQCPPRRRMLRRLWCPPHRRTFHRLTLLTSSKSRTQQKGRIMSSTWPSPY*       (AB003291  6722-7012   JAP)
    *****...*** * ***** .* ** *. ** * **  **** . . .   ***  *              *   .  *    . .  ***.   .
```

Fig. 2. Alignment of the deduced amino acid sequences from the overlapping ORF of 37 HGV isolates. *Boldface* characters indicate a putative nuclear targeting signal. *Asterisks* correspond to perfectly conserved amino acid residues. *Periods* indicate similar amino acid residues, in accordance with the PAM scoring matrix (Dayhoff 1978). Terminal stop codons or internal stop codons (see sequences 34, 35, and 36) are denoted by *asterisks*. The terminal codon of the first two sequences delimit a longer overlapping frame having the following position coordinates: 6855–7208 (AF081782) and 6755–7048 (AF006500). Data *in parentheses* report the GenBank/EMBL accession number, the boundaries of the genomic region encoding the reported amino acid sequence, and the geographical localization of each HGV isolate (CHI, China; JAP, Japan; EAF, East Africa; FRA, France; EAS, Southeast Asia; BOL, Bolivia; GHA, Ghana; WAF, West Africa).

Table 4. Mean number of synonymous substitutions per site (K_s index) in HGV gene regions

Gene/region[a]	Nucleotide position coordinates	Number of codons	Mean number of synonymous substitutions per site (K_s index)[b]	SD
E1	1–615	205	0.650	0.197
E2	616–1926	437	0.609	0.175
NS2	1927–2682	252	0.730	0.234
NS3	2683–4650	656	0.700	0.154
NS4A	4651–5283	211	0.547	0.139
NS4B	5284–5595	104	0.757	0.222
NS5A				
Nonoverlapping region	5596–6321	242	0.755	0.236
Putative overlapping region	6322–6576	85	0.227	0.120
Nonoverlapping region	6577–6840	88	0.486	0.159
NS5B	6841–8496[c]	552	0.525	0.112

[a] The genomic organization and the boundaries of the different coding regions were taken from Erker et al. (1996).

[b] The mean number of silent nucleotide substitutions for each gene region was obtained by a series of 666 comparisons among all pairs of the 37 HGV coding sequences under examination.

[c] The higher limit of the position coordinates is 30 nt lower than the typical length of the HGV ORF (8526 nt), because of the presence of a slightly shorter isolate (U75356). An insertion of 36 nt, occurring in the NS5A overlapping region of a single isolate (AB003291), was removed.

the presence of an internal stop codon in three HGV strains (see sequences 34, 35, and 36).

The speculation on the presence of a novel overlapping gene was also supported by an accurate evaluation of the evolutionary distance at synonymous sites among the 37 available HGV isolates. The genome of this virus is a positive-strand RNA, in which a single long ORF encodes a 2842-aa polyprotein. On the basis of a general similarity to the organization of hepatitis C virus, a genomic map of HGV was proposed by Erker and coworkers (1996). It consists of eight distinct coding regions (E1, E2, NS2, NS3, NS4A, NS4B, NS5A, and NS5B), with the structural and nonstructural proteins positioned at the amino terminus and carboxy terminus of the polyprotein, respectively. If a functional ORF is truly nested within the NS5A coding region, a local high degree of sequence conservation should be detectable, as the third-base position in the NS5A reading frame coincides with the second-base position in the +1 overlapping frame. The nucleotide coding sequences of 37 HGV isolates were aligned, and the evolutionary constraints acting on the synonymous sites of each gene region were estimated by the K_S index (Table 4). According to the hypothesis, the lowest mean value of the K_S index (0.227) was found in the fraction of NS5A frame for which a dual coding ability was supposed. Interestingly, this K_S value was about twofold and threefold lower than that obtained from the coding regions of the NS5A frame that are positioned downstream and upstream of the predicted region of overlap. Similarly, higher values of the mean number of silent substitutions per site were obtained from the remaining seven genes of HGV; they ranged from a K_S value of 0.525 (NS5B frame) to a K_S value of 0.757 (B frame). The significance ($P < 0.001$) of the lower variability at synonymous sites in the overlapping fraction of the NS5A gene, compared to each of the other genomic regions, was statistically demonstrated by the Wilcoxon test.

