OPIC
OFFICE DE LA PROPRIÉTÉ INTELLECTUELLE DU CANADA

CIPO
CANADIAN INTELLECTUAL PROPERTY OFFICE

(12) (19) (CA) **Demande-Application**

(21) (A1) **2,229,143**
(86) 1996/08/14
(87) 1997/02/27

(72) SIMONS, John N., US
(72) DESAI, Suresh M., US
(72) MUSHAHWAR, Isa K., US
(71) ABBOTT LABORATORIES, US

(51) Int.Cl.^{6} C12N 15/67, A61K 47/48, A61K 31/535, A61K 31/70
(30) 1995/08/14 (60/002,265) US
(30) 1995/12/21 (08/580,038) US
(30) 1996/04/19 (08/639,857) US

(54) **REACTIFS ET PROCEDES PERMETTANT DE MODULER LA TRADUCTION DE PROTEINES DE L'HEPATITE GBV**

(54) **REAGENTS AND METHODS USEFUL FOR CONTROLLING THE TRANSLATION OF HEPATITIS GBV PROTEINS**

(57) Ces réactifs et une composition permettant de moduler la traduction des peptides du virus de l'hépatite GB (VHGB)-A, -B ou -C à partir d'un acide nucléique viral. Ces réactifs et procédés concernent des éléments de modulation relevant de la région 5'NTR du génome viral de VHGB-A, -B ou -C.

(57) Reagents and composition for controlling the translation of hepatitis GB virus (HGBV)-A, -B or -C peptides from viral nucleic acid. These reagents and methods comprise control elements of the 5'NTR region of the HGBV-A, -B, or -C viral genome.

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :

C12N 15/86, 15/11, 15/40, 15/51, C07K 14/18 // C12Q 1/68

A1

(11) International Publication Number: **WO 97/07224**

(43) International Publication Date: 27 February 1997 (27.02.97)

(21) International Application Number: PCT/US96/13198

(22) International Filing Date: 14 August 1996 (14.08.96)

(30) Priority Data:

60/002,265	14 August 1995 (14.08.95)	US
08/580,038	21 December 1995 (21.12.95)	US
08/639,857	19 April 1996 (19.04.96)	US

(71) Applicant: ABBOTT LABORATORIES [US/US]; CHAD 0377/AP6D-2, 100 Abbott Park Road, Abbott Park, IL 60064-3500 (US).

(72) Inventors: SIMONS, John, N.; 738 N. Allegheny Road, Grayslake, IL 60030 (US). DESAI, Suresh, M.; 1408 Amy Lane, Libertyville, IL 60048 (US). MUSHAHWAR, Isa, K.; 18790 Arbor Boulevard, Grayslake, IL 60030 (US).

(74) Agents: POREMBSKI, Priscilla, E. et al.; Abbott Laboratories, CHAD 0377/AP6D-2, 100 Abbott Park Road, Abbott Park, IL 60064-3500 (US).

(81) Designated States: CA, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published

With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: REAGENTS AND METHODS USEFUL FOR CONTROLLING THE TRANSLATION OF HEPATITIS GBV PROTEINS

(57) Abstract

Reagents and composition for controlling the translation of hepatitis GB virus (HGBV)-A, -B or -C peptides from viral nucleic acid. These reagents and methods comprise control elements of the 5'NTR region of the HGBV-A, -B, or -C viral genome.

REAGENTS AND METHODS USEFUL FOR CONTROLLING THE TRANSLATION OF HEPATITIS GBV PROTEINS

Related Applications

This application is a continuation-in-part of U.S. Serial No. 08/580,038, filed December 21, 1995, which is a continuation-in-part application of, and claimed the benefit of, U.S. provisional application Serial No. 60/002,265 filed August 14, 1995, which is related to patent application U.S. Serial No. 60/002,255 filed August 14, 1995, which is related to patent applications U.S. Serial No. 08/480,995 filed June 7, 1995, U.S. Serial No. 08/473,475 filed June 7, 1995 and U.S. Serial No. 08/417,629, filed April 6, 1995, which are continuation-in-part applications of U.S. Serial No. 08/424,550 filed June 5, 1995, which is a continuation-in-part application of U.S. Serial No. 08/377,557 filed January 30, 1995, which is a continuation-in-part of U.S. Serial No. 08/344,185 filed November 23, 1994 and U.S. Serial No. 08/344,190 filed November 23, 1994, which are each continuation-in-part applications of 08/283,314 filed July 29, 1994, which is a continuation-in-part application of U.S. Serial No. 08/242,654, filed May 13, 1994, which is a continuation-in-part application of U.S. Serial No. 08/196,030 filed February 14, 1994, all of which enjoy common ownership and each of which is incorporated herein by reference.

Background of the Invention

This invention relates generally to the family of hepatitis GB viruses (HGBV) and more particularly, relates to reagents such as antisense nucleic acid sequences and methods utilizing these nucleic acid sequences which are useful for controlling translation of HGBV-A, -B, or -C, both *in vivo* and *in vitro*, by either increasing or decreasing the expressions of HGBV-A, -B or -C proteins.

Recently, a new family of flaviviruses detected in patients with clinically diagnosed hepatitis was reported. This new family of viruses has been named the "GB" viruses, after the initials of the patient first infected with the virus. These viruses have been reported by J. N. Simons et al., Proc. Natl. Acad. Sci. USA 92:3401-3405 (1995); and J. N. Simons et al., Nature Medicine 1(6):564-569 (1995). Three members of the family have been identified to date: GBV-A, GBV-B and GBV-C. T. P. Leary, et al., J. Med. Virol. 48:60-67 (1995) While HGBV-A appears at this time to be of non-human primate source, HGBV-C is clearly of human source. Currently, the source of HGBV-B is unknown. These viruses are thought to play a role in transmittable hepatitis disease of viral origin.

The GB viruses appear to be members of the *Flaviviridae* family. They possess RNA genomes approximately 9.5 kb in size which contain a single long open reading frame (ORF). Structural and nonstructural proteins are encoded in the N-terminal one-third and C-terminal two-thirds of the putative viral polyproteins, respectively. Phylogenetic analyses of the nonstructural helicase and replicase genes demonstrate that these viruses are related to, but distinct from, the HCV genus of the *Flaviviridae*. See, for example, T. P. Leary, et al., supra and A. S. Muerhoff et al., J. Virol. 69:5621-5630 (1995). Specifically, GBV-A and GBV-C appear most closely related as they share a common ancestor, while the GBV-A/C ancestor, GBV-B and HCV all appear to be equally divergent from other members of the *Flaviviridae*.

However, when the 5' nontranslated regions (NTRs) and structural genes are examined, a more striking division between the GB viruses and the other members of the *Flaviviridae* becomes apparent. GBV-B appears similar to the HCV and pestivirus genera of the *Flaviviridae*. Conserved sequences present in the 5' NTRs of HCV and pestiviruses are found in the 5' NTR of GBV-B, and GBV-B and HCV share closely related RNA secondary structures within the 5' NTR. (M. Honda, E. A. Brown and S. M. Lemon, manuscript submitted). Moreover, a basic (pI = 11.1) core protein is present at the N-terminus of the GBV-B putative polyprotein precursor, and two putative envelope glycoproteins with several potential N-linked glycosylation sites are located downstream of core in GBV-B. A. S. Muerhoff et al., supra. These structural proteins appear in all members of the *Flaviviridae* examined to date. See, for example, M. S. Collett et al., J. Gen. Virol. 69:2637-2643 (1989) and R. H. Miller and R. H. Purcell, Proc. Natl. Acad. Sci. USA 87:2057-2061 (1990).

In contrast to GBV-B, examination of GBV-A and GBV-C reveals marked differences between these viruses and other genera of the *Flaviviridae*. GBV-A and GBV-C contain long 5' NTRs that have limited sequence identity at the 5'NTR to each other but no identity to the 5' NTRs GBV-B, HCV or pestiviruses. GBV-A and GBV-C also encode putative envelope proteins that contain relatively few potential N-linked glycosylation sites. Most strikingly, clearly discernible basic core proteins are not found in the cDNA sequences cloned thus far from these viruses.

The absence of core proteins would distinguish GBV-A and GBV-C from other genera of the *Flaviviridae*. However, several important aspects of the structure of the GBV-A and GBV-C genomes remain undefined. Primary among these is the identification of the AUG codons at which translation of the viral polyproteins initiate. The sequence of GBV-A contains two potential in-frame initiator AUG codons 27 nucleotides (9 amino acids) and immediately upstream of the putative E1 signal sequence. Similarly, multiple GBV-C sequences possess two to three potential in-frame initiator AUG codons. *See,* T. P. Leary et al., supra; and J. Linnen et al., Science 271:505-508 (1996). However, none of these AUGs have been demonstrated to serve as the initiator codon, and initiation at any of these sites would result in a severely truncated core protein at best. It is conceivable that deletions during the cloning of these virus RNAs could have resulted in the elimination of core sequences or a disruption of the true ORF in this region of the genome, as suggested by Leary et al., supra. However, multiple RT-PCR products generated from the 5' ends of GBV-A and GBV-C using a variety of primers, polymerases and conditions (unpublished data), in addition to determining the 5' end sequences of over 35 separate GBV-C isolates (U.S. Serial No. 08/580,038, filed December 21, 1995, previously incorporated herein by reference) provide no support for the existence of additional sequence missing from the previously described cDNA clones. Thus, it is possible that the 5' ends of these viruses are complete (or nearly complete), and that GBV-A and GBV-C do not encode core proteins.

Of the genera that comprise the *Flaviviridae*, the viruses classified in the flaviviruses genus (e.g., yellow fever virus, dengue virus) contain relatively short 5' NTRs of 97 to 119 nucleotides. In these viruses, translational initiation is thought to utilize a conventional eukaryotic ribosome scanning mechanism in which ribosomes bind the RNA at a 5' cap structure and scan in a 3' direction until encountering an AUG codon in a favorable context for initiation. *See,* M. Kozak, Cell 44:283-292 (1989) and M. Kozak, J. Cell. Biol. 108:229-241 (1989).

In contrast to the flavivirus genus, genomic RNAs from members of the pestiviruses and HCV genera contain relatively long 5' NTRs of 341 to 385 nucleotides which in some ways are similar to those of picornaviruses. Extensive studies of the picornavirus 5' NTRs reveal that translation initiation occurs through a mechanism of internal ribosome entry. R.J. Jackson et al., Mol. Biol. Reports 19:147-159 (1994); K. Meerovitch and N. Sonnenberg, Semin. Virol. 67:3798-3807 (1993). This internal entry requires a defined segment of the viral 5' NTR known as an "internal ribosome entry site" (IRES) or "ribosome landing pad." The RNA comprising the cis-acting IRES forms highly ordered structures which interact with trans-acting cellular translation factors to bind the 40S ribosome subunit at an internal site on the viral message, often many hundreds of nucleotides downstream of the 5' end of the molecule. Such translation initiation functions in a 5' cap-independent fashion, and is generally not influenced by structure or sequence upstream of the IRES.

Practically, the ability of a sequence to function as an IRES is assessed by insertion of the sequence between two cistrons of a bicistronic RNA. If the intercistronic sequence contains an IRES, there is significant translation of the downstream cistron which is generally independent of the translational activity of the upstream cistron. Studies of the 5' NTRs of HCV and pestiviruses using bicistronic mRNAs demonstrate the presence of IRESs in these sequences. *See*, for example, T. L. Poole et al., Virology 206:750-754 (1995); R. Rijnbrand et al., FEBS Letters 365:115-119 (1995); K. Tsukiyama-Kohara et al., J. Virol. 66:1476-1483 (1992); and C. Wang et al., J. Virol. 67:3338-3344 (1993).

Structural changes in the IRES influence the rate of translation initiation. Thus, by modifying a virus's IRES, one can control the amount of viral protein being made. Control of the translation process of the nucleic acids of GB viruses could provide an effective means of treating viral disease. The ability to control translation could result in a decrease of the expression of viral proteins. Also, the ability to increase expression may prove useful by producing greater amounts of GB viral proteins which could be utilized in a variety of ways, both diagnostically and therapeutically. Further, the ability to increase translation of the GB viruses *in vivo* may provide a means for increasing immune stimulation in an individual.

It therefore would be advantageous to provide reagents and methods for controlling the translation of HGBV proteins from HGBV nucleic acids. Such reagents would comprise antisense nucleic acid sequences or other compound which may specifically destabilize (or stabilize) the IRES structure. Such nucleic acid sequences or compounds could greatly enhance the ability of the medical community to provide a means for treating an individual infected with GB virus(es). In addition, IRESs are among the most highly consereved nucleotide sequences. Identification of such a sequence immediately suggests a target for probe-based detection reagents. Diagnostic or screening tests developed from these reagents could provide a safer blood and organ supply by helping to eliminate GBV in these blood and organ donations, and could provide a better understanding of the prevalence of HGBV in the population, epidemiology of the disease caused by HGBV and the prognosis of infected individuals. Additionally, these consereved structures may provide a means for purifying GBV proteins for use in diagnostic assays.

Summary of the Invention

The present invention provides unique reagents comprising nucleic acid sequences for HGBV-A, -B or -C that are useful for controlling the translation of HGBV nucleic acids to proteins. These nucleic acid sequences may be DNA or RNA, derivatized DNA or RNA, PNA in both the antisense or sense orientations.

The present invention also provides a method for controlling the translation of HGBV nucleic acids to HGBV proteins, comprising contacting a first nucleic acid sequence with HGBV nucleic acid sequence under conditions which permit hybridization of the first nucleic acid sequence and the HGBV nucleic acid sequence, and altering the level of translation of the HGBV nucleic acid. The first nucleic acid sequence is an antisense nucleic acid sequence which is substantially complementary to a sequence of the sense strand within the 5' NTR region of the HGBV nucleic acid sequence. The sense strand is of genomic or messenger RNA that is subjected to the translation process. The method described herein is performed in an individual infected with HGBV.

The present invention also provides a method of enhancing the translation of a nucleic acid comprising operably linking a nucleic acid with a nucleic acid having a seeequence corresponding to the sequence of GBV-A, -B or -C 5' region, to form a combined nucleic acid capable of being translated.

