JOURNAL OF VIROLOGY, Jan. 2003, p. 452–459 Vol. 77, No. 1
0022-538X/03/$08.00+0 DOI: 10.1128/JVI.77.1.452–459.2003

Genome Variability and Capsid Structural Constraints of Hepatitis A Virus

Glòria Sánchez, Albert Bosch,* and Rosa M. Pintó

Grup Virus Entèrics, Department of Microbiology, University of Barcelona, 08028 Barcelona, Spain

Received 17 June 2002/Accepted 24 September 2002

The number of synonymous mutations per synonymous site (K_s), the number of nonsynonymous mutations per nonsynonymous site (K_a), and the codon usage statistic (N_c) were calculated for several hepatitis A virus (HAV) isolates. While K_s was similar to those of poliovirus (PV) and foot-and-mouth disease virus (FMDV), K_a was 1 order of magnitude lower. The N_c parameter provides information on codon usage bias and decreases when bias increases. The N_c value in HAV was about 38, while in PV and FMDV, it was about 53. The emergence of 22 rare codons in front of 8 in PV and 7 in FMDV was detected. Most of the conserved rare codons of the P1 region were strategically located at the carboxy borders of β barrels and α helices, their potential function being the assurance of proper folding of the capsid proteins through a decrease in the translation speed. This strategic location was not observed for amino acids encoded by the conserved rare codons of the 3D region. The percentage of bases with low pairing number values was higher in the latter region, suggesting a role of the conserved rare codons in the maintenance of RNA structure. Many of the rare codons in HAV are among the most frequent in humans, unlike in PV or in FMDV. This fact may be explained by the lack of cellular shutoff in HAV. One hypothesis is that HAV has evolved in order to avoid competition with its host for cellular tRNAs.

The high degree of conservation of the amino acid sequences of the capsid proteins of hepatitis A virus (HAV) correlates with a lack of antigenic diversity; thus, there is only a single serotype of human HAV. However, despite this limited amino acid heterogeneity, a significant degree of nucleic acid variability has been observed among different isolates from different regions of the world (3, 8, 23, 25). The molecular bases of this genetic variability may be the high error rate of the viral RNA-dependent RNA polymerase and the absence of proofreading mechanisms. Although no data exist on the error rate of the HAV polymerase, the mutation frequencies for a variety of different RNA viruses range from 10^{-4} to 10^{-5} substitution per base per round of copying (9). The reason why this nucleic acid heterogeneity does not correspond to amino acid heterogeneity should rely on the lack of nonsynonymous mutations, possibly due to their elimination by negative selection. However, the actual mode of transmission of very small HAV populations, frequently associated with contaminated foods, may lead to the accumulation of debilitating mutations (7). In this context, the strikingly low level of amino acid changes in the capsid region suggests strong structural constraints.

In the present work, we undertook an analysis of the nucleotide and amino acid changes in sequences representing the available strains from GenBank and isolates from a food-borne hepatitis A outbreak. Since HAV structural data exist only for the 3C protein (4), structural models for the VP2, VP3, VP1, and 3D proteins have been deduced from actual data for the structural proteins of poliovirus (13) and foot-and-mouth disease virus (1) and from the actual 3D polymerase of poliovirus (12).

* Corresponding author. Mailing address: Department of Microbiology, School of Biology, University of Barcelona, Av. Diagonal 645, 08028 Barcelona, Spain. Phone: (34) 934034620. Fax: (34) 934034629. E-mail: abosch@ub.edu.

MATERIALS AND METHODS

Viruses. The 15 complete HAV sequences available at GenBank were used throughout this study. These sequences represent a group of geographically and temporally diverse HAV strains (Table 1). Additionally, 18 strains were isolated from patients in an outbreak of acute hepatitis A associated with the consumption of coquina clams (24). Virus RNA was isolated from 60-μl serum samples by guanidine thiocyanate treatment as specified elsewhere (5, 24).

RT-PCR amplification and nucleotide sequencing. The complete P1-2A sequence of the HAV isolates was obtained after their amplification with the *Pwo* reverse transcription (RT)-PCR system (Roche) by following the manufacturer's specifications and with primers corresponding to the capsid protein genomic regions (Table 2). Sequencing of RT-PCR products in both directions was performed with a Thermo Sequenase II dye terminator cycle sequencing premix kit (Amersham Pharmacia Biotech) by following the manufacturer's instructions and with an ABI Prism 377 automated DNA sequencer (Perkin-Elmer).

Analysis of nucleotide and amino acid sequences. Alignment of multiple sequences was carried out with the ClustalW program (European Bioinformatics Institute). The number of synonymous mutations per synonymous site (K_s) and the number of nonsynonymous mutations per nonsynonymous site (K_a) were calculated by the Nei-Gojobori method (19) with the DnaSP program (http://www.ub.es/dnasp/) (University of Barcelona).

To create codon usage tables for HAV, poliovirus serotype 1 (PV-1), and foot-and-mouth disease virus serotype C (FMDV-C), the Cusp program (European Molecular Biology Open Software Suite) was used. A rare codon was defined as one whose frequency was less than 30% that of its most abundant synonym in each of the codon usage tables (11). The effective codon usage statistic (N_c) measures the codon bias (26). The N_c value is always between 20 (when only one codon is effectively used for each amino acid) and 61 (when codons are used randomly). The N_c value was calculated with the Chips program (European Molecular Biology Open Software Suite).