Though this evolutionary analysis agrees with the presence of a novel ORF, the relevance of the low K_S index value (0.227) peculiar to the critical region of the NS5A gene could be doubted, since an equally low K_S value may be associated with other regions with identical size (85 codons) but lacking overlapping ORFs. To meet this objection, an analysis of the distribution of the synonymous variability was carried out along the entire HGV coding sequence by a window scanning of 85 codons (Fig. 3). Indeed, three subregions with a highly decreased rate of synonymous substitutions were localized within the NS5B gene (a K_S value of 0.251), the NS4A gene (a K_S value of 0.296), and the E2 gene (a K_S value of 0.330); a search for the presence of conserved potential coding regions in the respective +1 frames gave negative results. On the other hand, this search pointed out an extremely higher content of stop codons in the three subregions compared to the NS5A gene (Table 5). The hypothesis that this anomaly may be due to selective pressures acting on the NS5A gene to prevent the occurrence of a stop codon in the +1 frame was tested by a simulation approach. For each of the 37 HGV strains, 1000 nucleotide sequences giving a 2842-aa protein identical to the real polyprotein were generated by a computer program; the assignment of synonymous codons to degenerate amino acids was not random, being based on the overall codon usage pattern of all HGV strains. By considering the four subregions with the lowest level of variability (NS5A, NS5B, NS4A, and E2), the number of stop codons in the +1 frame of the real HGV sequence was statistically compared with the mean and standard deviation of the number of stop codons in the +1 frame of the computer-simulated sequences (Table 5). Using the *z*-score test, it was demonstrated that the number of stop codons in the +1 frame of the NS5B, NS4A, and E2 subregions does not differ significantly from that occurring in the simulated sequences (a *z* value lower than the critical threshold of 3 SD was obtained).

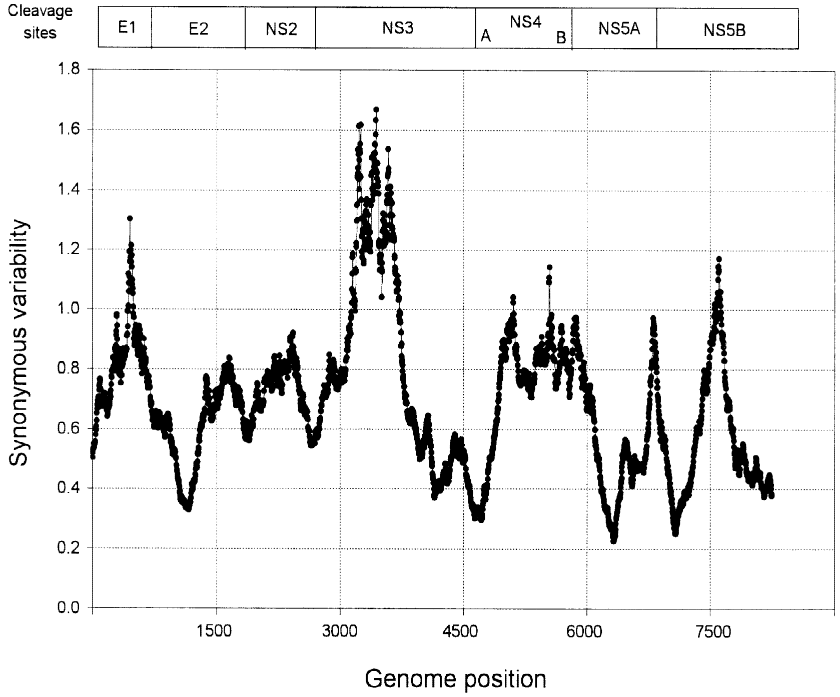

Fig. 3. Mean synonymous diversity along the polyprotein coding region of HGV genome, averaged over a sliding window of 85 codons.