Further, the invention herein provides a composition for enhancing the translation of a nucleic acid, which composition comprises a nucleic acid having a sequence corresponding to the squence of GBV-A, -B, or -C 5' region, for operable linkage to nucleic acid to be translated. Further, a composition for contrlling translation of hepatitis GB virus -A, -B, or -C from GBvirus -A, -B or -C nucleic acids is provided, which comprises a first non-naturally occurring nucleic acid having a sequence complementary to, or capable of being transcribed to form, a nucleic acid having a sequence complementary to, a sequence of the sense strand within the 5'-NTR region of HGBV-A, -B, or -C, wherein said first nucleic acid comprises a sequence selected from the 5' NTR region of GBV-A, -B, or -C, and a cleavage are at which the full length GBV-A, -B, or -C RNA is cleaved to form a subgenomic HGBV-A, -B, or -C RNA. The first nucleic acid can be a nucleic acid analog, and it can be linked to a cholesteryl moiety at the 3' end.

Brief Description of the Drawings

FIGURE 1 presents the alignment of GBV-A and GBV-C 5' sequences and amino acid alignment of their respective ORF's. The putative E1 signal sequence in GBV-C and the Asn-Cys-Cys motif are underlined.

FIGURE 2A presents a schematic representation of monocistronic T7 templates, wherein viral RNA sequence is represented as a bold line, the positions of the AUG codons (AUG) and ORFs (box) are indicated;

FIGURE 2B shows a PhosphorImager scan of products generated by IVTT reactions programmed with pA15-707/CAT (A15-707, lane 1), pC1-631/CAT (C1-631, lane 2), pHAV-CAT1 (HAV, lane 3), pC631-1/CAT (C631-1, lane 4), SspI-linearized pA15-707/CAT (A15-707-SspI, lane 5) and pC1-631/CAT (C1-631-SspI, lane 6).

FIGURE 3A presents the organization of site-specific mutants of GBV-CAT monocistronic templates;

FIGURE 3B shows a PhosphorImager scan of IVTT products generated from GBV-CAT mutant templates, wherein Lanes 1, 4 and 7 are control reactions programmed with pA15-707/CAT, pC1-631/CAT and pHAV-CAT1, respectively; products generated from reactions programmed with the mutant templates are found in lanes 2 (pAmut1/CAT), 3 (pAmut2/CAT), 5 (pCmut1/CAT) and 6 (pCmut2/CAT).

FIGURES 4A AND 4B show an Edman degradation of ^{3}H-Leu-labeled GBV-CAT fusion products, wherein IVTT reactions programmed with pA15-707/CAT are presented in 4A and those programmed with pC1-631/CAT are presented in 4B.

FIGURES 5A and 5B show the translation of monocistronic RNAs containing 3' GBV deletions. FIGURE 5A presents a schematic of the monocistronic templates and FIGURE 5B presents a PhosphorImager scan of IVTT products generated with pA15-665/CAT (lane 2), pA15-629/CAT (lane 3), pA15-596/CAT (lane 4), pC1-592/CAT (lane 6), pC1-553/CAT (lane 7) and pC1-526/CAT (lane 8). Control reactions are shown in lanes 1 (pA15-707/CAT), 5 (pC1-631/CAT) and 9 (pHAV-CAT1).

FIGURES 6A and 6B show the translation of bicistronic GBV and HCV vectors, wherein FIGURE 6A presents a schematic of the bicistronic T7 templates, and FIGURE 6B presents the luciferase activity (Luc-A, light units X 10^{-3}), luciferase protein (Luc-P, band volume X 10^{-3}) and protein production of IVTTs programmed with the bicistronic vectors.

FIGURE 7A, B, C and D presents a schematic that depicts a preliminary model of the secondary RNA structures which are present near the 5' end of the GBV-C genome (GenBank accession no. U36380) (SEQUENCE ID NO 3), wherein major putative structural domains are labeled I - V with roman numerals; base pairs which are sites of covariant nucleotide substitutions in different strains of GBV-C are shown in boxes; the putative initiator AUG codon (first in-frame AUG codon which is conserved in all GBV-C sequences) is located between domains IV and V (highlighted bases); (Inset) presents the preliminary model of GBV-A domain V; and covariance between GBV-A and sequences from GBV-A-like viruses found indigenous to tamarins are boxed.

Detailed Description of the Invention

The present invention provides reagents and methods useful for controlling the translation of HGBV-A, HGBV-B or HGBV-C nucleic acid to protein.

The term "Hepatitis GB Virus" or "HGBV", as used herein, collectively denotes a viral species which causes non-A, non-B, non-C, non-D, non-E hepatitis in man, and attenuated strains or defective interfering particles derived therefrom. This may include acute viral hepatitis transmitted by contaminated foodstuffs, drinking water, and the like; hepatitis due to HGBV transmitted via person to person contact (including sexual transmission, respiratory and parenteral routes) or via intraveneous drug use. The methods as described herein will allow the treatment of individuals who have acquired HGBV. Individually, the HGBV isolates are specifically referred to as "HGBV-A", "HGBV-B" and "HGBV-C." As described herein, the HGBV genome is comprised of RNA. Analysis of the nucleotide sequence and deduced amino acid sequence of the HGBV reveals that viruses of this group have a genome organization similar to that of the *Flaviridae* family. Based primarily, but not exclusively, upon similarities in genome organization, the International Committee on the Taxonomy of Viruses has recommended that this family be composed of three genera: Flavivirus, Pestivirus, and the hepatitis C group. Similarity searches at the amino acid level reveal that the hepatitis GB virus subclones have some, albeit low, sequence resemblance to hepatitis C virus. It now has been demonstrated that HGBV-C is not a genotype of HCV. See, for example, U.S. Serial No. 08/417,629, filed April 6, 1995, previously incorporated herein by reference.

The term "similarity" and/or "identity" are used herein to describe the degree of relatedness between two polynucleotides or polypeptide sequences. The techniques for determining amino acid sequence "similarity" and/or "identity" are well-known in the art and include, for example, directly determining the amino acid sequence and comparing it to the sequences provided herein; determining the nucleotide sequence of the genomic material of the putative HGBV (usually via a cDNA intermediate), and determining the amino acid sequence encoded therein, and comparing the corresponding regions. In general, by "identity" is meant the exact match-up of either the nucleotide sequence of HGBV and that of another strain(s) or the amino acid sequence of HGBV and that of another strain(s) at the appropriate place on each genome. Also, in general, by "similarity" is meant the exact match-up of amino acid sequence of HGBV and that of another strain(s) at the appropriate place, where the amino acids are identical or possess similar chemical and/or physical properties such as charge or hydrophobicity. The programs available in the Wisconsin Sequence Analysis Package, Version 8 (available from the Genetics Computer Group, Madison, Wisconsin, 53711), for example, the GAP program, are capable of calculating both the identity and similarity between two polynucleotide or two polypeptide sequences. Other programs for calculating identity and similarity between two sequences are known in the art.

Additionally, the following parameters are applicable, either alone or in combination, in identifying a strain of HGBV-A, HGBV-B or HGBV-C. It is expected that the overall nucleotide sequence identity of the genomes between HGBV-A, HGBV-B or HGBV-C and a strain of one of these hepatitis GB viruses will be about 45% or greater, since it is now believed that the HGBV strains may be genetically related, preferably about 60% or greater, and more preferably, about 80% or greater.

Also, it is expected that the overall sequence identity of the genomes between HGBV-A and a strain of HGBV-A at the amino acid level will be about 35% or greater since it is now believed that the HGBV strains may be genetically related, preferably about 40% or greater, more preferably, about 60% or greater, and even more preferably, about 80% or greater. In addition, there will be corresponding contiguous sequences of at least about 13 nucleotides, which may be provided in combination of more than one contiguous sequence. Also, it is expected that the overall sequence identity of the genomes between HGBV-B and a strain of HGBV-B at the amino acid level will be about 35% or greater since it is now believed that the HGBV strains may be genetically related, preferably about 40% or greater, more preferably, about 60% or greater, and even more preferably, about 80% or greater. In addition, there will be corresponding contiguous sequences of at least about 13 nucleotides, which may be provided in combination of more than one contiguous sequence. Also, it is expected that the overall sequence identity of the genomes between HGBV-C and a strain of HGBV-C at the amino acid level will be about 35% or greater since it is now believed that the HGBV strains may be genetically related, preferably about 40% or greater, more preferably, about 60% or greater, and even more preferably, about 80% or greater. In addition, there will be corresponding contiguous sequences of at least about 13 nucleotides, which may be provided in combination of more than one contiguous sequence.

A polynucleotide "derived from" a designated sequence for example, the HGBV cDNA, or from the HGBV genome, refers to a polynucleotide sequence which is comprised of a sequence of approximately at least about 6 nucleotides, is preferably at least about 8 nucleotides, is more preferably at least about 10-12 nucleotides, and even more preferably is at least about 15-20 nucleotides corresponding, i.e., similar to or complementary to, a region of the designated nucleotide sequence. Preferably, the sequence of the region from which the polynucleotide is derived is similar to or complementary to a sequence which is unique to the HGBV genome. Whether or not a sequence is complementary to or similar to a sequence which is unique to an HGBV genome can be determined by techniques known to those skilled in the art. Comparisons to sequences in databanks, for example, can be used as a method to determine the uniqueness of a designated sequence. Regions from which sequences may be derived include but are not limited to regions encoding specific epitopes, as well as non-translated and/or non-transcribed regions.

The derived polynucleotide will not necessarily be derived physically from the nucleotide sequence of HGBV, but may be generated in any manner, including but not limited to chemical synthesis, replication or reverse transcription or transcription, which are based on the information provided by the sequence of bases in the region(s) from which the polynucleotide is derived. In addition, combinations of regions corresponding to that of the designated sequence may be modified in ways known in the art to be consistent with an intended use.

The term "polynucleotide" as used herein means a polymeric form of nucleotides of any length, either ribonucleotides or deoxyribonucleotides. This term refers only to the primary structure of the molecule. Thus, the term includes double- and single-stranded DNA, as well as double- and single-stranded RNA. It also includes modifications, either by methylation and/or by capping, and unmodified forms of the polynucleotide.

The terms "polynucleotide," “oligomer,” "oligonucleotide," "oligo" and "primer" are used interchangeably herein.

"HGBV containing a sequence corresponding to a cDNA" means that the HGBV contains a polynucleotide sequence which is similar to or complementary to a sequence in the designated DNA. The degree of similarity or complementarity to the cDNA will be approximately 50% or greater, will preferably be at least about 70%, and even more preferably will be at least about 90%. The sequence which corresponds will be at least about 70 nucleotides, preferably at least about 80 nucleotides, and even more preferably at least about 90 nucleotides in length. The correspondence between the HGBV and the cDNA can be determined by methods known in the art, and include, for example, a direct comparison of the sequenced material with the cDNAs described, or hybridization and digestion with single strand nucleases, followed by size determination of the digested fragments.

"Purified viral polynucleotide" refers to an HGBV genome or fragment thereof which is essentially free, i.e., contains less than about 50%, preferably less than about 70%, and even more preferably, less than about 90% of polypeptides with which the viral polynucleotide is naturally associated. Techniques for purifying viral polynucleotides are well known in the art and include, for example, disruption of the particle with a chaotropic agent, and separation of the polynucleotide(s) and polypeptides by ion-exchange chromatography, affinity chromatography, and sedimentation according to density. Thus, "purified viral polypeptide" means an HGBV polypeptide or fragment thereof which is essentially free, that is, contains less than about 50%, preferably less than about 70%, and even more preferably, less than about 90% of cellular components with which the viral polypeptide is naturally associated. Methods for purifying are known to the routineer.

"Polypeptide" as used herein indicates a molecular chain of amino acids and does not refer to a specific length of the product. Thus, peptides, oligopeptides, and proteins are included within the definition of polypeptide. This term, however, is not intended to refer to post-expression modifications of the polypeptide, for example, glycosylations, acetylations, phosphorylations and the like.

A "polypeptide" or "amino acid sequence" derived from a designated nucleic acid sequence or from the HGBV genome refers to a polypeptide having an amino acid sequence identical to that of a polypeptide encoded in the sequence or a portion thereof wherein the portion consists of at least 3 to 5 amino acids, and more preferably at least 8 to 10 amino acids, and even more preferably 15 to 20 amino acids, or which is immunologically identifiable with a polypeptide encoded in the sequence.

A "recombinant polypeptide" as used herein means at least a polypeptide of genomic, semisynthetic or synthetic origin which by virtue of its origin or manipulation is not associated with all or a portion of the polypeptide with which it is associated in nature or in the form of a library and/or is linked to a polynucleotide other than that to which it is linked in nature. A recombinant or derived polypeptide is not necessarily translated from a designated nucleic acid sequence of HGBV or from an HGBV genome. It also may be generated in any manner, including chemical synthesis or expression of a recombinant expression system, or isolation from mutated HGBV.

The term "synthetic peptide" as used herein means a polymeric form of amino acids of any length, which may be chemically synthesized by methods well-known to the routineer. These synthetic peptides are useful in various applications.

"Recombinant host cells," "host cells," "cells," "cell lines," "cell cultures," and other such terms denoting microorganisms or higher eucaryotic cell lines cultured as unicellular entities refer to cells which can be, or have been, used as recipients for recombinant vector or other transfer DNA, and include the original progeny of the original cell which has been transfected.

As used herein "replicon" means any genetic element, such as a plasmid, a chromosome or a virus, that behaves as an autonomous unit of polynucleotide replication within a cell. That is, it is capable of replication under its own control.

A "vector" is a replicon in which another polynucleotide segment is attached, such as to bring about the replication and/or expression of the attached segment.

The term "control sequence" refers to polynucleotide sequences which are necessary to effect the expression of coding sequences to which they are ligated. The nature of such control sequences differs depending upon the host organism.

In prokaryotes, such control sequences generally include promoter, ribosomal binding site and terminators; in eukaryotes, such control sequences generally include promoters, terminators and, in some instances, enhancers. The term "control sequence" thus is intended to include at a minimum all components whose presence is necessary for expression, and also may include additional components whose presence is advantageous, for example, leader sequences.

"Operably linked" refers to a situation wherein the components described are in a relationship permitting them to function in their intended manner. Thus, for example, a control sequence "operably linked" to a coding sequence is ligated in such a manner that expression of the coding sequence is achieved under conditions compatible with the control sequences.

The term "open reading frame" or "ORF" refers to a region of a polynucleotide sequence which encodes a polypeptide; this region may represent a portion of a coding sequence or a total coding sequence.

A "coding sequence" is a polynucleotide sequence which is transcribed into mRNA and/or translated into a polypeptide when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a translation start codon at the 5' -terminus and a translation stop codon at the 3' -terminus. A coding sequence can include, but is not limited to, mRNA, cDNA, and recombinant polynucleotide sequences.

The term "immunologically identifiable with/as" refers to the presence of epitope(s) and polypeptide(s) which also are present in and are unique to the designated polypeptide(s), usually HGBV proteins. Immunological identity may be determined by antibody binding and/or competition in binding. These techniques are known to the routineer and also are described herein. The uniqueness of an epitope also can be determined by computer searches of known data banks, such as GenBank, for the polynucleotide sequences which encode the epitope, and by amino acid sequence comparisons with other known proteins.