For the rare codon location study, protein secondary structure wire plot models of VP2, VP3, and VP1 of HAV were calculated from a picornavirus alignment (16) in which actual structural data for the Mahoney strain of PV-1 (http://www.biochem.ucl.ac.uk/bsm/pdbsum/2plv/main.html) were added and aligned. The three-dimensional model for the HAV protomer was also deduced by Luo et al. (16; M. Luo, personal communication) from this alignment. To statistically confirm the locations or distributions of rare codons, a χ^2 analysis of frequencies was undertaken. The different proteins were divided into two regions: (i) the carboxy ends and borders of the highly structured elements (β

TABLE 1. Complete HAV sequences available at GenBank and used in the present study

Strain	Genotype	Geographical location	Date of isolation	Accession no.
LA	IA	United States	1975	K02990
HAS-15	IA	United States	1979	X15464
AH2	IA	Japan	1991	AB020565
AH1	IA	Japan	1992	AB020564
FH1	IA	Japan	1992	AB020567
AH3	IA	Japan	1993	AB020566
FH2	IA	Japan	1993	AB020568
FH3	IA	Japan	1994	AB020569
GBM	IA	Germany	1976	X75214
FG	IA	Italy	1988	X83302
MBB	IB	Northern Africa	1978	M20273
HM-175	IB	Australia	1976	M14707
HAF-203	IB	Brazil	1992	AF268396
L-A-1	IB			AF314208
SLF88	VII	Sierra Leone	1988	AY032861

barrels and α helices) and (ii) the remaining portions of the proteins. The so-called carboxy limits were defined as the third carboxy portion of the structural elements plus the five contiguous external residues. In some instances, when the β barrels or the α helices were shorter than three residues, only one internal residue was included; when there were fewer than five external joining residues before the next structural element, the totality of the joining region was included. As a cutoff value, only rare codons conserved in more than 50% of the sequences were included in this study. In the χ^2 test, the null hypothesis was the random distribution of the rare codons. The actual structures of the PV-1 and FMDV-C capsid proteins (http://www.biochem.ucl.ac.uk/bsm/pdbsum/2plv/main.html and http://www.biochem.ucl.ac.uk/bsm/pdbsum/1qgc/main.html), the actual structure of the recombinant HAV 3C protease (http://www.biochem.ucl.ac.uk/bsm/pdbsum/1hav/main.html), and a model of the HAV 3D polymerase deduced from the recombinant PV 3D protein (http://www.biochem.ucl.ac.uk/bsm/pdbsum/1rdr/main.html) were used for comparison analysis of the locations of the rare codons.

The pairing number (P-Num) values associated with different genomic regions of either the HM-175 strain of HAV or the Mahoney strain of PV-1 were obtained from http://www.bocklabs.wisc.edu/acp/.

Nucleotide sequence accession numbers. The P1-2A nucleotide sequences of the18 isolates from the clam-associated outbreak were deposited in GenBank and have been assigned accession numbers AF396391 to AF396408.

RESULTS

General characterization of the HAV isolates from the clam-associated hepatitis A outbreak. All 18 isolates belonged to genotype IB, although 6 out of the 18 samples had a conservative amino acid change (Val to Ile) at residue 72, in the middle of the VP3 sequence, involved in the immunodominant site (24). This mutation induces a loss of recognition by monoclonal antibody K34C8 and thus represents the occurrence of an antigenic variant. Two additional amino acid changes were detected, at position 40 of VP2 (Val to Ala in two strains) and position 28 of VP1 (Met to Val in one strain). The overall nucleotide homology among these 18 samples ranged from 97.85 to 100% in the P1-2A region, while the amino acid homology ranged from 99.75% to 100%.

TABLE 3. Analysis of mutations in HAV sequences

Genomic region (nucleotides)	K_s	K_a	K_s/K_a
VP0[a] (735–1469)	0.03806	0.00037	102.90
VP3[a] (1470–2207)	0.03832	0.00089	43.00
VP1-2A[a] (2208–3191)	0.05608	0.00015	373.90
P1-2A[a] (735–3191)	0.04767	0.00044	108.30
VP0[b] (735–1469)	0.28205	0.00499	56.52
VP3[b] (1470–2207)	0.29861	0.00272	109.80
VP1[b] (2208–3026)	0.29051	0.00244	119.06
P1[b] (735–3026)	0.29034	0.00336	86.40
2A[b] (3027–3242)	0.30325	0.01296	23.40
2B[b] (3243–3995)	0.32080	0.01009	31.80
2C[b] (3996–5000)	0.35559	0.01329	26.75
3A[b] (5001–5222)	0.24410	0.01291	18.90
3B[b] (5223–5291)	0.34555	0.00508	68.02
3C[b] (5292–5948)	0.27352	0.00472	57.95
3D[b] (5949–7415)	0.27100	0.01097	24.70
FMDV-C VP1[c]	0.29	0.03	9.70
FMDV-C P1[c]	0.28	0.02	14.00
PV-1 VP1[d]			2.50/18.40

[a] Hepatitis A outbreak sequence.
[b] HAV sequence from GenBank.
[c] From Martinez et al. (17) (accession numbers M84360, M90055, and M90367 to M90382 for VP1 and M90367, M90368, M90372, M90376, M90381, and L290061 for P1). (accession numbers AF238098, AF233099, AF233110, and AF233113).
[d] From Gavrilin et al. (11).

Frequencies and kinds of mutations in the capsid region. The K_s and K_a values were calculated for the two sets of HAV sequences in the capsid region (Table 3). The K_s values for the outbreak sequences were 7.5 times lower in the VP0 and VP3 capsid regions and 5 times lower in the VP1-2A region than the K_s values for the GenBank sequences, while the K_a values were 13.4, 3, and 16 times lower, respectively. These results indicate

TABLE 2. Primers and conditions used for RT-PCR amplification and sequencing of HAV

Target region	Sequence	Genomic position[a]	$[Mg^{2+}]$ (mM)	Hybridization temp (°C)
VP0 NH_2	CAGCTGGACTGTTCTTTGGG	648–667	2.0	55
VP0 NH_2	TCACCAGGAACCATAGCACAG	1198–1178	2.0	55
VP0 COOH	TACAATGAGCAGTTTGCTGT	1065–1084	2.0	55
VP0 COOH	GCTCTTGCATCTTCATAATTTG	1543–1522	2.0	55
VP3 NH_2	GGGACAGGAACTTCAGCTTATAC	1380–1402	2.0	50
VP3 NH_2	TCTACCTGAATGATATTTGG	1859–1840	2.0	50
Inner VP3	GTTATTCCAGTGGACCCATATT	1701–1722	2.0	50
Inner VP3	TCGTGTACCTATTCACTCTATA	2031–2010	2.0	50
VP3 COOH	TGTGCAGTGATGGATATTACAG	1938–1959	2.0	50
VP3 COOH	GTTGTTATGCCAACTTGGGGA	2287–2267	2.0	50
VP1 NH_2	AATGTTTATCTTTCAGCAAT	2136–2155	2.0	53
VP1 NH_2	ACAGCTCCAAGAGCAGTTTT	2751–2770	2.0	53
VP1 COOH	ATGGCCTGGTTTACTCCAG	2673–2691	2.5	45
VP1 COOH	CCCTTCATTTTCCTAGG	3229–3213	2.5	45

[a] In wild-type HM-175 (GenBank accession no. M14707).