Table 5. Computer simulation test at the four subregions of the HGV polyprotein coding sequence showing the lowest rate of synonymous substitutions

Gene region	Nucleotide position coordinates (255 nt)	K_s index	Number of stop codons +1 frame[a]	Mean number of stop codons +1 frame[b]	z value
NS5A	6322–6576	0.227	3	112.7 (7.3)	15.1
NS5B	7078–7332	0.251	159	134.4 (8.3)	2.9
NS4A	4717–4971	0.296	148	142.8 (8.1)	0.6
E2	1159–1413	0.330	144	164.3 (8.1)	2.5

[a] The number of stop codons was obtained by analysis of the +1 frame in the aligned set of the nucleotide coding sequences of 37 HGV isolates.
[b] The mean number of stop codons was obtained by analysis of the +1 frame of 1000 set of computer-generated HGV coding sequences; the standard deviation is given in parentheses.

An opposite result was obtained, instead, from analysis of the putative region of overlap lying in the NS5A gene (a z value of 15 SD); this clearly indicates a modulation of its codon usage pattern to allow the expression of a continuous overlapping coding region by the +1 frame. This conclusion was confirmed by evaluating the degree of correlation between the codon usage pattern of the entire HGV coding sequence and that peculiar to each of the four subregions. The NS5A subregion showed an extremely low value of the correlation coefficient r (0.075), whereas higher and statistically significant r values (0.45, 0.67, and 0.76) were obtained, respectively, for the NS5B, NS4A, and E2 subregions.

Discussion

The acquisition of overlapping expression in viruses has been described as an episodic event, giving rise to novel ORFs that are exposed to selection for additional specialized functions (Keese and Gibbs 1992). Translation of overlapping genes has been shown to be mediated by ribosomal frameshifting (reviewed by Brierley 1995) or simply by internal initiation in an alternative frame (Chang et al. 1989). Clues on the evolutionary history of overlapping genes can persist in the present-day sequences. This is the case of bacteriophages, where striking differences between the codon usage strategies of overlapping and nonoverlapping coding regions have been documented (Pavesi et al. 1997).

Thus, the identification of discriminant features peculiar to overlapping genes is a crucial topic, since additional and potentially functional frames can be hidden in the vast amount of virus nucleotide sequence data. Some characteristics of both the heptanucleotide sequence of ribosomal slippage and the secondary structure around the frameshift signal have been described for various viral bicistronic mRNAs (Ten Dam et al. 1990). A dissection of the signals that control initiation of translation has been performed using a plant luteovirus mRNA on which protein synthesis initiates at two out-of-frame AUG start codons (Dinesh-Kumar and Miller 1993). The effects of both leader length and AUG codon context on the translational regulation of the p20/p21 bifunctional mRNA of cucumber necrosis virus have been elucidated (Johnston and Rochon 1996). Despite their relevance, these clues on the sequence signals mediating the overlapping expression do not seem to be sufficient for the development of a computer algorithm to analyze large sets of data in a sensitive and selective way.

Here the search for signature sequences in overlapping ORFs was first carried out at the amino acid sequence level, by comparison with the nonoverlapping counterpart in a sample set of seven virus groups. By this analysis, a first discriminant feature of overlapping genes—the expression of short homopolymers of leucine or arginine residues—was documented (Table 1). Interestingly, a relevant role in the infectious phase of the virus life cycle was previously ascribed to such sequence motifs. For example, the polyleucine region of the overlapping E protein of bacteriophages lies within a transmembrane domain causing host cell lysis (Buckley and Hayashi 1986). The clusters of arginines contained in the core protein of hepatitis B virus are part of a signal that is involved in nuclear targeting of the protein (Eckhardt et al. 1991). A similar function was documented for the polyarginine motifs occurring in the Tat and Rev proteins of lentiviruses (Truant and Cullen 1999). A critical role in the recognition, and nucleocytoplasmic transport, of unspliced viral mRNAs was demonstrated for an arginine-rich domain located in the center of the Rev protein (Zapp et al. 1991). Similarly, the trans-activation of viral gene expression by the Tat protein requires a conserved arginine-rich region, which is responsible for the specific binding to a RNA stem-and-loop structure (Delling et al. 1991). In plant viruses, a deletion analysis encompassing the arginine-rich motif contained in the amino-terminal region of the coat protein pointed out a key role of this region in RNA packaging and spread of infection in the host (Rao and Grantham 1996).