As used herein, "epitope" means an antigenic determinant of a polypeptide. Conceivably, an epitope can comprise three amino acids in a spatial conformation which is unique to the epitope. Generally, an epitope consists of at least five such amino acids, and more usually, it consists of at least eight to ten amino acids. Methods of examining spatial conformation are known in the art and include, for example, x-ray crystallography and two-dimensional nuclear magnetic resonance.

The term "individual" as used herein refers to vertebrates, particularly members of the mammalian species and includes but is not limited to domestic animals, sports animals, primates and humans; more particularly the term refers to tamarins and humans.

A polypeptide is "immunologically reactive" with an antibody when it binds to an antibody due to antibody recognition of a specific epitope contained within the polypeptide. Immunological reactivity may be determined by antibody binding, more particularly by the kinetics of antibody binding, and/or by competition in binding using as competitor(s) a known polypeptide(s) containing an epitope against which the antibody is directed. The methods for determining whether a polypeptide is immunologically reactive with an antibody are known in the art.

As used herein, the term "immunogenic polypeptide containing an HGBV epitope" means naturally occurring HGBV polypeptides or fragments thereof, as well as polypeptides prepared by other means, for example, chemical synthesis or the expression of the polypeptide in a recombinant organism.

The term "transformation" refers to the insertion of an exogenous polynucleotide into a host cell, irrespective of the method used for the insertion. For example, direct uptake, transduction, or f-mating are included. The exogenous polynucleotide may be maintained as a non-integrated vector, for example, a plasmid, or alternatively, may be integrated into the host genome.

"Treatment" refers to prophylaxis and/or therapy.

The term "plus strand" (or "+") as used herein denotes a nucleic acid that contains the sequence that encodes the polypeptide. The term "minus strand" (or "-") denotes a nucleic acid that contains a sequence that is complementary to that of the "plus" strand.

"Positive stranded genome" of a virus denotes that the genome, whether RNA or DNA, is single-stranded and encodes a viral polypeptide(s).

The term "test sample" refers to a component of an individual's body which is the source of the analyte (such as, antibodies of interest or antigens of interest). These components are well known in the art. These test samples include biological samples which can be tested by the methods of the present invention described herein and include human and animal body fluids such as whole blood, serum, plasma, cerebrospinal fluid, urine, lymph fluids, and various external secretions of the respiratory, intestinal and genitorurinary tracts, tears, saliva, milk, white blood cells, myelomas and the like; biological fluids such as cell culture supernatants; fixed tissue specimens; and fixed cell specimens.

"Purified HGBV" refers to a preparation of HGBV which has been isolated from the cellular constituents with which the virus is normally associated, and from other types of viruses which may be present in the infected tissue. The techniques for isolating viruses are known to those skilled in the art and include, for example, centrifugation and affinity chromatography.

"PNA" denotes a "peptide nucleic analog" which may be utilized in various diagnostic, molecular or therapeutic methods. PNAs are neutrally charged moieties which can be directed against RNA or DNA targets. PNA probes used in assays in place of, for example, DNA probes, offer advantages not achievable when DNA probes are used. These advantages include manufacturability, large scale labeling, reproducibility, stability, insensitivity to changes in ionic strength and resistance to enzymatic degradation which is present in methods utilizing DNA or RNA. These PNAs can be labeled with such signal generating compounds as fluorescein, radionucleotides, chemiluminescent compounds, and the like. PNAs or other nucleic acid analogs such as morpholino compounds thus can be used in various methods in place of DNA or RNA. It is within the scope of the routineer that PNAs or morpholino compounds can be substituted for RNA or DNA with appropriate changes if and as needed in reagents and conditions utilized in these methods.

The detection of HGBV in test samples can be enhanced by the use of DNA hybridization assays which utilize DNA oligomers as hybridization probes. Since the amount of DNA target nucleotides present in a test sample may be in minute amounts, target DNA usually is amplified and then detected. Methods for amplifying and detecting a target nucleic acid sequence that may be present in a test sample are well-known in the art. Such methods include the polymerase chain reaction (PCR) described in U.S. Patents 4,683,195 and 4,683,202 which are incorporated herein by reference, the ligase chain reaction (LCR) described in EP-A-320 308, gap LCR (GLCR) described in European Patent Application EP-A-439 182 and U.S. Patent No. 5,427,930 which are incorporated herein by reference, multiplex LCR described in International Patent Application No. WO 93/20227, NASBA and the like. These methods have found widespread application in the medical diagnostic field as well as in the fields of genetics, molecular biology and biochemistry.

The reagents and methods of the present invention are made possible by the provision of a family of closely related nucleotide sequences present in the plasma, serum or liver homogenate of an HGBV infected individual, either tamarin or human. This family of nucleotide sequences is not of human or tamarin origin, since it hybridizes to neither human nor tamarin genomic DNA from uninfected individuals, since nucleotides of this family of sequences are present only in liver (or liver homogenates), plasma or serum of individuals infected with HGBV. In addition, the family of sequences has shown no significant identity at the nucleic acid level to sequences contained within the HAV, HBV, HCV, HDV and HEV genome, and low level identity, considered not significant, as translation products. Infectious sera, plasma or liver homogenates from HGBV infected humans contain these polynucleotide sequences, whereas sera, plasma or liver homogenates from non-infected humans has not contained these sequences. Northern blot analysis of infected liver with some of these polynucleotide sequences has demonstrated that they are derived from a large RNA transcript similar in size to a viral genome. Sera, plasma or liver homogenates from HGBV-infected humans contain antibodies which bind to this polypeptide, whereas sera, plasma or liver homogenates from non-infected humans do not contain antibodies to this polypeptide; these antibodies are induced in individuals following acute non-A, non-B, non-C, non-D and non-E hepatitis infection. By these criteria, it is believed that the sequence is a viral sequence, wherein the virus causes or is associated with non-A, non-B, non-C, non-D and non-E hepatitis.

Using determined portions of the isolated HGBV nucleic acid sequences as a basis, oligomers of approximately eight nucleotides or more can be prepared, either by excision or synthetically, which hybridize with the HGBV genome and are useful in identification of the viral agent(s), further characterization of the viral genome, as well as in detection of the virus(es) in diseased individuals. The natural or derived probes for HGBV polynucleotides are a length which allows the detection of unique viral sequences by hybridization. While six to eight nucleotides may be a workable length, sequences of ten to twelve nucleotides are preferred, and those of about 20 nucleotides may be most preferred. These sequences preferably will derive from regions which lack heterogeneity. These probes can be prepared using routine, standard methods including automated oligonucleotide synthetic methods. A complement of any unique portion of the HGBV genome will be satisfactory. Complete complementarity is desirable for use as probes, although it may be unnecessary as the length of the fragment is increased.

Synthetic oligonucleotides may be prepared using an automated oligonucleotide synthesizer such as that described by Warner, DNA 3:401 (1984).

If desired, the synthetic strands may be labeled with ^{32}P by treatment with polynucleotide kinase in the presence of ^{32}P-ATP, using standard conditions for the reaction. DNA sequences including those isolated from genomic or cDNA libraries, may be modified by known methods which include site directed mutagenesis as described by Zoller, Nucleic Acids Res. 10:6487 (1982). Briefly, the DNA to be modified is packaged into phage as a single stranded sequence, and converted to a double stranded DNA with DNA polymerase using, as a primer, a synthetic oligonucleotide complementary to the portion of the DNA to be modified, and having the desired modification included in its own sequence. Culture of the transformed bacteria, which contain replications of each strand of the phage, are plated in agar to obtain plaques. Theoretically, 50% of the new plaques contain phage having the mutated sequence, and the remaining 50% have the original sequence. Replicates of the plaques are hybridized to labeled synthetic probe at temperatures and conditions suitable for hybridization with the correct strand, but not with the unmodified sequence. The sequences which have been identified by hybridization are recovered and cloned.

Polymerase chain reaction (PCR) and ligase chain reaction (LCR) are techniques for amplifying any desired nucleic acid sequence (target) contained in a nucleic acid or mixture thereof. In PCR, a pair of primers are employed in excess to hybridize at the outside ends of complementary strands of the target nucleic acid. The primers are each extended by a polymerase using the target nucleic acid as a template. The extension products become target sequences themselves, following dissociation from the original target strand. New primers are then hybridized and extended by a polymerase, and the cycle is repeated to geometrically increase the number of target sequence molecules. PCR is disclosed in U.S. patents 4,683,195 and 4,683,20, previously incorporated herein by reference.

LCR is an alternate mechanism for target amplification. In LCR, two sense (first and second) probes and two antisense (third and fourth) probes are employed in excess over the target. The first probe hybridizes to a first segment of the target strand and the second probe hybridizes to a second segment of the target strand, the first and second segments being positioned so that the primary probes can be ligated into a fused product. Further, a third (secondary) probe can hybridize to a portion of the first probe and a fourth (secondary) probe can hybridize to a portion of the second probe in a similar ligatable fashion. If the target is initially double stranded, the secondary probes will also hybridize to the target complement in the first instance. Once the fused strand of sense and antisense probes are separated from the target strand, it will hybridize with the third and fourth probes which can be ligated to form a complementary, secondary fused product. The fused products are functionally equivalent to either the target or its complement. By repeated cycles of hybridization and ligation, amplification of the target sequence is achieved. This technique is described in EP-A-320,308, hereby incorporated by reference. Other aspects of LCR technique are disclosed in EP-A-439,182, which is incorporated herein by reference.

The 5'-NTR region of HGBV-A is approximately 592 nucleotides long (SEQUENCE ID NOs 23). This region in HGBV-B is approximately 445 nucleotides long (SEQUENCE ID NO 32), and the 5'NTR region of HGBV-C is approximately 533 nucleotides in length (SEQUENCE ID NO 4). To functionally characterize the 5' ends of GBV-A and GBV-C RNAs, the sites and mechanism of translation initiation of both monocistronic and bicistronic RNAs were examined in a cell-free *in vitro* translation system. Weak IRES elements were found to be present in the 5' RNAs of GBV-A and GBV-C suggesting that these sequences are complete or nearly complete. In addition, the position of the initiating AUG codons in the monocistronic RNAs, and presumably in the viral genomic RNA as well, demonstrated that GBV-A and GBV-C do not contain core proteins at the N-termini of their polyproteins. Thus, GBV-A and GBV-C appear unique from other members of the *Flaviviridae* and may constitute a separate group within this family. Consistent with this hypothesis, we also discovered that the secondary structures of the 5' ends of these viruses are different from the conserved structures present in the 5' NTRs of the pestiviruses, HCV and GBV-B.

The present invention provides nucleic acids that are capable of interacting with distinct cis-acting control elements of HGBV and thus are capable of blocking, enhancing or suppressing the translation of HGBV nucleic acids.

In a first embodiment, a method for controlling the translation of HGBV nucleic acids to proteins is provided. This method comprises the steps of contacting a first *non-naturally occurring* nucleic acid with HGBV nucleic acid. This first nucleic acid has a sequence that is complementary to a sequence of the sense strand within the 5'NTR region of HGBV-A, -B or -C. This first nucleic acid is contacted with an HGBV nucleic acid for times and under conditions suitable for hybridization to occur, and thus form a hybridization product. The hybridization results in the alteration of the level of translation of the HGBV nucleic acid.

The antisense nucleic acid of the present invention is RNA, DNA or a modified nucleic acid such as a PNA or morpholino compound, degradation-resistant sulfurized and thiophosphate derivatives of nucleic acids, and the like. Modified nucleic acids preferably will be able to increase the intracellular stability and/or permeability of the nucleic acid, increase the affinity of the nucleic acid for the sense strand or decrease the toxicity of the nucleic acid. Such advantages are well known in the art, and are described in, for example, S. T. Crooke et al., eds., Antisense Research and Applications, CRC Press (1993).

Antisense nucleic acids thus can be modified or altered to contain modified bases, sugars or linkages, be delivered in specialized systems such as liposomes or by gene therapy, or may have attached moieties. Such attached moieties, such as hydrophobic moieties such as lipids and in particular, cholesterols, can enhance the interaction of the nucleic acid with cell membranes. In addition, such attached moieties can act as charge neutralizers of the phosphate backbone (for example, polycationic moieties such as polylysine). These moieties can be attached at either the 5' or the 3' end of the nucleic acids, and also can be attached through a base, sugar or internucleotide linkage. Other moieties can act as capping groups which are specifically placed at the 3' or the 5' ends of the nucleic acids to prevent exonucease degradation. These capping groups include, for example, hydroxyl protecting groups including glycols such a polyethylene glycols (PEG), tetraethylene glycol (TEG) and the like.

The first nucleic acid will have at least 10 nucleotides in a sequence substantially complementary to a sequence of the sense strand within the 5'NTR region of HGBV-A, -B, or -C. Preferably, the first nucleic acid has about 12 nucleotides in such a complementary sequence; more preferably, the first nucleic acid has about 15 nucleotides; and still more preferably, the first nucleic acid has about 20 nucleotides. It is preferred that such a first nucleic acid have less than 100 nucleotides in such a complementary sequence, and more preferably, a first nucleic acid will have less than 50 nucleotides. Most preferably, the first nucleic acid will have between 20 to 30 nucleotides that are capable of forming a stable hybridization product with a sense sequence of the 5'NTR region of HGBV-A, -B or -C.

The 5'NTR region of HGBV-A is set forth in SEQUENCE ID NO 23; the 5'NTR region of HGBV-B is set forth in SEQUENCE ID NO 32; and the 5'NTR region of HGBV-C is set forth in SEQUENCE ID NO 4. The nucleic acid can be placed in the cell through several ways known to those in the art. For example, cells can be transfected with a second nucleic acid capable of generating the first nucleic acid as a transcription product (for example, by including the second nucleic acid in a viral carrier as detailed by U.S. Patent 4,493,002, incorporated herein by reference, or by gene therapy methods such as including the second nucleic acid in a retroviral vector). Gene therapy methods are known to those of skill in the art.

The present invention further encompasses means for placing the first nucleic acid or the second nucleic acid into cells infected with HGBV-A, -B or -C or into cells which are to be protected from HGBV infection. Examples of such means include but are not limited to vectors, liposomes and lipid suspensions, such as N-(1-(2,3-dioleoyloxy)propul)-N, N, N-thrimethylammonium methylsulfate (DOTAP), N-[1-(2,3-dioleyloxy)propul]-N, N, N-trimethylammonium chloride (DOTMA), and the like. The lipid may be covalently linked directly to the first nucleic acid in an alternative embodiment.