TABLE 4. Percentages of codons with regard to the most abundant synonym in HAV and human cell codon usage tables

Amino acid	Codon[a]	Anticodons	% Occurrence in:		
			HAV		Human cells
			GenBank	Outbreak (P1-2A)	
Arg	AGA	UCU, UCI	100.0	100.0	88
	AGG	UCC, UCU	29.2	52.6	98
	CGC	GCG, GCI	2.7	0.0	100
	CGU	GCA, GCG, GCI	3.5	4.5	42
	CGA	GCU, GCI	3.2	0.0	48
	CGG	GCC, GCU	0.6	0.0	92
Leu	UUG	AAC, AAU	100.0	100.0	26
	UUA	AAU, AAI	52.0	58.4	12
	CUU	GAA, GAG, GAI	46.0	32.8	25
	CUG	GAC, GAU	25.3	35.0	100
	CUA	GAU, GAI	9.2	4.6	15
	CUC	GAG, GAI	7.0	10.4	47
Ser	UCU	AGA, AGG, AGI	100.0	100.0	71
	UCA	AGU, AGI	84.5	72.5	50
	AGU	UCA, UCG, UCI	31.8	18.2	50
	UCC	AGG, AGI	27.9	45.3	95
	UCG	AGC, AGU	4.7	11.3	22
	AGC	UCG, UCI	4.7	3.6	100
Thr	ACU	UGA, UGG, UGI	100.0	100.0	55
	ACA	UGU, UGI	90.8	95.7	63
	ACC	UGG, UGI	18.0	21.2	100
	ACG	UGC, UGU	4.5	6.7	29
Pro	CCU	GGA, GGG, GGI	100.0	100.0	78
	CCA	GGU, GGI	89.5	74.4	73
	CCC	GGG, GGI	20.0	21.0	100
	CCG	GGC, GGU	1.8	0.0	33
Ala	GCU	CGA, CGG, CGI	100.0	100.0	67
	GCA	CGU, CGI	60.7	73.2	48
	GCC	CGG, CGI	30.7	35.0	100
	GCG	CGC, CGU	1.4	4.8	25
Gly	GGA	CCU, CCI	100.0	100.0	67
	GGU	CCA, CCG, CCI	58.4	79.1	44
	GGG	CCC, CCU	34.4	42.7	68
	GGC	CCG, CCI	28.1	39.1	100
Val	GUU	CAA, CAG	100.0	100.0	34
	GUG	CAC, CAU	46.3	39.5	100
	GUA	CAU, CAI	17.7	8.0	19
	GUC	CAG, CAI	12.8	9.2	53
Lys	AAA	UUU, UUI	100.0	100.0	64
	AAG	UUC, UUI	58.8	56.5	100
Asn	AAU	UUA, UUG, UUI	100.0	100.0	73
	AAC	UUG, UUI	19.1	29.0	100
Gln	CAA	GUU, GUI	88.0	100.0	33
	CAG	GUC, GUU	100.0	91.6	100
His	CAU	GUA, GUG, GUI	100.0	100.0	65
	CAC	GUG, GUI	22.1	17.4	100
Glu	GAA	CUU, CUI	100.0	100.0	65
	GAG	CUC, CUU	76.7	92.2	100
Asp	GAU	CUA, CUG, CUI	100.0	100.0	75
	GAC	CUG, CUI	19.2	17.0	100
Tyr	UAU	AUA, AUG, AUI	100.0	100.0	66
	UAC	AUG, AUI	26.5	18.0	100
Cys	UGU	ACA, ACG, ACI	100.0	100.0	68
	UGC	ACG, ACI	26.5	12.0	100
Phe	UUU	AAA, AAG, AAI	100.0	100.0	70
	UUC	AAG, AAI	27.0	25.0	100
Ile	AUU	UAA, UAG, UAI	100.0	100.0	61
	AUA	UAU, UAI	31.7	29.0	24
	AUC	UAG, UAI	15.9	12.0	100

[a] Codons shown in bold type were rare in both sets of HAV sequences and were considered in the location study.

the higher divergence of the GenBank sequences than of the cluster of outbreak sequences. The K_s/K_a ratios for these outbreak sequences were extremely high for the VP0 and VP1-2A regions and considerably high for the VP3 region. The lower ratio observed for the VP3 region reflects the relatively abundant (6 of 18 sequences) occurrence of the antigenic variant described above. For the GenBank sequences, this ratio was also very high for all of the analyzed genomic regions. When the complete capsid regions (P1 regions) of HAV and FMDV-C were compared, it was found that while the K_s values were similar in both viruses, the K_a value was 1 order of magnitude higher in FMDV-C. Consequently, the K_s/K_a ratio was significantly higher in HAV (Table 3). A similar conclusion was drawn after a comparison of the values for the VP1 regions of HAV, FMDV-C, and PV-1 (Table 3).

Codon usage. Since synonymous mutations are the most prevalent in HAV and the occurrence of such mutations is subject to the influence of codon usage, an analysis of this usage in HAV was undertaken. The complete coding genome was studied for the GenBank sequences. Fifteen out of the 18 amino acid families containing synonymous codons showed the use of rare codons (Table 4). Overall, 25 rare codons were detected. Similarly, analysis of the codon usage in the P1-2A region of the sequences from the outbreak revealed the existence of 20 rare codons in 14 out of the 18 amino acid families (Table 4). Eighteen out of these 20 rare codons were in common with those of the GenBank sequences, and 4 more (Arg: CGC; Arg: CGA; Arg: CGG; Pro: CCG) were not found in the outbreak sequences. However, three (Arg: CGC; Arg: CGG; Pro: CCG) out of these four codons were not detected in the capsid region of the GenBank sequences. Two codons that were rare in the outbreak sequences were not rare in the GenBank group (Ser: AGU; Ile: AUA), while three codons that were rare in the GenBank sequences were not rare in the outbreak group (Arg: AGG; Ser: UCC; Leu: CUG).