In the following step in the analysis, relevant constraints on the use of synonymous codons for the Arg amino acid were elucidated in overlapping regions encoding polyarginine sequences (Table 2). The proposed discriminant feature was a cluster of arginine residues encoded by a frame, combined with a stretch of acidic residues encoded by the corresponding overlapping frame. It was detected in most of the overlapping genes belonging to a sample set of 21 viruses. When a larger set of data, including the complete sequence of 422 limited-size viral genomes, was subjected to a computer search algorithm, this feature was found in other known overlapping-gene systems and appeared as a common characteristic of virus species having distantly related life cycles. The main limit of the algorithm was its high rate of false-positive instances, this preventing an extensive search for overlapping frames in the vast amount of all virus nucleotide sequences determined to date.

Interestingly, this approach led to the identification of a new potential overlapping gene in hepatitis G virus, whose product would correspond to a small highly basic protein. Based on the notion that HGV does not contain a clearly discernible core gene in its 5′ genomic region (Muerhoff et al. 1996), and also considering that flavivirus core proteins are usually highly basic, the above feature suggests that a core-like protein could be encoded by the newly identified ORF. However, the ability of the 5′ region of HGV to encode a core protein was recently proposed by Xiang and co-workers (1998). It was motivated by the conservation in five HGV strains of a short ORF (93 nt) which is in frame and upstream, with respect to the E1 protein coding region, and by a peptide immunoassay demonstrating the actual expression of this ORF. An accurate analysis at the 5′ end of the 37 HGV isolates used in this study revealed a poor degree of conservation for this ORF, with premature termination codons occurring in 16 of 37 sequences. Therefore, the expression of the additional 5′ region proposed by Xiang and co-workers could be related to a translation event, specific to some HGV strains, starting from an AUG upstream of the E1 frame, rather than to the encoding of a functional core protein. Alternatively, a core-like function was ascribed to an expressed ORF which was localized in the antigenome sequence of HGV (Kondo et al. 1998). By a sequence analysis of the antisense genome of 37 HGV isolates, I found a strong evolutionary conservation of this ORF. Its product, a 118-aa protein showing rich hydrophilicity, showed nearly half the content of basic residues (15.6%) compared to that occurring in the amino acid sequence from the ORF predicted here (29.2%). This observation, together with the presence of a bipartite nuclear targeting sequence, could favor the hypothesis of more striking core-like characteristics for the ORF identified in this work.

The presence of a bifunctional coding region in HGV was also supported by the good degree of conservation (68 of the 73 NS5A coding sequences stored in the EMBL database show a 255-bp continuous ORF in the +1 frame) and by the strong reduction of its rate of synonymous substitutions, as evaluated by the K_S index (Table 4). Interestingly, the ratios between the K_S value of the presumed region of overlap (NS5A gene) and the K_S values of the remaining HGV genes are well comparable to those previously observed using the well-known overlapping and nonoverlapping coding regions of human HBV (Mizokami et al. 1997). By a simulation approach, selection pressures affecting the codon usage pattern of the NS5A gene, and thus preventing the occurrence of stop codons in the +1 frame, were evidenced. Of note, these pressures were absent in other gene regions of HGV showing an almost equally low value of the K_S index and lacking overlapping frames (Table 5).

As proposed by Simmonds and Smith (1999), the reduction in the synonymous substitution rate peculiar to these regions, which were localized in the NS5B, NS4A, and E2 genes, is likely due to restrictions on sequence change imposed by RNA secondary structures.