The antisense nucleic acid also may be linked to moieties that increase cellular uptake of the nucleic acid. Such moieties may be hydrophobic (such as, phospholipids or lipids such as steroids [for example, cholesterol]) or may be polycationic moieties that are attached at any point to the antisense nucleic acid, including at the 5' or 3' ends, base, sugar hydroxyls and internucleoside linkages. A moiety known to increase uptake is a cholesteryl group, which may be attached through an activated cholesteryl chloroformate or cholic acid, by means known in the art.

Further, enhancement of translation may allow for stronger immune responses. Blocking or decreasing translation of viral nucleic acid may decrease the pathology of the viral infection.

Nucleic acid or nucleic acid analogs can be provided as compositions for pharmaceutical administration. Injection preparations and suppositories may usually contain 1-10 mg of the nucleic acid or nucleic acid analog per dose (ampule or capsule). For humans the daily dose of about 0.1 to 1000 mg, preferably 1-100 mg (from about 10-20 mg/kg to 1000 to 2000 mg/kg body weight) is the daily dosage. As is known to those in the art, however, a particular dose for a particular individual depends on a variety of factors, including but not limited to, effectiveness of the particular nucleic acid or nucleic acid analog used, the age, weight and general state of health of the individual, the diet and sex of the individual, the mode of administration of the dosage, the rate of elimination and half life of the composition, whether this composition is used in combination with other medications and the clinical severity of the individual's disease. Such compositions which are pharmaceutical articles of manufacture include articles whose active ingredients are contained in an effective amount of attain the intended purpose. A preferred range has been described hereinabove, and determination of the most effective amounts for treatment of each HGBV infection is well within the skill of the rountineer.

In addition to the nucleic acid and nucleic acid analogs of the present invention, contemplated pharmaceutical preparations may contain suitable excipients and auxiliaries which facilitate processing of the active compounds. These preparations can be administered orally, rectally, parenterally, bucally or sublingually. All may contail from 0.1 to 99% by weight of active ingredients, together with an excipient. A preferred method of administration is parenteral, especially intraveneous administration.

Suitable formulations for parenteral administration include aqueous solutions of the active compounds in water-soluble or water-dispersible form. Additionally, suspensions of the active compounds as appropriate oily injection suspensions may be administered. Suitable lipophilic solvents or vehicles include fatty oils (for example, sesame oil or synthetic fatty acid esters such as ethyloleate or triglycerides). Aqueous injection suspensions may contain substances which increase the viscosity of the suspension, for example, sodium carboxymethyl cellulose, sorbitol, and/or dextran. The suspension also may contain stabilizers.

It is within the scope of the present invention that the compositions described herein may be administered encapsulated in liposomes, pharmaceutical compositions wherein the active ingredient is contained either dispersed or variously present in corpuscles consisting of aqueous concentric layers adherent to lipidic layers. Methods of utilizing this technology are known in the art.

The present invention will now be described by way of examples, which are meant to illustrate, but not to limit, the spirit and scope of the invention.

EXAMPLES

Example 1. Internal ribosome entry site in 5' NTR of GBV-B

Several positive strand RNA viruses, such as picornaviruses and pestiviruses, possess large 5' nontranslated regions (NTRs). These large NTRs control the initiation of cap-independent translation by functioning as internal ribosome entry sites (IRESs) (Pelletier and Sonenberg, Nature (London) 334:320-

325). The IRES is thought to form a specific RNA structure which allows ribosomes to enter and begin translation of an RNA without using the cellular machinery required for cap-dependent translation initiation. The large 5' NTR of HCV has been shown to possess an IRES (Tsukiyama-Kohara et al. J. Virol. 66:1476-1483, 1992; Wang *et al.* J. Virol. 67:3338-3344, 1993; Rijnbrand et al. FEBS Letters 365:115-119, 1995). Due to the high level of sequence conservation between the 5' NTRs of GBV-B and HCV, it was reasoned that GBV-B may also contain an IRES.

To test for IRES function in GBV-B (SEQUENCE ID NO 32), the 5' NTR of this virus was used to replace the 5' NTR of hepatitis A virus (HAV) in the pLUC-HAV-CAT plasmid described by Whetter *et al.* (J. Virol. 68:5253-5263, 1994). The 5' NTR of GBV-B was amplified from a plasmid clone using SEQUENCE ID NO. 58 (UTR-B.1) and SEQUENCE ID. NO. 59 (NTR-B-a1) as primers Briefly, a 50 µl PCR was set up using a Perkin-Elmer PCR kit as described by the manufacturer with 1 µM primers, 2 mM $MgCl_2$ and approximately 10 ng of plasmid. This reaction was amplified for 20 cycles (94°C, 20 sec; 55°C, 30 sec; 72°C, 30 sec) followed by a final extension at 72°C for 10 min. The completed reaction then was held at 4°C. This product was extracted with phenol:chloroform and precipitated as described in the art. The 3' terminal adenosine residues added by the AmpliTaq® polymerase were removed from this product by incubation with T4 DNA polymerase and deoxynucleotide triphosphates as described (Sambrook *et al.*, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Press, 1989). After heat inactivation, the product was digested with Xba I and gel purified as described in the art. The purified product was ligated to pHAV-CAT1 (Whetter *et al.* J. Virol. 68:5253-5263, 1994) that had been cut with HindIII, end-filled with Klenow polymerase and deoxynucleotide triphosphates, heat-inactivated, digested with Xba I, treated with bacterial alkaline phosphatase, extracted with phenol:chloroform, and precipitated as described in the art. The constructed plasmid, pGBB-CAT1, was digested with Sac I, blunt-ended with T4 DNA polymerase and deoxynucleotide triphosphates, heat-inactivated, and digested with Not I as described in the art. The 1.3 kbp product from these reactions was gel purified and cloned into pLUC-HAV-CAT (Whetter *et al.* J. Virol. 68:5253-5263, 1994) that had been digested with HindIII, end-filled with Klenow polymerase and deoxynucleotide triphosphates, heat-inactivated, digested with Not I, treated with bacterial alkaline phosphatase, extracted with phenol:chloroform, and precipitated as described in the art. The resultant plasmid, pLUC-GBB-CAT was used in *in vitro* transcription-translation experiments to test for an IRES function.

An *in vitro* transcription-translation assay was performed using the TNT™ T7 coupled reticulocyte lysate system from Promega (Madison, WI) as described by the manufacturer. The plasmids tested were pLUC-GBB-CAT (described above), pLUC-HAV-CAT (positive control from Whetter *et al.* J. Virol. 68:5253-5263, 1994), and pLUC-Δ355-532 (negative control from Whetter *et al.* J. Virol. 68:5253-5263, 1994). The products (labeled with ^{35}S-methionine) were run on a 10% Laemmli gel as described in the art. The gel was fixed in 10% methanol, 20% acetic acid for 10 minutes, dried down and exposed to a PhosphoImager® screen (Molecular Dynamics, Sunnyvale, CA). The products were visualized with the PhosphoImager®. In addition, the reactions were examined for Luc and CAT activity using commercially available kits (Promega, Madison, WI)(data not shown).

All three reactions contained luciferase activity and a band consistent with the size expected for luciferase (transcribed from the LUC gene in the plasmid). LUC expression, which is a measure of the level of translation that initiates from the 5' end of the mRNA, appeared to be equivalent in the three reactions. Thus, equivalent amounts of RNA templates were present in a translatable form in these three reactions. The pLUC-HAV-CAT and the pLUC-GBB-CAT reactions also possessed chloramphenicol acetyltransferase (CAT) activity and contained a band consistent with the size expected for CAT (from the the CAT gene in the plasmid). This band is not seen in the pLUC-Δ355-532 negative control. CAT expression measures the level of internal translation initiation. Because translation of the CAT gene requires the existence of an IRES in this plasmid construct, the 5' NTR of GBV-B must be providing this function. Therefore, similar to HCV, GBV-B's 5'NTR contains an IRES. Further studies of these plasmids, both *in vitro* and *in vivo* are ongoing to better characterize the IRES in GBV-B.

Example 2. Internal ribosome entry site in 5' NTR of GBV-A and -C

A. Plasmids. Various monocistronic and bicistronic plasmids were constructed with PCR-amplified sequences of GBV-A and GBV-C. PCRs utilized components of the GeneAmp PCR Kit with AmpliTaq (Perkin-Elmer) as directed by the manufacturer with final reaction concentrations of 1 μM for oligonucleotide primers and 2 mM $MgCl_2$. PCR products were digested with restriction endonucleases, gel purified and cloned using standard procedures as described by J. Sambrook et al., Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor (1989). Monocistronic fusions between GBV sequences and bacterial chloramphenicol acetyltransferase (CAT) were generated by replacing the hepatitis A virus (HAV) HindIII/XbaI fragment of pHAV-CAT1 (described by L. E. Whetter et al., J. Virology 68:5253-5263 (1994) with PCR-amplified cDNA from the 5' ends of GBV-A and GBV-C. The bicistronic constructs were generated in pT7/CAT/ICS/Luc, described by D. Macejak et al., in M. A. Brinton et al., eds., New Aspects of Positive-Strand RNA Viruses, American Society for Microbiology, Washington, D. C.,1990, p. 152-157, and provided as a gift by P. Sarnow, in a two step procedure. First, monocistronic fusions between GBV and luciferase (Luc) were constructed by inserting GBV sequences into the HindIII/NcoI-cut pT7/CAT/ICS/Luc. Bicistronic vectors were constructed by cloning the HindIII/blunt/SacI GBV fragment from these monocistronic vectors into pT7/CAT/ICS/Luc which had been digested with SalI (blunt) and SacI. The sequence of the cloned inserts and ligation junctions were confirmed by dsDNA sequencing (Sequenase 2.0, USB, Cleveland). Nomenclature (e.g. A15-707) describes the source (GBV-A) and range (nts 15 to 707) of sequence incorporated into the various vectors.

GBV-A sequences (GenBank accession no. U22303) were amplified from a plasmid clone. PCRs for the GBV-A monocistronic and bicistronic constructs utilized the sense primer 5'-TATAATAAGCTTGCCCCGGACCTCCCACCGAG-3' (HindIII site underlined) (SEQUENCE ID NO 5) coupled with 5'-GCTCTAGATCGGGAACAACAATTGGAAAG (SEQUENCE ID NO 6), 5'-GCTCTAGAGCACTGGTGCCGCGAGT (SEQUENCE ID NO 11), 5'-GCTCTAGAGAGGGGGAAGCAAACCA (SEQUENCE ID NO 12) and 5'-GCTCTAGACATGGTGAATGTGTCGACCAC (Xba I sites underlined) (SEQUENCE ID NO 13) for the monocistronic vectors pA15-707/CAT, pA15-665/CAT, pA15-629/CAT and pA15-596/CAT, respectively; and 5' CCATAATCATGAGGGAACAACAATTGGAAAG (SEQUENCE ID NO 17) , 5'-CCATAATCATGAGCCGCGAGTTGAAGAGCAC (SEQUENCE ID NO 24), and 5' GCCAAGCCATGGTGAATGTG 3' (BspHI or NcoI sites underlined) (SEQUENCE ID NO 25) for the bicistronic vectors pCAT/A15-705/Luc, pCAT/A15-657/Luc and pCAT/A15-596/Luc, respectively. In addition, a GBV-A sequence amplified with 5'-TATAATAAGCTTGCCGCGAGTTGAAGAGCAC (SEQUENCE ID NO 21) and 5'-

CCATAATCATGAGCCCCGGACCTCCCACCGAG (SEQUENCE ID NO 22) were used to construct pCAT/A657-15/Luc which contain GBV-A sequences in the antisense orientation.

GBV-C sequences were amplified from a plasmid generated during the cloning of GBV-C 5' sequences, as described in U.S. Serial No. U.S. Serial No. 08/580,038, previously incorporated herein by reference. The sequence of this GBV-C cDNA (nts 1 to 631, SEQUENCE ID NO 4) corresponds to nts 30 to 659 of GenBank accession no. U44402, the longest GBV-C isolate reported to date and nts 13 to 643 of SEQUENCE ID NO. 3. PCRs for the GBV-C monocistronic and bicistronic plasmids utilized the sense primer 5'-TATAATAAGCTTCACTGGGTGCAAGCCCCA (HindIII site underlined) (SEQUENCE ID NO 7) coupled with 5'-GCTCTAGAGGCGCAACAGTTTGTGAGGAA (SEQUENCE ID NO 8), 5'-GCTCTAGAACAAGCGTGGGTGGCCGGGG (SEQUENCE ID NO 14), 5'-GCTCTAGAGACCACGAGAAGGAGCAGAAG (SEQUENCE ID NO 15) and 5'-GCTCTAGACATGATGGTATAGAAAAGAG (Xba I site underlined) (SEQUENCE ID NO 16) for the monocistronic vectors pC1-631/CAT, pC1-592/CAT, pC1-553/CAT and pC1-526/CAT, respectively; and 5'-CATGCCATGGCGCAACAGTTTGTGAGGAA (SEQUENCE ID NO 18), 5'-GTATTGCGCCATGGCTCGACAAGCGTGGGTGGCCGGGG (SEQUENCE ID NO 26), and 5'-GGACTGCCATGGTGGTATAGAAAAGAG (NcoI sites underlined) (SEQUENCE ID NO 27) for the bicistronic vectors pCAT/C1-629/Luc, pCAT/C1-596/Luc and pCAT/C1-526/Luc, respectively. Additional GBV-C sequences were amplified with 5'-GCTCTAGACACTGGGTGCAAGCCCCA (XbaI site underlined) (SEQUENCE ID NO 9) and 5'-TATAATAAGCTTGGCGCAACAGTTTGTGAG (HindIII site underlined) (SEQUENCE ID NO 10) for the monocistronic pC631-1/CAT plasmid, and 5'- TATAATAAGCTTCTCGACAAGCGTGGGTGGCCGGGG 3' (HindIII site underlined) (SEQUENCE ID NO 28) and 5'-GTATTGCGCCATGGCACTGGGTGCAAGCCCCAGAA (NcoI site underlined) (SEQUENCE ID NO 29) for the bicistronic pCAT/C596-1/Luc plasmid. Both of these plasmids contain GBV-C sequences in the antisense orientation.

HCV sequences were amplified from a plasmid clone of a genotype 1a isolate using the sense primer 5'-TATAATAAGCTTCACTCCCCTGTGAGGAACTAC (HindIII site underlined) (SEQUENCE ID NO 19) coupled with 5'-

GTATTGCGTCATGATGGTTTTTCTTTGGGGTTTAG (SEQUENCE ID NO 20) or 5'-CCATAATCATGATGCACGGTCTACGAGACCT (BspHI sites underlined) (SEQUENCE ID NO 30) to generate the bicistronic vectors pCAT/HCV39-377/Luc and pCAT/HCV39-345/Luc, respectively.