No significant differences could be detected in the frequencies of rare codons between the capsid region and the nonstructural region of the genome among the GenBank sequences. These frequencies, defined as the number of rare codons versus the total number of codons, were 9.1% in the former genomic region and of 8.2% in the second. However, as mentioned above, some rare codons were absent from the capsid region. The overall heterogeneity could be expressed as the effective N_c values, which were 39 for the total coding region, 38.8 for the capsid region, and 38.7 for the nonstructural region. The N_c value for the capsid region in the outbreak sequences was 38.9. These values indicate that there was no significant difference in the codon usage bias among the closely related outbreak sequences and the GenBank sequences. The N_c values were also calculated for two other picornaviruses, PV-1 and FMDV-C (Table 5). For both of these viruses, markedly higher N_c values were obtained (52.6 and 52.1, respectively, compared to 37.2 for VP1 of HAV; 53.3 and 38.8 for the entire P1 regions of FMDV-C and HAV, respectively), indicating that HAV has a much higher bias in codon usage. This fact was further confirmed by the number of amino acid families containing rare codons or by the clearly higher total number of rare codons in HAV than in the other viruses (Table 5).

Additionally, another important and surprising difference among HAV and the other picornaviruses (PV-1 or FMDV-C)

TABLE 5. Codon usage for three picornaviruses

Virus	Effective N_c for:		No. of amino acid families with rare codons[a]	Total no. of rare codons[b]
	VP1	P1		
HAV	38.1	38.8	14	22
PV-1[c]	52.6		5	8
FMDV-C[d]	52.1	53.3	7	7

[a] The number of amino acid families containing synonymous codons was 18.
[b] The number of synonymous codons was 59.
[c] Data were calculated with sequences from Gavrilin et al. (11) (accession numbers AF238098, AF233099, AF233110, and AF233113).
[d] Data were calculated with sequences from Martinez et al. (17) (accession numbers M84360, M90055, and M90367 to M90382 for VP1 and M90367, M90368, M90372, M90376, M90381, and L290061 for P1).

could be observed. While the codon usage of PV-1 and FMDV-C was mostly coincident with that of their hosts, the HAV codon usage was quite antagonistic to that of human cells. The codons most abundantly used for Arg, Ser, Thr, Pro, Asn, His, Asp, Tyr, Cys, Phe, and Ile by human cells were rare codons for HAV. Leu, Ala, Gly, Val, and Lys codons were not as rare as the codons just listed but clearly were less abundant. Only 5 out of the 22 HAV rare codons were also rare in human cells (UCG, CUA, ACG, GCG, and GUA). Consequently, the most abundant HAV codons were not the most abundant human codons. Only Gln and Glu were mostly encoded by the same triplets. A closer analysis of this situation reveals that for most twofold degenerate amino acids, namely, Asn, Hys, Asp, Tyr, Cys, and Phe, one codon may bind two different tRNAs, each bearing a different anticodon, and the other codon may bind these same two tRNAs and a third tRNA with a third and different anticodon; human cells more frequently use the first of these codons, while HAV clearly uses the second, which contains the unshared anticodon (Table 4). For the three remaining twofold degenerate amino acids, Lys, Gln, and Glu, each codon may bind two tRNAs and only one anticodon is shared; no clear codon preference is shown by HAV. For the threefold degenerate Ile, the most abundant human codon may bind two tRNAs, whose anticodons are shared with the two alternative codons; the second most abundant codon may bind three anticodons (two shared and one unshared); and the least abundant codon may bind two anticodons (one shared and one unshared). In this situation, HAV chooses as the most abundant codon a codon that has one unshared anticodon and that is not rare for the host cell. The same strategy is observed for most of the four- and sixfold degenerate amino acids, although in some, such as Val and Leu, HAV selects a rare human codon as its most abundant, since the alternative codons with an unshared anticodon correspond to the most abundant human codons, thus representing strong competition.

Locations of rare codons. The locations of the rare codons in the GenBank sequences were studied, although codons considered rare were only those that were present and rare in both the GenBank and the hepatitis outbreak sequences (Table 4). Only rare codons conserved in more than 50% of the analyzed sequences were considered significant in the location study. Overall, 60 rare codons out of 764 total codons were conserved in the P1 region, representing 7.8% conserved rare codons. When the stringency in conservation was increased to either 85 or 100%, 2.7 or 0.9%, respectively, of the total P1 codons were rare codons. It should be noted that 7 out of these 60 rare codons (11.6%) were conserved in the entire group of sequences. In some instances, sequences lacking an individual rare codon had an alternative rare codon in close proximity (distance of one to three codons), increasing the percent conservation. Accordingly, 26 highly conserved positions that should be very critical were recognized in the structural genome.

A tendency to be located within the carboxy limits of the highly structured elements (β barrels and α helices) was observed (Fig. 1). Overall, 52.7% of the rare codons of the capsid region were located in the carboxy limits of the β barrels and α helices, while the residues contained in these limits represented 37.6% of the total polyprotein. This tendency to be located within the carboxy limits was statistically significant ($P < 0.05$). These same determinations were calculated for the P1 region of FMDV-C and for the VP1 region of PV-1. A significant nonrandom location could not be detected in either FMDV-C or PV-1, although in both viruses a preference for the carboxy limits was also observed.