The likelihood of the putative overlapping genes being expressed may be tested experimentally. In fact, the multiple alignment of the amino acid sequences of the deduced gene product (Fig. 2) could be used to design a synthetic peptide, selected on the basis of its antigenic properties and conservation among the reported HGV isolates. The detection of antibodies against the synthetic peptide in serum of patients with chronic HGV infection should reveal an active expression of the predicted overlapping gene.

From an evolutionary point of view, the nature of selective pressures acting on the overlapping fraction of the NS5A gene makes this region an attractive marker to reconstruct the phylogenetic history of HGV. This critical region, in fact, should be preferred to the most variable ones, where the drawback of the multiple substitutions per nucleotide site can alter the actual relatedness among HGV strains. These considerations are of particular interest, if we consider the likely African origin of human HGV (Tanaka et al. 1998) and its proposed role for clarifying the migration patterns of human populations (Suzuki et al. 1999).

Though the features of overlapping genes described in the present work are of limited generality, their simplicity made feasible a computer scan of a relatively large set of virus genomic sequences. Since short clusters of basic or acidic amino acid residues were found to be encoded in a combined fashion by several overlapping ORFs, a further, more general question can be addressed. It concerns the possibility that the overlapping expression, besides being a tool to expand the genetic information of limited-size genomes, is also a way to express, within two overlapping proteins, local clusters of amino acid residues with opposite physicochemical properties. A Fourier analysis comparing the parametric profiles deriving from the primary structure of known overlapping proteins could provide a further key to characterize such a peculiar evolutionary strategy.

Acknowledgments. I am grateful to Professor Franco Conterio for support and encouragement. Insightful suggestions from the anonymous referees and the critical comments of Elena Maestri are also gratefully acknowledged. I thank Paola Pizzetti for preparation of the manuscript. This work was supported by the National Research Council of Italy and by the Ministry of University and Scientific and Technological Research.

References

Barrell BG, Air GM, Hutchinson III CA (1976) Overlapping genes in bacteriophage ϕX174. Nature 264:34–41

Beck DL, Guilford PJ, Voot DM, Andersen MT, Forest RL (1991) Triple gene block proteins of white clover mosaic potexvirus are required for transport. Virology 183:695–702

Bilsel PA, Rowe JE, Fitch WM, Nichol ST (1990) Phosphoprotein and nucleocapsid protein evolution of vesicular stomatitis virus New Jersey. J Virol 64:2498–2504

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 (1995) Ribosomal frameshifting viral RNAs. J Gen Virol 76:1885–1992

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 translation initiation not ribosomal frameshifting. Nature 337:364–368

Dayhoff MO (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, Vol 5, Suppl 3. National Biomedical Research Foundation, Washington DC, pp 345–358

Delling U, Roy S, Sumner-Smith M, Barnett R, Reid L, Rosen CA, Sonenberg N (1991) The number of positively charged amino acids in the basic domain of Tat is critical for transactivation and complex formation with TAR RNA. Proc Natl Acad Sci USA 88:6234–6238

Dinesh-Kumar SP, Miller WA (1993) Control of start codon choice on a plant viral RNA encoding overlapping genes. Plant Cell 5:679–692

Ding SW, Anderson BJ, Haase HR, Symons RH (1994) New overlapping gene encoded by the cucumber mosaic virus genome. Virology 198:593–601

Ding SW, Li WX, Symons RH (1995) A novel naturally occurring hybrid gene encoded by a plant RNA virus facilitates long distance virus movement. EMBO J 23:5762–5772