Site-specific nucleotide changes were generated in pA15-707/CAT and pC1-631/CAT using the MORPH™ site-specific plasmid DNA mutagenesis kit (5 Prime --> 3 Prime, Inc., Boulder, CO) as directed by the manufacturer. Nucleotide changes were confirmed by dsDNA sequencing as described above.

B. *In vitro* transcription/translation. *In vitro* transcription/translation (IVTT) reactions were performed with the TNT™ T7 Coupled Reticulocyte Lysate System (Promega) according to manufacturer's instructions. Reactions (25 µl) contained 20 units rRNasin (Promega), 20 µCi ^{35}S-cysteine (1000 Ci/mmol, Amersham), and 0.5 µg of plasmid template. After incubation at 30°C for 60 minutes, 5 µl aliquots were denatured (5 minutes, 99°C) in an equal volume of 2X SDS/PAGE loading buffer (125 mM Tris, pH 6.8, 4% SDS, 20% glycerol, 10% 2-mercaptoethanol and 0.2 mg/ml bromophenol blue) and electrophoretically separated on 10 to 20% SDS-polyacrylamide gels (Bio-Rad). The gels were fixed in 10% methanol, 20% acetic acid, dried and analyzed with a PhosphorImager SI™ using ImageQuaNT™ software (Molecular Dynamics, Inc.). Image exposure time, white-black range and product quantitations are presented hereinbelow corresponding figure descriptions.

C. Reporter gene enzymatic assays. Luciferase assays were performed by mixing 50 µl of 1X Luciferase Assay Reagent (Promega) with 1 µl of a 10-fold dilution of a rabbit reticulocyte lysate reaction. Activity was assayed immediately by a 5 second count in a Clinilumat LB9502 Luminometer (Berthold Systems Inc., Pittsburgh). CAT assays were completed with a commercially available kit (Promega) according to manufacturer's instructions. Briefly, 5 µl of lysate was incubated with [^{3}H]chloramphenicol and n-butyryl CoA in a 125 µl reaction for one hour at 37 °C. Butyrylated [^{3}H]chloramphenicol products were isolated by xylene extraction and quantitated by liquid scintillation counting.

D. Secondary RNA structure. A model of the secondary structure of the 5' nontranslated RNA of the GBV-C genome was constructed using a combination of phylogenetic and thermodynamic approaches. A first level phylogenetic analysis considered nucleotide sequences representing the 5' RNA of GBV-C strains present in 35 different patient sera, as presented in U.S. Serial No. 08/580,038, filed December 21, 1995, previously incorporated herein by reference. These were aligned with the program PILEUP (Wisconsin Sequence Analysis Package, version 8, September 1994; Genetics Computer Group, Madison, Wisconsin) and subjected to a manual search for covariant nucleotide substitutions indicative of conserved helical structures. In addition to canonical Watson-Crick base pairs, G-U base pairs were considered acceptable for this analysis. Conserved helical structures identified by the presence of one or more covariant nucleotide substitutions were forced to base pair in the subsequent computer-based folding of the prototype GBV-C sequence (GenBank accession no. U36380) (SEQUENCE ID NO 3) which used the program MFOLD. Separate MFOLD analyses were carried out with sequences representing nts 1-611, 43-522 (both closed at 273-418), 273-418, and 43-180 of SEQUENCE ID NO 3. MFOLD predicts a series of alternative structures with different predicted folding energies. These were reviewed to determine which predicted structures were most permissive for covariant and noncovariant nucleotide substitutions present in the other GBV-C sequences. Where no predicted structure could accommodate most nucleotide substitutions, the sequence was left single stranded in the final model. A second level phylogenetic analysis involved the alignment of GBV-C sequences with the 5' RNA sequences of 5 separate GBV-A strains (as described in G. G. Schlauder et al., Lancet 346:447 [1995] and J. N. Simons et al., Proc Natl. Acad. Sci. USA 92:3401-3405 [1995]), followed by a manual search for covariant substitutions indicative of similar structures in the 5' sequences of these related viruses.

E. Results.

1. Translation of monocistronic transcripts containing 5' GBV RNA. A common Asn-Cys-Cys motif homologous to the HCV E1 Asn-Ser-Cys motif is found near the N-termini of the putative E1 proteins of GBV-A, GBV-B and GBV-C (T. P. Leary et al., supra and FIGURE 1). Located near the N-termini of the GBV-A and GBV-C large ORFs, this tripeptide sequence appears to be the 5' most conserved motif between HCV and the GB viruses. Because it is within the coding regions of GBV-B and HCV and in-frame with the long ORF, this sequence was believed likely to be translated in GBV-A and GBV-C as well. To determine whether the 5' ends of GBV-A and -C could direct translation, nts 15 to 707 of GBV-A (SEQUENCE ID NO 23) and nts 1 to 631 of GBV-C (SEQUENCE ID NO 4) were cloned into plasmid vectors to create pA15-707/CAT and pC1-631/CAT, respectively. These vectors contained a T7 promoter driving transcription of the 5' GBV sequences, which were ligated in-frame (relative to the Asn-Cys-Cys motif) with the bacterial chloramphenicol acetyltransferase (CAT)

gene, as shown in FIGURE 2A. For GBV-C, only AUGs conserved in all isolated examined are depicted.

In vitro transcription-translation (IVTT) reactions containing rabbit reticulocyte lysates were programmed with pA15-707/CAT, pC1-631/CAT and a positive control plasmid, pHAV-CAT1, which contained the 5' NTR of hepatitis A virus (HAV) inserted upstream of CAT. All three plasmid DNAs directed the translation of discreet products migrating with somewhat different molecular masses in SDS-PAGE, as shown in FIGURE 2B. Referring to the FIGURE 2B, the image was generated from a 16 h exposure with a linear range of 7 to 200. GBV-CAT product in lanes 1 and 2 are present at 26 to 27% of the level of the CAT product made from pHAV-CAT1 (lane 3) when the number of Cys residues have been normalized for each product. The products derived from pA15-707/CAT and pC1-631/CAT were slightly larger than that derived from pHAV-CAT1, indicating that translation was initiating upstream of the site of GBV-CAT fusion. In contrast, no product was detected in IVTT reactions programmed with pC631-1/CAT which contained the GBV-C sequences inserted in the antisense orientation relative to CAT. Only the pHAV-CAT1-programmed reaction possessed detectable CAT activity (data not shown). The absence of activity in the products of reactions programmed with pA15-707/CAT and pC1-631/CAT was likely to be due to the misfolding of the CAT protein as a result of its fusion with the N-terminal segment of the GBV polyprotein.

To confirm that the products of the reactions programmed with pA15-707/CAT and pC1-631/CAT were in fact GBV-CAT fusion proteins, the pA15-707/CAT and pC1-631/CAT plasmids were digested with *SspI* prior to being used to program reactions. *SspI* linearized these plasmids within the CAT coding region so that run-off transcripts produced from these plasmids would lack sequences encoding the C-terminal 45 amino acids of CAT. As expected, reactions programmed with the *SspI*-digested pA15-707/CAT and pC1-631/CAT DNAs (FIGURE 2B, lanes 5 and 6, respectively) contained products that were approximately 5 kDa smaller than those found in reactions programmed with undigested pA15-707/CAT and pC1-631/CAT plasmids (lanes 1 and 2 of FIGURE 2B, respectively).

2. <u>Site of translation initiation in GBV-A and GBV-C</u>. The apparent molecular masses of the GBV-CAT fusion proteins shown in FIGURE 2B suggested possible sites of translation initiation. As indicated in FIGURE 1, the GBV-A and GBV-C ORFs that were ligated to CAT in pA15-707/CAT and pC1-

631/CAT each contained two in-frame AUG codons that might serve as potential sites of translation initiation within the sequence immediately upstream of CAT. These were the fourth and fifth AUG codons in each of the GBV-A and GBV-C sequences (see FIGURE 2A). If initiation occurred at the fourth AUG, the resultant fusion proteins would contain 46 amino acids of GBV-A (adding 5.1 kDa to the 24 kDa of CAT) (SEQUENCE ID NO 30) or 67 amino acids of GBV-C (adding 7.5 kDa to CAT) (SEQUENCE ID NO 31), respectively. In contrast, initiation at the fifth AUG in these transcripts would produce CAT fusion proteins containing 38 and 36 amino acids of GBV-A and GBV-C encoded protein, respectively, adding 4.1 kDa to CAT. The apparent molecular mass of the ~28 kDa fusion proteins detected in the reactions programmed with pA15-707/CAT and pC1-631/CAT suggested that translation initiates at the fifth AUG in each transcript (i.e., the second in-frame Met codons in the long ORF, which are located at nt 594 of the GBV-A sequence [SEQUENCE ID NO 23] and nt 524 of the GBV-C sequence [SEQUENCE ID NO 4]). To identify the sites of translation initiation, the first and second in-frame AUG codons in GBV-A (SEQUENCE ID NO 23) and GBV-C (SEQUENCE ID NO 4)were changed to UAG stop codons producing pAmut1/CAT, pAmut2/CAT, pCmut1/CAT and pCmut2/CAT, as shown in FIGURE 3A. These plasmids were used to program IVTT reactions.

GBV-CAT fusion proteins were detected in reactions programmed with pAmut1/CAT and pCmut1/CAT, as shown in FIGURE 3B, lanes 2 and 5, respectively). Referring to FIGURE 3B, image characteristics are identical to those of FIGURE 2B. The GBV-CAT proteins in Lanes 1 and 4 are present at 35 to 41% of the level of CAT produced from pHAV-CAT1 template (lane 7). Amut 1 (lane 2) is 94% of A15-707 (lane 1); Cmut1 (lane 5) is 42% of C1-631 (lane 4). Reactions programmed with pAmut2/CAT and pCmut2/CAT (FIGURE 3B, lanes 3 and 6, respectively) did not produce detectable quantities of fusion protein. Thus, because the 28 kDa GBV-CAT protein was detected when the first in-frame AUG codon (nt. 570 in GBV-A [SEQUENCE ID NO 23] and nt. 431 in GBV-C [SEQUENCE ID NO 4]) was replaced with a stop codon, initiation did not occur at this position. However, mutation of the second in-frame AUG codon (nt. 594 in GBV-A [SEQUENCE ID NO 23] and nt. 524 in GBV-C [SEQUENCE ID NO 4]) completely abrogated protein production directed by these constructs, consistent with the second in-frame AUG being the site of translation initiation in both GBV-A (SEQUENCE ID NO 23) and GBV-C (SEQUENCE ID NO 4). In a related experiment, IVTT reactions programmed with a plasmid containing GBV-C sequence with an AUG to ACG change at the position of the second in-frame AUG (nt 524) produced protein of identical size to pC1-631/CAT, although at a diminished level (data not shown). Because initiation has been found to occur with lower efficiency at ACG codons in other mRNAs (R. Böck et al., EMBO J 13:3608-3617 [1994]), these data are consistent with translation of the GBV-C/CAT fusion protein initiating at the ACG codon.

The number and position of Leu residues immediately downstream of the initiator Met in both GBV-A (SEQUENCE ID NO 23) and GBV-C (SEQUENCE ID NO 4) provided a biochemical method to confirm the position of the initiation site in the GBV-CAT fusion proteins. IVTT reactions containing ^{3}H-Leu were programmed with pA15-707/CAT and pC1-631/CAT. Reaction products were separated by SDS-PAGE, transferred onto a solid support, and the 28 kDa protein bands were excised. The N-terminal amino acids of the resultant GBV-CAT fusion proteins were sequentially removed by Edman degradation and each fraction was analyzed by scintillation counting. These results are shown in FIGURE 4A and 4B. The ^{3}H-Leu profile obtained from the pA15-707/CAT product was consistent with the expected sequence of GBV-A downstream of the second in-frame AUG, as shown in FIGURE 4A) assuming that the N-terminal Met residue is removed (see, F. Sherman et al., Bioessays 3:27-31 [1985]). Some trailing of the ^{3}H signal was noted which may be attributed to incomplete removal of the N-terminal Met. However, for the pC1-631/CAT product, the ^{3}H-Leu profile exactly matched the expected amino acid sequence downstream of the second in-frame AUG for GBV-C, as shown in FIGURE 4B). Referring to FIGURE 4B, CPM following each degradation cycle is plotted above the predicted N-terminal sequences (minus initiator Met) of HGBV-A (SEQUENCE ID NO 30) and GBV-C (SEQUENCE ID NO 31). These experiments thus confirm that translation is initiated at nt 594 of the GBV-A sequence (SEQUENCE ID NO 23) and nt 524 of the GBV-C sequence (SEQUENCE ID NO 4). The relative length of the 5' nontranslated RNA segments and the multiple AUG codons (some of which are in good context for translation initiation) upstream of the authentic initiator AUG in these transcripts both suggest that translation is initiated on these RNAs by internal ribosomal entry, rather than by a conventional 5' scanning mechanism. Thus, we concluded that it is likely that the GBV-A and GBV-C 5' sequences contain an IRES.

3. GBV coding sequence is required for efficient translation of monocistronic RNAs. The results of the *in vitro* translation reactions described

above demonstrated that initiation begins at the Met residue positioned immediately upstream of the putative E1 signal sequence in both pA15-707/CAT and pC1-631/CAT. To determine the 3' limits of the apparent IRES in GBV-A and GBV-C, and whether any amount of GBV sequence is necessary for protein production in the IVTT assays, several 3' deletions were made which reduced the amount of GBV sequence in the GBV-CAT fusion proteins. A schematic of these constructs is shown in FIGURE 5A. Protein production was observed in reactions programmed with the deletion constructs pA15-665/CAT and pC1-592/CAT, which encode 72 and 69 nucleotides of the GBV-A (SEQUENCE ID NO 23) and GBV-C (SEQUENCE ID NO 4) coding sequence fused to CAT, respectively, and as shown in FIGURE 5B, lanes 2 and 6). Referring to FIGURE 5B, image characteristics are identical to those of FIGURE 2B. GBV-CAT protein (lanes 1, 2, 5 and 6) is present at 20 to 36% of the level of CAT produced from the pHAV-CAT1 template (lane 9). In contrast, no protein was detected in reactions programmed with the deletion constructs pA15-596/CAT, pC1-526/CAT, pA15-629/CAT or pC1-553/CAT which contain three (pA15-596/CAT, pC1-526/CAT), 36 (pA15-629/CAT) or 30 (pC1-553/CAT) nts of the GBV coding sequence ligated in-frame with CAT. These results demonstrate, rather surprisingly, that sequences downstream of the predicted initiator AUG are necessary for efficient translation initiation *in vitro*. Given that the authentic initiator codons are in good context in both GBV-A (SEQUENCE ID NO 23) and GBV-C (SEQUENCE ID NO 4), these data provide further evidence that translation is not initiated by a conventional 5' scanning mechanism.