At certain positions, the conserved rare codons were clustered (Fig. 1). A cluster was defined as a group of at least two contiguous rare codons or a group of at least two rare codons at a distance of one to three codons. In the VP0 region, three clear clusters were observed. The first was located in the VP4 region, for which no structural data exist. The second was located right at the carboxy border of the βD barrel, and the third extended all along the short joining sequence between the βF and the βG1 barrels. In the VP3 region, one cluster was detected starting just before the βG2 barrel and finishing just after this structure. In the VP1 region, one cluster was located at the amino terminus of the protein, and two contiguous clusters were located at the end of the βE barrel and at the joining sequence between the βE barrel and the αB helix. The outbreak sequences were not included in this analysis to avoid the potential bias due to their close relationships; however, they were used to confirm the existence of critical positions (Fig. 1). Eighty-one percent of the highly conserved rare codon locations detected in the GenBank sequences were also detected as highly conserved in the outbreak sequences. A high degree of correlation was also observed among the rare codon clusters, with the exception of those of VP1. However, since these clusters were not highly conserved in the GenBank sequences, they should not be regarded as actual clusters.

All of the amino acids encoded by these clusters were located in exposed regions of the capsid (Fig. 2), and these clusters included very rare codons. The frequencies of these very rare codons were below 20% that of the most common codon of their families or even below 5% in some instances, such as that of the βG2 cluster of VP3. It should be noted that not only the clusters but also several single rare codons encoded residues located on the surface. Overall, 67% of the total rare codons encoded residues exposed on the capsid surface, more precisely, 81, 72, and 64% of the VP2, VP3, and VP1 rare codons, respectively. On the other hand, 60% of these rare codons were highly conserved in at least 85% of the GenBank sequences.

The occurrences and locations of rare codons in the 3C and 3D regions of the GenBank sequences of HAV were comparatively analyzed. The 3C coding region contained 1.82, 1.37,

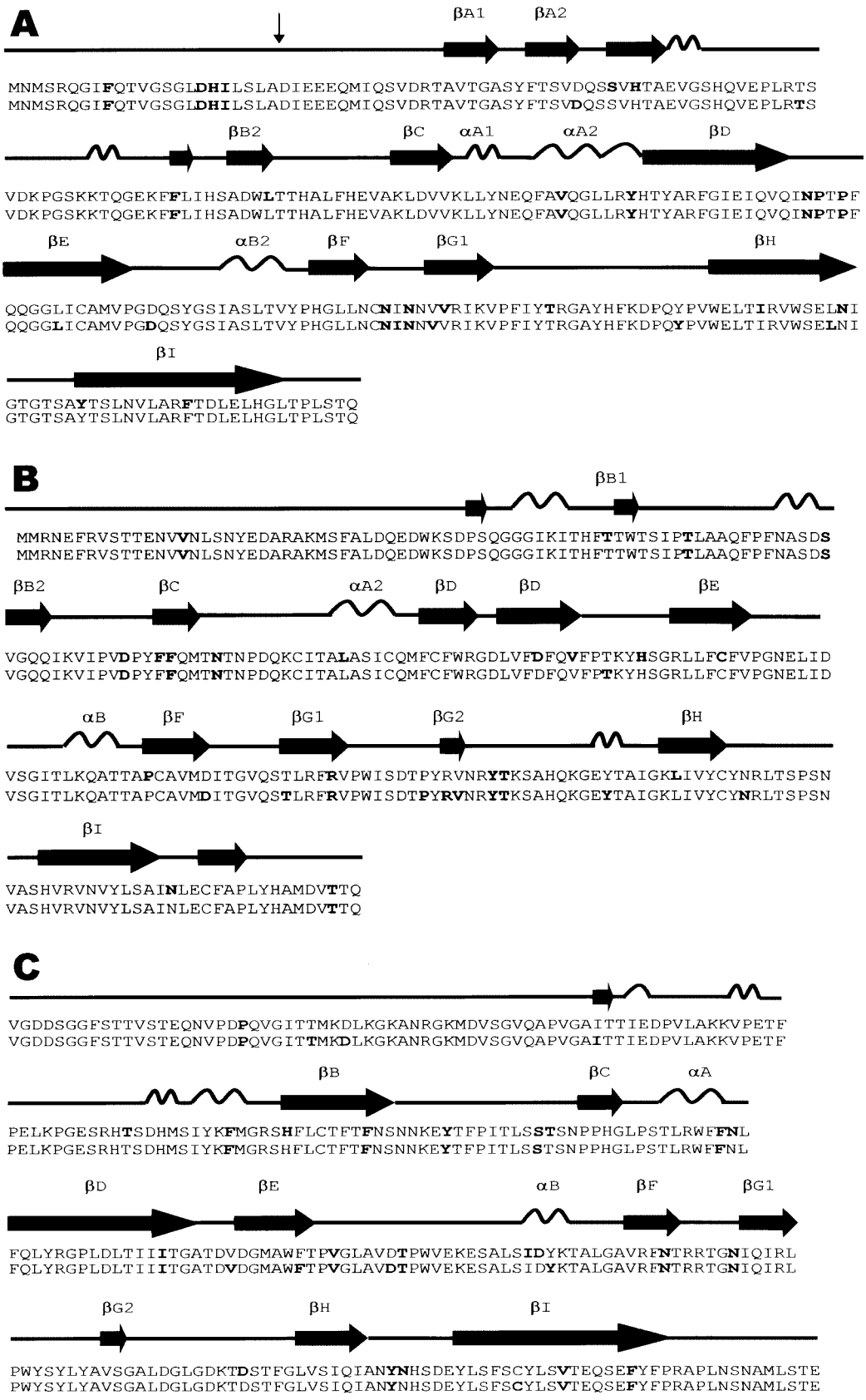

FIG. 1. VP2, VP3, and VP1 structural wire plot models deduced from actual data for PV-1 (Mahoney). (A) VP2. (B) VP3. (C) VP1. The first and second rows correspond to the outbreak and GenBank consensus sequences, respectively. Bold type indicates amino acids encoded by conserved rare codons.