Erker JC, Simons JN, Muerhoff S, Leary TP, Chalmers ML, Desai SM, Mushahwar IK (1996) Molecular cloning and characterization of a GB virus C isolate from a patient with non-A-E hepatitis. J Gen Virol 77:2713–2720

Fiddes JC, Godson GN (1979) Evolution of the three overlapping gene systems in G4 and ϕX174. J Mol Biol 133:19–43

Hofmann K, Bucher P, Falquet L, Bairoch A (1999) The PROSITE database, its status in 1999. Nucleic Acids Res 27:215–219

Johnston JC, Rochon DM (1996) Both codon context and leader length contribute to efficient expression of two overlapping reading frames of a cucumber necrosis virus bifunctional subgenomic mRNA. Virology 221:232–239

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

Kondo Y, Mizokami M, Nakano T, Kato T, Tanaka Y, Hirashima N, Ueda R, Kunimatsu M, Sasaki M, Yasuda K, Iino S (1998). Analysis of conserved ambisense sequences within GB virus C. J Infect Dis 178:1185–1188

Mayo MA, Robinson DJ, Jolly CA, Hyman L (1989) Nucleotide sequence of potato leafroll luteovirus RNA. J Gen Virol 70:1037–1051

Mizokami M, Orito E, Ohba KI, Ikeo K, Lau JYN, Gojobori T (1997) Constrained evolution with respect to gene overlap of hepatitis B virus. J Mol Evol 44 (Suppl 1):S83–S90

Morch MD, Boyer JC, Haenni AL (1988) Overlapping open reading frames revealed by complete nucleotide sequencing of turnip yellow mosaic virus genomic RNA. Nucleic Acids Res 16:6157–6173

Muerhoff AS, Simons JN, Leary TP, Erker JC, Chalmers ML, Pilot-Matias TJ, Dawson GJ, Desai SM, Mushahwar IK (1996) Sequence heterogeneity within the 5′ terminal region of the hepatitis GB virus C genome and evidence for genotypes. J Hepatol 25:379–384

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Pavesi A, De Iaco B, Granero MI, Porati A (1997) On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. J Mol Evol 44:625–631

Peluso RW, Richardson JC, Talon J, Lock M (1996) Identification of a set of proteins (C′ and C) encoded by the bicistronic P gene of the Indiana serotype of vesicular stomatitis virus and analysis of their effect on transcription by the viral RNA polymerase. Virology 218: 335–342

Rao AL, Grantham GL (1996) Molecular studies on bromovirus capsid protein. Functional analysis of the amino-terminal arginine-rich motif and its role in encapsidation, movement, and pathology. Virology 226:294–305

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

Siegel S (1956) Nonparametric statistics for the behavioral sciences. McGraw–Hill, New York, pp 75–83

Simmonds P, Smith DB (1999) Structural constraints on RNA virus evolution. J Virol 73:5787–5794

Snedecor GW, Cochran WG (1967) Statistical methods. Iowa State University Press, Ames, pp 215–217

Spiropoulou CF, Nichol ST (1993) A small highly basic protein is encoded in overlapping frame within the P gene of vesicular stomatitis virus. J Virol 67:3103–3110

Suzuki Y, Katayama K, Fukushi S, Kageyama T, Oya A, Okamura H, Tanaka Y, Mizokami M, Gojobori T (1999) Slow evolutionary rate of GB virus C/hepatitis G virus. J Mol Evol 48:383–389

Tanaka Y, Mizokami M, Orito E, Ohba K, Kato T, Kondo Y, Mboudjeka I, Zekeng L, Kaptue L, Bikandou B, M’Pele P, Takehisa J, Hayami M, Suzuki Y, Gojobori T (1998) African origin of GB virus C/hepatitis G virus. FEBS Lett 423:143–148

Ten Dam EB, Pleij CWA, Bosch L (1990) RNA pseudoknots: Translational frameshifting and readthrough on viral RNAs. Virus Genes 4:121–136

Thompson JD, Higgins DG, Gibson TJ (1994) Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Truant R, Cullen BR (1999) The arginine-rich domain present in human immunodeficiency tpe 1 Tat and Rev function as direct importin beta-dependent nuclear localization signals. Mol Cell Biol 19:1210–1217

Wright F (1990) The “effective number of codons” used in a gene. Gene 87:23–29

Xiang J, Klinzman D, McLinden J, Schmidt WN, LaBrecque DR, Gish R, Stapleton JT. (1998) Characterization of hepatitis G virus (GB-C virus) particles: Evidence for a nucleocapsid and expression of sequences upstream of the E1 protein. J Virol 72:2738–2744

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