The quantity of CAT produced from the control plasmid, pHAV-CAT1 (seen in FIGURE 5B, lane 9), was considerably greater than that produced from either the GBV-A (SEQUENCE ID NO 23) or GBV-C (SEQUENCE ID NO 4) monocistronic constructs. This is of interest, because the HAV IRES has been known to direct the internal initiation of translation with very low efficiency relative to other picornaviral IRES elements (L. E. Whetter et al., J. Virol. 68:5253-5263 [1994]). The low production of GBV-CAT proteins was believed not likely to be due to differences in T7 transcriptional efficiency in these IVTT assays, as similar results were obtained with reactions programmed with equal amounts of RNA (data not shown). Thus, it appears that the level of GBV-CAT protein reflects the extremely low efficiency with which the GBV IRESs direct internal initiation *in vitro*.

4. Translation of bicistronic GBV RNAs. In an effort to formally demonstrate that the 5' RNA sequences of GBV-A and GBV-C contain IRESs, these sequences were inserted between CAT and luciferase (Luc) genes to create bicistronic T7 transcriptional units. These results are graphically shown in FIGURE 6A. IVTT reactions programmed with the bicistronic constructs produced equivalent amounts of CAT activity and CAT protein, as shown in FIGURE 6B). Referring to FIGURE 6B, CAT activity was equivalent in the reactions shown (157,000 ± 3,550 cpm). The PhosphorImager scan was generated from a 72 h exposure with a linear range of 25 to 600. Band volumes are reported in FIGURE 6B without background subtraction. This confirmed that essentially equivalent amounts of RNA were being transcribed in each reaction. In contrast, the level of Luc activity and amount of Luc protein produced was dependent on the sequence cloned into the intercistronic space upstream of Luc. Although much less than the level of Luc produced from two positive control plasmids containing the IRES of HCV in the intercistronic space (270,000 to 540,000 light units, FIGURE 6B), detectable levels of Luc activity were produced only in reactions programmed with GBV bicistronic constructs containing GBV-A (SEQUENCE ID NO 23) and GBV-C sequences (SEQUENCE ID NO 4) in the sense orientation (10,300 to 13,300 light units, FIGURE 6B). Although the quantities of Luc produced were barely detectable by SDS-PAGE, PhosphorImager analysis of these gels indicated that Luc enzymatic activity did not correlate with the protein detected in the IVTT assays (FIGURE 6B, Luc-A versus Luc-P). This was most likely due to altered activity as a result of the GBV fusion. Of greater importance, however, was the fact that no detectable protein and only minimal Luc activities (130 and 2020 light units) were produced in reactions programmed with bicistronic constructs containing GBV-A (SEQUENCE ID NO 23) and GBV-C sequences (SEQUENCE ID NO 4) in the antisense orientation. These results suggest that these viruses utilize internal ribosome entry for initiation of translation, but the extraordinarily low activities of the putative GBV IRES elements when placed in a bicistronic context raises a number of issues which are discussed hereinbelow.

5. Secondary structure of the 5' NTR of GBV-C. The results presented above suggested that translation of the GBV-A and GBV-C polyproteins is initiated by an unusual mechanism of internal ribosomal entry, which is likely to be controlled by RNA structures within the 5' nontranslated RNA, and which is also dependent upon sequence downstream of the initiator AUG (see FIGURE 5).

Thus, we attempted to characterize the secondary structure near the 5' end of GBV-C RNA using a combination of phylogenetic analysis and thermodynamic predictions. Covariant nucleotide substitutions indicative of conserved base-pair interactions were identified by manual search of an alignment of 41 different GBV-C sequences. These were used to constrain the folding of the RNA by the computer program, MFOLD. Alternative structures were reviewed to determine which were most permissive for observed variations in the nucleotide sequence, resulting in the model for secondary structure shown in FIGURE 7A-D. Referring to FIGURE 7A-D, the model structure resulted from a combination of phylogenetic analysis and computational thermodynamic prediction. With minor variation, the structure shown can be assumed by all available known GBV-C sequences. The predicted secondary structure of the 5' NTR of GBV-C is very different from that of HCV (E. A. Brown et al., Nuc. Acid Res. 20:5041-5045 [1992] and M. Honda et al., manuscript submitted) suggesting that the 5' NTRs of these viruses have distinctly different evolutionary histories.

The model suggests that the 5' RNA of GBV-C contains 4 major secondary structure domains upstream of the authentic initiator AUG at nt 524 which is conserved in all GBV-C sequences (domains I - IV in FIGURE 7A-D). Domain I consists of an extended stem-loop structure, which is highly conserved in nucleotide sequence between nts 68-152, but which contains several covariant nucleotide substitutions within the flanking RNA segments near its base (FIGURE 7A-D, boxed base pairs). The predicted structure of the conserved sequence between nts 68-152 is confirmed by the presence of covariant nucleotide substitutions in alignments of GBV-C with GBV-A, which shares a very similar overall 5' NTR secondary structure (not shown). Domain II contains two small stem-loops (IIa and IIb), both of which are supported by the presence of covariant substitutions in different GBV-C strains. The larger, complex stem-loops which comprise domains III and IV of the model structure are also well supported by covariant substitutions among different GBV-C strains (FIGURE 7A-D). Of particular interest, given the requirement for the inclusion of coding sequence for efficient translation of monocistronic GBV transcripts (FIGURE 5), is evidence suggesting the existence of a very stable, conserved stem-loop containing 9-10 G-C base-pairs within the ORF, downstream of the putative 5' NTR (see below) (FIGURE 7A-D). The existence of this stable helical structure is supported by the presence of a single covariant substitution among different GBV-C strains. This stem-loop appears to be an extension of a larger, well conserved structure (domain

V, FIGURE 7A-D), located 20 nts downstream of the putative initiator AUG. Importantly, a very similar structure is present near the 5' end of the ORF of GBV-A (FIGURE 7A-D, inset).

F. Discussion.

Monocistronic mRNAs containing the 5' ends of the GBV-A and GBV-C genomic RNAs fused to CAT directed the production of GBV-CAT fusion proteins in IVTT reactions. Site-specific mutagenesis and Edman degradation of the translation products indicated that translation of these transcripts, and presumably GBV-A and GBV-C genomic RNAs as well, initiates immediately upstream of the putative E1 envelope signal sequence, at the AUG located at nt 594 in GBV-A (SEQUENCE ID NO 23) and nt 524 in the GBV-C sequence (SEQUENCE ID NO 4). The site of initiation identified in GBV-C is corroborated by analysis of the 5' RNA sequences obtained from 35 different GBV-C positive individuals. When these sequences are aligned, the only conserved AUG codon which is in-frame with the GBV-C polyprotein is the AUG at nt 524. Downstream of this AUG codon, nucleotide substitutions in the different GBV-C strains generally result in either silent or conservative amino acid changes. In contrast, upstream of this AUG codon nucleotide substitutions, deletions and insertions drastically change the encoded amino acid sequence in different strains. These data suggest that there is a selective pressure acting downstream of the AUG at nt 524 to maintain a protein coding sequence while no selective pressure exists to maintain such a sequence upstream of this codon.

The fact that translation initiates at the fifth AUG codon in both viral RNAs, many hundreds of nucleotides from the 5' end, is strongly reminiscent of translation in the picornaviruses and HCV, and suggests that translation may be initiated by binding of the 40S ribosomal subunit at an internal site on the RNA. Thus, it seems likely that the 5' NTRs of these viruses may contain an IRES. Because the functional activities of the IRES elements of HCV and the picornaviruses are known to be highly dependent on RNA secondary structure within the 5' NTR, we sought evidence for conserved secondary RNA structures within the 5' NTRs of these viruses. Although the 5' nucleotide sequences of the GBV-C and GBV-A virus genomes have only ~50% nucleotide identity within the 500 nts preceding the initiator AUG of GBV-C, we found the secondary structures of these RNAs to be remarkably similar. Each of the major secondary structural domains shown for GBV-C in FIGURE 7A-D is conserved in the structure of GBV-A with only minimal changes (data not shown). However, both the GBV-A and GBV-C 5' NTR structures are very different from those of the pestiviruses, HCV, and GBV-B, despite the fact that these viruses share a common genome organization as well as multiple sequence motifs within their nonstructural proteins (T. P. Leary et al., supra and A. S. Muerhoff et al., supra). While the 5' NTRs of GBV-B, HCV and the pestiviruses are particularly closely related to each other at the structural level (E. A. Brown et al., supra and M. Honda et al., supra), the prominent domain III pseudoknot and complex stem-loop III structures of these viruses are completely lacking in GBV-C and GBV-A. In addition there is no clear-cut structural relatedness to HCV or the pestiviruses in any of the upstream secondary structures of GBV-A and GBV-C. Thus, similar to the existence of two distinct types of 5' NTR structures among the picornaviruses (one in the cardioviruses, aphthoviruses, and hepatoviruses, and another in the enteroviruses and rhinoviruses [R. J. Jackson et al., Mol. Biol. Reports 19:147-159 {1994}]), there are two distinct types of 5' NTR structures present in the flaviviruses. This has interesting implications for the evolution of these agents.

A prominent feature of the 5' NTR sequences of GBV-C and GBV-A is the presence of a short oligopyrimidine tract located just upstream of the initiator AUG. While this tract is somewhat variable in sequence, it is present in all of the GBV-C sequences and is positioned approximately 21 nts upstream of the initiator AUG. Thus, this region of the 5' NTR bears remarkable similarity to the "box A" / "box B" motif identified at the 3' end of picornaviral 5' NTRs by Pilipenko et al. (E. V. Pilipenki et al., Cell 68:119-131 [1992]), including the distance (20 to 25 nts) between the start of the pyrimidine tract and the first downstream AUG in GBV-C (the initiator AUG), which Pilipenko et al. found to be critical to poliovirus IRES-directed translation. It is interesting that the segment intervening between the oligopyrimidine tract and the first downstream AUG is somewhat shorter in the GBV-A viruses (approximately 17 nts). By analogy with the picornaviruses (Pilipenko et al., supra), this might be expected to result in a preference for initiation of translation at the second in-frame AUG codon in GBV-A (nt +25 with respect to the first AUG). We confirmed this experimentally (see FIGURE 4A). The striking differences between the 5' NTR structures of these viruses and that of HCV, coupled with these similarities between the translation of GBV-A and GBV-C and picornaviral 5' NTRs, suggests that the mechanism of translation might be closer to that of picornaviruses than HCV. In HCV, relatively strong evidence supports the concept that the 40S ribosomal subunit binds RNA directly at the site of translation initiation (Honda et al., supra). In contrast, the

40S subunit appears to scan for a variable distance from an upstream primary binding site to the initiator AUG in some picornaviruses (R. J. Jackson et al, supra). Given the variable distances between the authentic initiator codons and the upstream oligopyrimidine tracts in GBV-A and GBV-C, this appears likely to be the case with GBV-A (and possibly also GBV-C).

Both GBV-A and GBV-C contain a very stable stem-loop structure within the translated open reading frame (domain V, FIGURE 7D.). This conserved structure is located about 20 nts downstream of the initiator AUG in GBV-C, although it is possible that additional, less well conserved base-pair interactions may bring the base of this structure closer to the AUG. It is tempting to speculate that this stem-loop may function to enhance initiation by a scanning 40S ribosomal subunit, much as M. Kozak, Proc. Natl. Acad. Sci USA 87:8301-8305 (1990) has shown that stable stem-loops placed downstream of an AUG can result in a "pausing" of the ribosome over the AUG, enhancing the likelihood of initiation at that codon. This phenomenon may explain why the efficient translation of reporter proteins fused to the 5' NTR requires inclusion of the most 5' sequence of the GBV-C open reading frame. If so, this would provides a novel mechanism by which sequence within the open-reading frame can contribute to regulation of translation in flaviviruses. Both HCV and the GBV-B viruses differ from GBV-A and GBV-C in that their initiator AUG is located within the loop segment of a stem-loop which straddles the 5' end of the open reading frame (M. Honda, supra). Initiation of translation of these viral RNAs is thus dependent upon melting of this stem-loop while, in the case of GBV-A and GBV-C, initiation of translation is likely to be dependent on maintenance of the integrity of the domain V stem-loop.

The domain V stem-loop for which is required for efficient translation of the monocistronic transcripts does not appear to be required for efficient translation in the bicistronic transcripts (compare FIGURES 5 and 6). This apparent discrepancy may be a result of the different reporter genes being utilized in these transcripts. Similar findings have been reported for HCV. Specifically, Reynolds et al., supra, using bicistronic vectors with the IRES-dependent reporter genes secreted alkaline phosphatase or a truncated influenza virus nonstructural protein, show efficient translation directed by the 5' end of HCV requires the inclusion of coding sequences. In contrast, Wang et al., supra, using monocistronic and bicistronic vectors with luciferase as the IRES-dependent reporter gene, find the inclusion of HCV coding sequences is not necessary for efficient translation.

Addressing these conflicting results, Reynolds et al., supra, hypothesize that the 5' end of the luciferase gene may complement the function provided by the HCV coding sequences. A similar argument may explain the discordance between the results obtained with the monocistronic GBV-CAT constructs and the bicistronic GBV-Luc constructs.