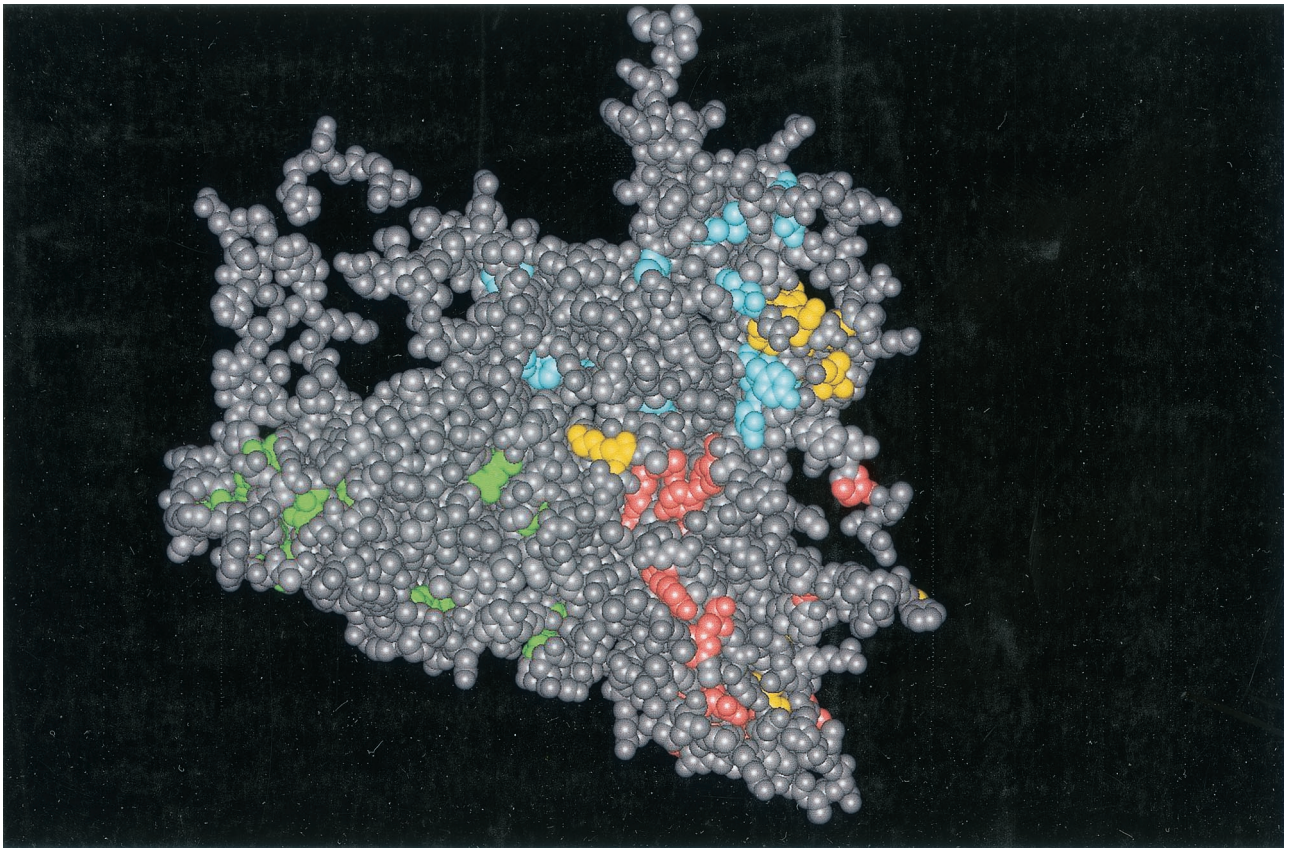

FIG. 2. Capsid surface location of amino acids encoded by conserved rare codons in an HAV protomer refined model (Luo, personal communication). Green, pink, and blue spheres indicate residues of VP2, VP3, and VP1, respectively, encoded by rare codons. Yellow spheres correspond to residues implicated in antigenic sites.

and 0.91% rare codons conserved in 50, 85, and 100% of the sequences, respectively. For the 3D region, 6.7% of the total codons were rare codons conserved in 50% of the sequences, 2.86% of the total codons were rare codons conserved in 85% of the sequences, and 1.84% of the total codons were rare codons conserved in 100% of the sequences. While the patterns of conservation of the rare codons were completely different in the 3C and P1 regions, that of the 3D region did not differ significantly from that of the P1 region. However, the strategic location of the HAV P1 region rare codons contrasted with the data obtained for the 3D region, whose rare codons were randomly distributed, instead of being accumulated at the carboxy limits. Overall, the carboxy limits represented 30.5% of the 3D protein, and 33.3% of the rare codons were located at these limits. Thus, it could be concluded that no clear preference for the carboxy limit location exists in the 3D polymerase. For the 3C protein, the statistical analysis was hampered by the low number of rare codons.

RNA secondary structure. Although the dynamic nature of the RNA genome avoids an accurate prediction of its secondary structure, P-Num values provide a quantitative estimation of the propensity of a base to pair with alternative partners in a collection of suboptimal folds (20). RNA regions having abundant bases with low P-Num values (P-Num, <100) are predicted to contain secondary structures (20). This parameter was calculated either for the total genome of the HM-175 strain of HAV or for partial RNA regions. Although the percentage of bases with P-Num values of <100 was 24.15%, a distinct pattern was observed among the different genomic regions (Table 6). Remarkably, significantly higher percentages of bases with low P-Num values were observed in the P3 region than in the P1 region and even than in the noncoding regions, suggesting a tighter structure in the P3 region. The ratios between the P-Num values of different regions were calculated for HAV and PV-1. The P3 region/P1 region ratios were 2.3 and 1.7, respectively, the P3 region/5′ noncoding region ratios were 1.6 and 1.7, respectively, and the P1 region/5′ noncoding region ratios were 0.7 and 1, respectively. These ratios suggested that the RNA of the HAV P1 region had a comparatively lower P-Num value and correspondingly a relatively looser structure.

DISCUSSION

HAV has low antigenic variability, as reflected by the existence of a single serotype (14). However, antigenic variants have been selected for their resistance to different MAbs (18, 21). Among the group of isolates from the clam-associated outbreak, a natural antigenic variant has been detected which induces a loss of recognition by MAb K34C8 and a second variation in a linear epitope of VP1 (24). However, the frequency of nonsynonymous mutations observed in HAV is sig-

TABLE 6. Effective N_c values, percentages of rare codons conserved in at least half of the GenBank sequences of HAV, and P-Num values for the HM-175 strain of HAV in association with different genomic regions

Genomic region	N_c	% of:	
		Conserved rare codons	Bases with P-Num values of <100[a]
Whole genome		6.55	24.1
5′ Noncoding region			23.4 (7.1)
P1	38.8	7.85	16.1 (7.1)
VP0	38.7	8.16	9.2
VP3	38.3	7.32	20.9
VP1	38.1	8.05	18.1
P2	39.0	6.37	18.4 (8.2)
2A	35.8	5.55	6.5
2B	34.3	5.55	27.0
2C	39.6	7.16	14.5
P3	37.9	5.46	37.0 (12.5)
3A	33.6	9.46	68.5
3B	29.1	4.35	94.2
3C	34.8	1.82	33.9
3D	37.1	6.74	30.9
3′ Noncoding region			11.1 (94.2)

[a] P-Num values for the PV-1 Mahoney strain are shown in parentheses.

nificantly lower than those found in other picornaviruses, such as PV-1 and FMDV. Overall, one position in VP4 (residue 5), four positions in VP2 (residues 40, 44, 89, and 180), six positions in VP3 (residues 32, 45, 72, 92, 145, and 239), and nine positions in VP1 (residues 25, 28, 148, 156, 174, 208, 216, 241, and 271) show variability in the GenBank and outbreak sequences. Assuming that the tolerance to amino acid substitutions is higher at surface protein sites free of structural constraints, it seems reasonable to assume that all of these residues are located at the capsid surface, even more so when it is considered that most of these substitutions are nonconservative. Since antigenic sites are frequently located at the surface, it can be expected that several of these substitutions are located at HAV antigenic sites. However, only three of these substitutions have been found directly involved in HAV epitopes, i.e., positions 25 and 28 of VP1 (10, 15) and position 72 of VP3 (18, 21, 24). Additionally, position 174 of VP1 is located in the middle of the sequence from residues 171 to 176 of this protein, which is part of the immunodominant site of HAV, although no data exist on the implication of this position for the epitope structure (18, 21). The same situation applies to residues 148, 156, and 241 of VP1 and residue 145 of VP3, which are located close to residues that are part of the HAV immunodominant site in the HAV structural model (16; Luo, personal communication). Furthermore, residues 208 and 216 of VP1 are located close to residue 221, which defines a second antigenic site of HAV (18, 21). However, despite the existence of these potential antigenic variants, different serotypes have not been defined, possibly because these substitutions lead to losses of only single epitopes from complex antigenic sites, as is the case for the variation in residue 72 of VP3 (24). It can then be expected that more extensive substitutions are required for the emergence of a new serotype and that such replacements are hampered by strong structural constraints.

The general idea of strong capsid structural constraints may be reinforced by the profuse use and conservation of rare codons, whose strategic locations have been postulated to create codon context variations along the mRNA, inducing a decrease in translation speed and allowing proper protein folding. The N_c values obtained for different picornaviruses indicate a significantly higher bias in codon usage in HAV than in PV-1 or FMDV-C. Genes with lower N_c values are considered to be restricted in the use of synonymous codons compared to genes with higher N_c values, which have greater flexibility in the use of synonymous codons (26). It may be assumed that low N_c values imply the use of preferred codons and, consequently, the existence of rare codons. There are indications that synonymous codon usage may be biased toward rare codons in segments connecting domains and regular secondary structure blocks (2, 11). In fact, a statistically significant nonrandom distribution of rare codons could be observed in the capsid region, the preferred locations being the carboxy ends and borders of the highly structured protein elements. In contrast, this tendency has not been observed in the 3C region or even in the 3D region, which is richer in 100% conserved rare codons than the P1 region. However, by using the P-Num value of the HM-175 strain as a reference for HAV, it was observed that for the 3D region, the RNA secondary structure is tighter; consequently, its sequence should be less prone to variability, as has been suggested for other viruses (11). Thirty percent and 64% of the P1 region and 3D region rare codons, respectively, were immersed in RNA regions with low P-Num values (data not shown). Consequently, it can be considered that these rare codons play a dual role in maintaining both the RNA and the protein structures, being more important at the protein level in the P1 region and at the RNA level in the 3D region.

The occurrence of surface residues encoded by rare codons, which account for approximately 15% of the total surface residues (Fig. 2 and data not shown), could contribute to the low variability of the HAV capsid, since it is quite unlikely that the occurrence of a nucleotide substitution in a rare codon will give rise to a new codon of similar rarity (Table 4), in order to maintain the translation kinetics for correct folding without a loss of efficiency. In fact, among the previously mentioned capsid substitutions, only that at position 25 of VP1 (Ile to Met) changed from quasi-rare to non-rare and that at position 271 of VP1 (Ser to Pro) changed from quasi-rare to unmistakably rare. All of the other substitutions affected non-rare codons.

An intriguing issue, however, is the antagonism in the codon usage of HAV and human cells, since the availability of tRNAs is host dependent. The picornavirus models used throughout this study showed codon usage very similar to that of their hosts. However, this situation implies the occurrence of competition for tRNAs, among other factors. For PV-1 and FMDV-C, this competition is avoided by the induction of cellular shutoff of protein synthesis through carboxy cleavage of component eIF4G of the translation initiation complex by 2A and L proteases, respectively (22). The cleaved eIF4G factor is still active for the internal ribosome entry site-dependent initiation of translation of most picornaviruses, although that of HAV requires an intact eIF4G factor (6). The latter is a plausible explanation for why shutoff has not been described for HAV-infected cells. Consequently, HAV competes poorly for cellular factors, among them tRNAs; therefore, the most abun-

dant codons of its host are not its most abundant and, in several instances, even are rare codons.

The lack of a specific shutoff-inducing mechanism and the occurrence of long extracorporeal periods are concordant with a special codon usage which prevents direct competition with the host cell system and concomitantly allows a highly compact capsid that ensures a high level of environmental persistence.

ACKNOWLEDGMENTS

We acknowledge the skillful assistance of Àngels Rabassó and the technical expertise of the Serveis Científic-Tècnics of the University of Barcelona.