Although all of these observations suggest the strong likelihood that GBV-A and GBV-C translation is initiated by internal ribosomal entry, only minimal translation of the downstream cistron was noted from bicistronic transcripts containing the 5' NTRs of these viruses in the intercistronic space. Translation directed by the GBV-A and GBV-C 5' NTRs within a bicistronic context was only 2 to 5% that of the HCV IRES in rabbit reticulocyte lysates *in vitro* (FIGURE 6). The very low activities of the GBV-A and GBV-C IRESs suggest several possibilities. First, it is possible that these viruses may in fact have IRES elements with extraordinarily low activity. This is supported by a very low level of translation directed by monocistronic transcripts containing the 5' ends of GBV-A and GBV-C in the *in vitro* system. Specifically, after adjustment for the number of Cys residues in each construct, GBV-CAT fusion proteins were translated from pA15-707/CAT and pC1-631/CAT transcripts at only 20 to 41% of the level produced by the IRES of HAV. The HAV IRES is known to have very low activity, in the range of 2% of the Sabin poliovirus type I IRES within HAV permissive cells (see, D. E. Schultz et al., J. Virol. 70:1041-1049 [1996] and L. E. Whetter et al., supra). Thus, the low GBV IRES activity noted *in vitro* may be a true reflection of the strength of these translation elements. Limiting production of viral proteins within an infected host might act to reduce recognition of the infection by the immune system and thus promote viral persistence. Alternatively, it is possible that the low IRES activity detected in reticulocyte lysates reflects a requirement for a specific host cell translation factor which is absent in reticulocyte lysates. The nuclear autoantigen, La, is an example of such a specific cellular factor. It is required for efficient translation directed by the poliovirus IRES, but is not present in sufficient amounts in reticulocyte lysates. K. Meerovitch et al., J. Virol. 67:3798-3807 (1993). It is difficult to comment more specifically on this possibility, since the cellular tropisms of GBV-A and GBV-C are unknown. Yet a third possibility is that the low translational activity of the GBV-A and GBV-C 5' NTRs may reflect a requirement for additional, yet to be identified 5' viral sequences that may be present in these viral genomes. It is also conceivable that translation is initiated by a mechanism distinct from both the classic 5' scanning and IRES-directed translation initiation mechanism. For example, relatively efficient translation initiation at an internal site in monocistronic transcripts but low translational activity in the bicistronic context could be explained by a mechanism involving "ribosome shunting" (J. Fütterer et al., Cell 73:789-802 [1993]) following recognition of the 5' end of the RNA by the 40S ribosome subunit. Further studies will be required to distinguish between these different possibilities.

The proteins located at or near the amino termini of the polyproteins of yellow fever virus (protein C), a flavivirus, bovine viral diarrhea virus, a pestivirus, and HCV (core) are small and highly basic (Q.-L. Choo et al., Proc. Natl. Acad. Sci. USA 88:2451-2455 [1991]; M.S. Collett et al., supra; R. H. Miller et al., Proc. Natl. Acad. Sci. USA 87:2057-2061 [1990]). Because GBV-A and GBV-C are phylogenetically related to these viruses (12, 18) it was expected that such a protein would be encoded in these viruses. However, the position of the initiation codons in GBV-A and GBV-C eliminates the possibility of a basic core protein being located at the N-termini of the viral polyproteins. The possibility that the core coding sequences may have been deleted during RT-PCR amplification or cloning of the 5' ends of GBV-A and GBV-C is unlikely for several reasons. First, identical deletions would have had to occur consistently in each of the several clones generated during the sequencing of GBV-A and GBV-C, in addition to the 42 separate GBV-C isolates described by U.S. Serial No. 08/580,038, filed December 21, 1995 and previously incorporated herein by reference, and the 2 HGV isolates described by Linnen et al., supra. This consistency, in addition to the correspondence between PCR and infective titers for GBV-A (G. G. Schlauder et al., J. Med. Virol. 46:81-90 [1995] and J. N. Simons , Proc. Natl. Acad. Sci USA, supra), argues against GBV-A and GBV-C sequences being derived from defective interfering particles in the cloning sources. Second, the deletion of core sequences would have had to occur without disturbing the translational activity of the 5' ends of these viruses. But because proper initiation requires sequences located in the coding regions of GBV-A and GBV-C, the coupling between the translational activity and the coding regions appear to make this an impossibility. Finally, several RT-PCR experiments using different virus isolates, different primer combinations, and different RT-PCR conditions and polymerases provide no evidence for additional virus sequence (data not shown).

The lack of a core-like protein at the N-terminus of the viral polyprotein distinguishes GBV-A and GBV-C from all other members of the *Flaviviridae*. In fact, searches of all six potential reading frames of the three full length GBV-C sequences (T. P. Leary et al., supra and L. Linnen et al., supra) or the GBV-A sequence (SEQUENCE ID NO. 23) present in GenBank does not reveal a conserved open reading frame encoding a core-like protein. Thus, these viruses appear distinct from enveloped viruses in general as they do not appear to encode a basic protein which mediates the packaging of the viral nucleic acid into the virion envelope. Core-less infectious particles have been generated artificially using the vesicular somatitis virus glycoprotein. M. M. Rolls et al., Cell 79:497-506 (1994). Thus, it is possible that GBV-A and GBV-C may be truly "core-less" enveloped viruses. However, it is possible that a cellular RNA-binding protein has been appropriated by these viruses to facilitates the specific and efficient packaging of the virion RNA into the envelope. Whether GBV-A and GBV-C contain core proteins and the source of these cores awaits the biochemical characterization of these viruses.

The present invention is intended to be limited only by the appended claims.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Simons, J. N.
Desai, S. M
Mushahwar, I. K.

(ii) TITLE OF INVENTION: REAGENTS AND METHODS USEFUL FOR CONTROLLING THE TRANSLATION OF HEPATITIS GB PROTEINS

(iii) NUMBER OF SEQUENCES:32

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: Abbott Laboratories
(B) STREET: 100 Abbott Park Rd
(C) CITY: Abbott Park
(D) STATE: IL
(E) COUNTRY: USA
(F) ZIP: 60064

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.30

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER:
(B) FILING DATE:
(C) CLASSIFICATION:

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: Porembski, Priscilla E.
(B) REGISTRATION NUMBER: 33,207
(C) REFERENCE/DOCKET NUMBER: 5793.US.P1

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 708-937-0378
(B) TELEFAX: 708-938-2623

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

CCACAAACAC TCCAGTTTGT TAC 23

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

GCTCTAGACA TGTGCTACGG TCTACGAG 28

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 9126 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