This study was supported in part by grants ERB3514PL973098, QLRT-1999-0634, and QLRT-1999-0594 from the European Union; BIO99-0455 from the CICYT, Ministry of Science and Technology, Spain; and 1997SGR 00224 from the Generalitat de Catalunya.

REFERENCES

1. **Acharya, R., E. Fry, D. Stuart, G. Fox, D. Rowlands, and F. Brown.** 1989. The three dimensional structure of foot-and-mouth disease virus at 2.8 Å resolution. Nature **337:**709–716.
2. **Adzhubei, A. A., I. A. Adzhubei, I. A. Krasheninnikov, and S. Neidle.** 1996. Non-random usage of "degenerate"codons is related to protein three-dimensional structure. FEBS Lett. **399:**78–82.
3. **Arauz-Ruiz, P., L. Sundqvist, Z. García, L. Taylor, K. Visoná, H. Norder, and L. O. Magnius.** 2001. Presumed common source outbreaks of hepatitis A in an endemic area confirmed by limited sequencing within the VP1 region. J. Med. Virol. **65:**449–456.
4. **Bergmann, E. M., S. C. Mosimann, M. M. Chernaia, B. A. Malcolmand M. N. James.** 1997. The refined crystal structure of the 3C gene product from hepatitis A virus: specific proteinase activity and RNA recognition. J. Virol. **71:**2436–2448.
5. **Boom, R., C. J. A. Sol, M. M. M. Salimans, C. L. Jansen, P. M. E. Wertheim-van Dillen, and J. Van der Noordaa.** 1990. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. **28:**495–503.
6. **Borman, A. M., and K. M. Kean.** 1997. Intact eukaryotic initiation factor 4G is required for hepatitis A virus internal initiation of translation. Virology **237:**129–136.
7. **Chao, L.** 1990. Fitness of RNA virus decreased by Muller's ratchet. Nature **348:**454–455.
8. **Costa-Mattioli, M., V. Ferre, S. Monpoeho, L. García, R. Colina, S. Billaudel, I. Vega, R. Pérez-Bercoff, and J. Cristina.** 2001. Genetic variability of hepatitis A virus in South America reveals heterogeneity and co-circulation during epidemic outbreaks. J. Gen. Virol. **82:**2647–2652.
9. **Domingo, E.** 1996. Biological significance of viral quasispecies. Viral Hep. Rev. **2:**247–261.
10. **Emini, E. A., J. V. Hughes, D. S. Perlow, and J. Boger.** 1985. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J. Virol. **55:**836–839.
11. **Gavrilin, G. V., E. A. Cherkasova, G. Y. Lipskaya, O. M. Kew, V. I. Agol.** 2000. Evolution of circulating wild poliovirus and of vaccine-derived poliovirus in an immunodeficient patient: a unifying model. J. Virol. **74:**7381–7390.
12. **Hansen, J. L., A. M. Long, and S. C. Shultz.** 1997. Structure of the RNA-dependent RNA polymerase of poliovirus. Structure **5:**1109–1122.
13. **Hogle, J. M., M. Chow, and D. J. Filman.** 1985. Three-dimensional structure of poliovirus at 2.9 Å resolution. Science **229:**1358–1365.
14. **Hollinger, F. B., and S. U. Emerson.** 2001. Hepatitis A virus, p. 799–840. *In* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Straus (ed.), Fields virology, 4th ed., vol. 1. Lippincott Williams & Wilkins, Philadelphia, Pa.
15. **Ivanov, V. S., L. N. Kulik, A. E. Gabrielian, L. D. Tehikin, A. T. Kozhich, and V. T. Ivanov.** 1994. Synthetic peptides in the determination of hepatitis A T-cell epitopes. FEBS Lett. **345:**159–161.
16. **Luo, M., M. G. Rossmann, and A. C. Palmenberg.** 1988. Prediction of three-dimensional models for foot-and-mouth disease virus and hepatitis A virus. Virology **166:**503–514.
17. **Martinez, M. A., J. Dopazo, J. Hernandez, M. G. Mateu, F. Sobrino, E. Domingo, and N. J. Knowles.** 1992. Evolution of the capsid protein genes of foot-and mouth disease virus: antigenic variation without accumulation of amino acid substitutions over six decades. J. Virol. **66:**3557–3565.
18. **Nainan, O. V., M. A. Brinton, and H. S. Margolis.** 1992. Identification of amino acids located in the antibody binding sites of human hepatitis A virus. Virology **191:**984–987.
19. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the number of synonymous and non synonymous nucleotide substitutions. Mol. Biol. Evol. **3:**418–426.
20. **Palmemberg, A. C., and J.-Y. Sgro.** 1997. Topological organization of picornaviral genomes: statistical prediction of RNA structural signals. Semin. Virol. **8:**231–241.
21. **Ping, L.-H., and S. M. Lemon.** 1992. Antigenic structure of human hepatitis A virus defined by analysis of escape mutants selected against murine monoclonal antibodies. J. Virol. **66:**2208–2216.
22. **Racaniello, V. R.** 2001. *Picornaviridae*: the viruses and their replication, p. 685–722. *In* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Straus (ed.), Fields virology, 4th ed., vol. 1. Lippincott Williams & Wilkins, Philadelphia, Pa.
23. **Robertson, B. H., R. W. Jansen, B. Khanna, A. Totsuka, O. V. Nainan, G. Siegl, A. Widell, H. S. Margolis, S. Isomura, K. Ito, T. Ishiku, Y. Moritsugu, and S. M. Lemon.** 1992. Genetic relatedness of hepatitis A virus strains recovered from different geographical regions. J. Gen. Virol. **73:**1365–1377.
24. **Sánchez, G., R. M. Pintó, H. Vanaclocha, and A. Bosch.** 2002. Molecular characterization of hepatitis A virus isolates from a transcontinental shellfishborne outbreak. J. Clin. Microbiol. **40:**4148–4155.
25. **Taylor, M. B.** 1997. Molecular epidemiology of South African strains of hepatitis A virus: 1982–1996. J. Med. Virol. **51:**273–279.
26. **Wright, F.** 1990. The effective number of codons used in a gene. Gene **87:**23–29.