CCCCCCCCCC GGCACTGGGT GCAAGCCCCA GAAACCGACG CCTACTGAAG TAGACGTAAT 60

GGCCCCGCGC CGAACCGGCG ACCGGCCAAA AGGTGGTGGA TGGGTGATGA CAGGGTTGGT 120

AGGTCGTAAA TCCCGGTCAT CCTGGTAGCC ACTATAGGTG GGTCTTAAGG GGAGGCTACG 180

GTCCCTCTTG CGCATATGGA GGAAAAGCGC ACGGTCCACA GGTGTTGGTC CTACCGGTGT 240

AATAAGGACC CGGCGCTAGG CACGCCGTTA AACCGAGCCC GTTACTCCCC TGGGCAAACG 300

ACGCCCACGT ACGGTCCACG TCGCCCTTCA ATGTCTCTCT TGACCAATAG GCGTAGCCGG 360

CGAGTTGACA AGGACCAGTG GGGGCCGGGC GGGAGGGGGA AGGACCCCCA CCGCTGCCCT 420

TCCCGGGGAG GCGGGAAATG CATGGGGCCA CCCAGCTCCG CGGCGGCCTA CAGCCGGGGT 480

AGCCCAAGAA CCTTCGGGTG AGGGCGGGTG GCATTTCTTT TCCTATACCG ATCATGGCAG 540

TCCTTCTGCT CCTACTCGTG GTGGAGGCCG GGGCTATTTT AGCCCCGGCC ACCCATGCTT 600

GTAGCGCGAA AGGGCAATAT TTBCTCACAA ACTGTTGCGC CCTGGAGGAC ATAGGCTTCT 660

```
GCCTGGAGGG CGGATGCCTG GTGGCTCTGG GGTGCACCAT TTGCACCGAC CGCTGCTGGC      720
CACTGTATCA GGCGGGTTTG GCCGTGCGGC CCGGCAAGTC CGCCGCCCAG TTGGTGGGGG      780
AACTCGGTAG TCTCTACGGG CCCTTGTCGG TCTCGGCTTA TGTGGCCGGG ATCCTGGGGC      840
TTGGGGAGGT CTACTCGGGG GTCCTCACCG TCGGGGTGGC GTTGACGCGC AGGGTCTACC      900
CGGTCCCGAA CCTGACGTGT GCAGTAGAGT GTGAGTTGAA GTGGGAAAGT GAGTTTTGGA      960
GATGGACTGA ACAGCTGGCC TCAAACTACT GGATTCTGGA ATACCTCTGG AAGGTGCCTT     1020
TCGACTTTTG GCGGGGAGTG ATGAGCCTTT CTCCTCTCTT GGTGTGCGTG GCGGCCCTCC     1080
TCCTGCTGGA GCAGCGTATT GTCATGGTCT TCCTCCTGGT CACTATGGCG GGCATGTCAC     1140
AAGGCGCGCC CGCCTCAGTG TTGGGGTCAC GGCCTTTCGA GGCCGGGCTG ACTTGGCAGT     1200
CTTGTTCTTG CAGGTCGAAC GGGTCCCGCG CGCCGACAGG GGAGAGGGTT TGGGAACGTG     1260
GGAACGTCAC ACTTTTGTGT GACTGCCCCA ACGGTCCTTG GGTGTGGGTC CCGGCCCTTT     1320
GCCAGGCAAT CGGATGGGGC GACCCTATCA CTCATTGGAG CCACGGACGA AATCAGTGGC     1380
CCCTTTCTTG TCCCCAATTT GTCTACGGCG CCGTTTCAGT GACCTGCGTG TGGGGTTCTG     1440
TGTCTTGGTT TGCTTCCACT GGGGGTCGCG ACTCCAAGGT TGATGTGTGG AGTTTGGTTC     1500
CAGTTGGCTC TGCCAGCTGT ACCATAGCCG CACTGGGATC TTCGGATCGC GACACAGTGG     1560
TTGAGCTCTC CGAATGGGGA ATCCCCTGCG CCACTTGTAT CCTGGACAGG CGGCCTGCCT     1620
CGTGTGGCAC CTGTGTGAGG GACTGCTGGC CCGAGACCGG GTCGGTACGT TTCCCATTCC     1680
ACAGGTGTGG CGCGGGACCG AGGCTGACCA GAGACCTTGA GGCTGTGCCC TTCGTCAATA     1740
GGACAACTCC CTTCACCATA AGGGGGCCCC TGGGCAACCA GGGGCGAGGC GACCCGGTGC     1800
GGTCGCCCTT GGGTTTTGGG TCCTACACCA TGACCAAGAT CCGAGACTCC TTACACTTGG     1860
TGAAATGTCC CACCCCAGCC ATTGAGCCTC CCACCGGAAC GTTTGGGATC TTCCCAGGAG     1920
TCCCCCCCCT TAACAACTGC ATGCTTCTCG GCACTGAGGT GTCAGAGGTA TTGGGTGGGG     1980
CGGGCCTCAC TGGGGGGTTT TACGAACCTC TGGTGCGGCG GTGTTCAGAG CTGATGGGTC     2040
GGCGGAATCC GGTCTGCCCG GGGTTTGCAT GGCTCTCTTC GGGACGGCCT GATGGGTTCA     2100
TACATGTACA GGGCCACTTG CAGGAGGTGG ATGCGGGCAA CTTCATTCCG CCCCCACGCT     2160
GGTTGCTCTT GGACTTTGTA TTTGTCCTGT CATACCTGAT GAAGCTGGCA GAGGCACGGT     2220
TGGTCCCGCT GATCCTCCTC CTGCTATGGT GGTGGGTGAA CCAGTTGGCG GTCCTTGKAC     2280
TGSCGGCTGC KCRCGCCGCC GTGGCTGGAG AGGTGTTTGC GGGCCCTGCC TTGTCCTGGT     2340
GTCTGGGCCT ACCCTTCGTG AGTATGATCC TGGGGCTAGC AAACCTGGTG TTGTACTTCC     2400
```

```
GCTGGATGGG TCCTCAACGC CTGATGTTCC TCGTGTTGTG GAAGCTCGCT CGGGGGGCTT    2460
TCCCGCTGGC ATTACTGATG GGGATTTCCG CCACTCGCGG CCGCACCTCT GTGCTTGGCG    2520
CCGAATTCTG CTTTGATGTC ACCTTTGAAG TGGACACGTC AGTCTTGGGT TGGGTGGTTG    2580
CTAGTGTGGT GGCTTGGGCC ATAGCGCTCC TGAGCTCTAT GAGCGCGGGG GGGTGGAAGC    2640
ACAAAGCCAT AATCTATAGG ACGTGGTGTA AAGGGTACCA GGCYCTTCGC CAGCGCGTGG    2700
TGCGTAGCCC CCTCGGGGAG GGGCGGCCCA CCAAGCCGCT GACGATAGCC TGGCGTCTGG    2760
CCTCTTACAT CTGGCCGGAC GCTGTGATGT TGGTGGTTGT GGCCATGGTC CTCCTCTTCG    2820
GCCTTTTCGA CGCGCTCGAT TGGGCCTTGG AGGAGCTCCT TGTGTCGCGG CCTTCGTTGC    2880
GTCGTTTGGC AAGGGTGGTG GAGTGTTGTG TGATGGCGGG CGAGAAGGCC ACTACCGTCC    2940
GGCTTGTGTC CAAGATGTGC GCGAGAGGGG CCTACCTGTT TGACCACATG GGGTCGTTCT    3000
CGCGCGCGGT CAAGGAGCGC TTGCTGGAGT GGGACGCGGC TTTGGAGMCC CTGTCATTCA    3060
CTAGGACGGA CTGCCGCATC ATACGAGACG CCGCCAGGAC TCTGAGCTGC GGCCAATGCG    3120
TCATGGGCTT GCCCGTGGTG GCTAGGCGCG GCGATGAGGT CCTGGTTGGG GTCTTTCAGG    3180
ATGTGAACCA CTTGCCTCCG GGGTTTGYTC CTACAGCGCC TGTTGTCATC CGTCGGTGCG    3240
GAAAGGGCTT CCTCGGGGTC ACTAAGGCTG CCTTGACTGG TCGGGATCCT GACTTACACC    3300
CAGGAAACGT CATGGTTTTG GGGACGGCTA CCTCGCGCAG CATGGGAACG TGCTTAAACG    3360
GGTTGCTGTT CACGACATTC CATGGGGCTT CTTCCCGAAC CATTGCGACA CCTGTGGGGG    3420
CCCTTAACCC AAGGTGGTGG TCGGCCAGTG ATGACGTCAC GGTCTATCCC CTCCCCGATG    3480
GAGCTAACTC GTTGGTTCCC TGCTCGTGTC AGGCTGAGTC CTGTTGGGTC ATYCGATCCG    3540
ATGGGGCTCT TTGCCATGGC TTGAGCAAGG GGGACAAGGT AGAACTGGAC GTGGCCATGG   ·3600
AGGTTGCTGA CTTTCGTGGG TCGTCTGGGT CTCCTGTCCT ATGCGACGAG GGGCACGCTG    3660
TAGGAATGCT CGTGTCCGTC CTTCATTCGG GGGGGAGGGT GACCGCGGCT CGATTCACTC    3720
GGCCGTGGAC CCAAGTCCCA ACAGACGCCA AGACTACCAC TGAGCCACCC CCGGTGCCAG    3780
CTAAAGGGGT TTTCAAAGAG GCTCCTCTTT TCATGCCAAC AGGGGCGGGG AAAAGCACAC    3840
GCGTCCCTTT GGAATATGGA AACATGGGGC ACAAGGTCCT GCTTCTCAAC CCGTCGGTTG    3900
CCACTGTGAG GGCCATGGGC CCTTACATGG AGAAGCTGGC GGGGAAACAT CCTAGCATTT    3960
TCTGTGGACA CGACACAACA GCTTTCACAC GGATCACGGA CTCTCCATTG ACGTACTCTA    4020
CCTATGGGAG GTTTCTGGCC AACCCGAGGC AGATGCTGAG GGGAGTTTCC GTGGTCATCT    4080
```

```
GTGATGAGTG CCACAGTCAT GACTCAACTG TGTTGCTGGG TATAGGCAGG GGCAGGGAGC   4140
TGGCGCGGGG GTGTGGAGTG CAATTAGTGC TCTACGCTAC TGCGACTCCC CCGGGCTCGC   4200
CTATGACTCA GCATCCATCC ATAATTGAGA CAAAGCTGGA CGTCGGTGAG ATCCCCTTTT   4260
ATGGGCATGG TATCCCCCTC GAGCGTATGA GGACTGGTCG CCACCTTGTA TTCTGCCATT   4320
CCAAGGCGGA GTGCGAGAGA TTGGCCGGCC AGTTCTCCGC GCGGGGGGTT AATGCCATCG   4380
CCTATTATAG GGGTAAGGAC AGTTCCATCA TCAAAGACGG AGACCTGGTG GTTTGTGCGA   4440
CAGACGCGCT CTCTACCGGG TACACAGGAA ACTTCGATTC TGTCACCGAC TGTGGGTTAG   4500
TGGTGGAGGA GGTCGTTGAG GTGACCCTTG ATCCCACCAT TACCATTTCC TTGCGGACTG   4560
TCCCTGCTTC GGCTGAATTG TCGATGCAGC GGCGCGGACG CACGGGGAGA GGTCGGTCGG   4620
GCCGCTACTA CTACGCTGGG GTCGGTAAGG CTCCCGCGGG GGTGGTGCGG TCTGGTCCGG   4680
TCTGGTCGGC AGTGGAAGCT GGAGTGACCT GGTATGGAAT GGAACCTGAC TTGACAGCAA   4740
ACCTTCTGAG ACTTTACGAC GACTGCCCTT ACACCGCAGC CGTCGCAGCT GACATTGGTG   4800
AAGCCGCGGT GTTCTTTGCG GGCCTCGCGC CCCTCAGGAT GCATCCCGAT GTTAGCTGGG   4860
CAAAAGTTCG CGGCGTCAAT TGGCCCCTCC TGGTGGGTGT TCAGCGGACG ATGTGTCGGG   4920
AAACACTGTC TCCCGGCCCG TCGGACGACC CTCAGTGGGC AGGTCTGAAA GGCCCGAATC   4980
CTGCCCCACT ACTGCTGAGG TGGGGCAATG ATTTGCCATC AAAAGTGGCC GGCCACCACA   5040
TAGTTGACGA TCTGGTCCGT CGGCTCGGTG TGGCGGAGGG ATACGTGCGC TGTGATGCTG   5100
GRCCCATCCT CATGGTGGGC TTGGCCATAG CGGGCGGCAT GATCTACGCC TCTTACACTG   5160
GGTCGCTAGT GGTGGTAACA GACTGGAATG TGAAGGGAGG TGGCAATCCC CTTTATAGGA   5220
GTGGTGACCA GGCCACCCCT CAACCCGTGG TGCAGGTCCC CCCGGTAGAC CATCGGCCGG   5280
GGGGGGAGTC TGCGCCAGCG GATGCCAAGA CAGTGACAGA TGCGGTGGCA GCCATCCAGG   5340
TGAACTGCGA TTGGTCTGTG ATGACCCTGT CGATCGGGGA AGTCCTCACC TTGGCTCAGG   5400
CTAAGACAGC CGAGGCCTAC GCAGCTACTT CCAGGTGGCT CGCTGGCTGC TACACGGGGA   5460
CGCGGGCCGT CCCCACTGTA TCAATTGTTG ACAAGCTCTT CGCCGGGGGT TGGGCCGCCG   5520
TGGTGGGTCA CTGTCACAGC GTCATTGCTG CGGTGGTGGC TGCCTATGGG GTTTCTCGAA   5580
GTCCTCCACT GGCCGCGGCG GCATCCTACC TCATGGGGTT GGGCGTCGGA GGCAACGCAC   5640
AGGCGCGCTT GGCTTCAGCT CTTCTACTGG GGGCTGCTGG TACGGCTCTG GGGACCCCTG   5700
TCGTGGGACT CACCATGGCG GGGGCCTTCA TGGGCGGTGC CAGCGTGTCC CCCTCGCTCG   5760
TCACTGTCCT ACTTGGGGCT GTGGGAGGTT GGGAGGGCGT TGTCAACGCT GCCAGTCTCG   5820
```

```
TCTTCGACTT CATGGCTGGG AAACTTTCAA CAGAAGACCT TTGGTATGCC ATCCCGGTAC    5880
TCACTAGTCC TGGRGCGGGC CTCGCGGGGA TTGCCCTTGG TCTGGTTTTG TACTCAGCAA    5940
ACAACTCTGG CACTACCACA TGGCTGAACC GTCTGCTGAC GACGTTGCCA CGGTCATCTT    6000
GCATACCCGA CAGCTACTTC CAACAGGCTG ACTACTGCGA CAAGGTCTCG GCAATGCTGC    6060
GCCGCCTGAG CCTTACTCGC ACCGTGGTGG CCCTGGTCAA CAGGGAGCCT AAGGTGGATG    6120
AGGTCCAGGT GGGGTACGTC TGGGATCTGT GGGAGTGGGT AATGCGCCAG GTGCGCATGG    6180
TGATGTCTAG ACTCCGGGCC CTCTGCCCTG TGGTGTCACT CCCCTTGTGG CACCGCGGGG    6240
AGGGGTGGTC CGGTGAATGG CTTCTCGATG GGCACGTGGA GAGTCGTTGT CTGTGCGGGT    6300
GTGTAATCAC CGGCGACGTC CTCAATGGGC AACTCAAAGA TCCAGTTTAC TCTACCAAGC    6360
TGTGCAGGCA CTACTGGATG GGAACTGTGC CGGTCAACAT GCTGGGCTAC GGGGAAACCT    6420
CACCTCTTCT CGCCTCTGAC ACCCCGAAGG TGGTACCCTT CGGGACGTCG GGGTGGGCTG    6480
AGGTGGTGGT GACCCCTACC CACGTGGTGA TCAGGCGCAC GTCCTGTTAC AAACTGCTTC    6540
GCCAGCAAAT TCTTTCAGCA GCTGTAGCTG AGCCCTACTA CGTTGATGGC ATTCCGGTCT    6600
CTTGGGAGGC TGACGCGAGA GCGCCGGCCA TGGTCTACGG TCCGGGCCAA AGTGTTACCA    6660
TTGATGGGGA GCGCTACACC CTTCCGCACC AGTTGCGGAT GCGGAATGTG GCGCCCTCTG    6720
AGGTTTCATC CGAGGTCAGC ATCGAGATCG GGACGGAGAC TGAAGACTCA GAACTGACTG    6780
AGGCCGATTT GCCACCAGCG GCTGCTGCCC TCCAAGCGAT AGAGAATGCT GCGAGAATTC    6840
TCGAACCGCA CATCGATGTC AYCATGGAGG ATTGCAGTAC ACCCTCTCTC TGTGGTAGTA    6900
GCCGAGAGAT GCCTGTGTGG GGAGAAGACA TACCCCGCAC TCCATCGCCT GCACTTATCT    6960
CGGTTACGGA GAGCAGCTCA GATGAGAAGA CCCTGTCGGT GACCTCCTCG CAGGAGGACA    7020
CCCCGTCCTC AGACTCATTT GAAGTCATCC AAGAGTCTGA TACTGCTGAA TCAGAGGAAA    7080
GCGTCTTCAA CGTGGCTCTT TCCGTACTAA AAGCATTATT TCCACAGAGC GTTGCCACAC    7140
GAAAGCTAAC GGTTAAGATG TCTTGCTGTG TTGAGAAGAG CGTAACACGC TTCTTTTCTT    7200
TAGGGTTGAC CGTGGCTGAC GTGGCTAGCC TGTGTGAGAT GGAGATCCAG AACCATACAG    7260
CCTATTGTGA CAAGGTGCGC ACTCCGCTCG AATTGCAAGT TGGGTGCTTG GTGGGCAATG    7320
AACTTACCTT TGAATGTGAC AAGTGTGAGG CACGCCAAGA GACCCTTGCC TCCTTCTCCT    7380
ACATATGGTC CGGGGTCCCA CTTACTCGGG CCACTCCGGC CAAACCACCA GTGGTGAGGC    7440
CGGTGGGGTC CTTGTTGGTG GCAGACACCA CCAAGGTCTA CGTGACCAAT CCGGACAATG    7500
```

```
TTGGGAGGAG GGTTGACAAG GTGACTTTCT GGCGCGCTCC TCGGGTACAC GACAAGTTCC   7560

TCGTGGACTC GATCGAGCGC GCTCGGAGAG CTGCTCAAGG CTGCCTAAGC ATGGGTTACA   7620

CTTATGAGGA GGCAATAAGG ACTGTTAGGC CGCATGCTGC CATGGGCTGG GGATCTAAGG   7680

TGTCGGTCAG GGACTTGGCC ACCCCTGCGG GGAAGATGGC TGTTCATGAC CGGCTTCAGG   7740

AGATACTTGA AGGGACTCCA GTCCCTTTTA CCCTGACTGT CAAAAAGGAG GTGTTCTTCA   7800

AAGATCGTAA GGAGGAGAAG GCCCCCCGCC TCATTGTGTT CCCCCCCCTG GACTTCCGGA   7860

TAGCTGAAAA GCTCATTCTG GGAGACCCGG GGCGGGTTGC AAAGGCGGTG TGGGGGGGGG   7920

CTTACGCCTT CCAGTACACC CCCAACCAGC GGGTTAAGGA GATGCTAAAG CTGTGGGAAT   7980

CAAAGAAGAC CCCGTGCGCC ATCTGTGTGG ATGCCACTTG CTTCGACAGT AGCATTACTG   8040

ARGAGGACGT GGCACTAGAG ACAGAGCTTT ACGCCCTGGC CTCGGACCAT CCAGAATGGG   8100

TGCGCGCCCT GGGGAAATAC TRTGCCTCTG GCACAATGGT GACCCCGGAA GGGGTGCCAG   8160

TGGGCGAGAG GTATTGTAGG TCCTCGGGTG TGTTAACCAC AAGTGCTAGC AACTGTTTGA   8220

CCTGCTACAT CAAAGTGAGA GCCGCCTGTG AGAGGATCGG ACTGAAAAAT GTCTCGCTTC   8280

TCATCGCGGG CGATGACTGC TTAATTGTGT GCGAGAGGCC TGTATGCGAC CCTTGCGAGG   8340

CCCTGGGCCG AGCCCTGGCT TCGTACGGGT ACGCGTGTGA GCCCTCGTAT CACGCTTCAC   8400

TGGACACAGC CCCCTTCTGC TCCACTTGGC TTGCTGAGTG CAATGCGGAT GGGRAAAGGC   8460

ATTTCTTCCT GACCACGGAC TTTCGGAGAC CACTCGCTCG CATGTCGAGC GAGTACAGTG   8520

ACCCTATGGC TTCGGCCATT GGTTACATTC TCCTCTATCC CTGGCRTCCC ATCACACGGT   8580

GGGTCATCAT CCCGCATGTG CTAACATGCG CTTCTTTCCG GGGTGGTGGC ACACSGTCTG   8640

ATCCGGTTTG GTGTCAGGTT CATGGTAACT ACTACAAGTT TCCCCTGGAC AAACTGCCTA   8700

ACATCATCGT GGCCCTCCAC GGACCAGCAG CGTTGAGGGT TACCGCAGAC ACAACCAAAA   8760

CAAAGATGGA GGCTGGGAAG GTTCTGAGCG ACCTCAAGCT CCCTGGTCTA GCCGTCCACC   8820

GCAAGAAGGC CGGGGCATTG CGAACACGCA TGCTCCGGTC GCGCGGTTGG GCGGAGTTGG   8880

CTAGGGGCCT GTTGTGGCAT CCAGGACTCC GGCTTCCTCC CCCTGAGATT GCTGGTATCC   8940

CAGGGGGTTT CCCTCTGTCC CCCCCCTACA TGGGGGTGGT TCATCAATTG GATTTCACAG   9000

CSCAGCGGAG TCGCTGGCGG TGGTTGGGGT TCTTAGCCCT GCTCATCGTA GCGCTCTTTG   9060

GGTGAACTAA ATTCATCTGT TGCGGCCGGA GTCAGACCTG AGCCCCGTTC AAAAGGGGAT   9120

TGAGAC                                                              9126
```

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 635 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

```
CACTGGGTGC AAGCCCCAGA AACCGACGCC TATCTAAGTA GACGCAATGA CTCGGCGCCA      60

ACTCGGCGAC CGGCCAAAAG GTGGTGGATG GGTGATGACA GGGTTGGTAG GTCGTAAATC     120

CCGGTCACCT TGGTAGCCAC TATAGGTGGG TCTTAAGAGA AGGTTAAGAT TCCTCTTGTG     180

CCTGCGGCGA GACCGCGCAC GGTCCACAGG TGTTGGCCCT ACCGGTGTGA ATAAGGGCCC     240

GACGTCAGGC TCGTCGTTAG ACCGAGCCCG TCACCCACCT GGGCAAACGT CGCCCACGTA     300

CGGTCCACGT CGCCCTTCAA TGTCTCTCTT GACCAATAGG CTTAGCCGGC CGAGTTGACA     360

AGGACCAGTG GGGGTCGGGG GCTTGGGGAG GGACCCCAAG TCCTGCCCTT CCCGGTGGGC     420

CGGGAAATGC ATGGGGCCAC CCAGCTCCGC GGCGGCCTGC AGCCGGGGTA GCCCAAGAAT     480

CCTTCGGGTG AGGGCGGGTG GCATTTCTCT TTTCTATACC ATCATGGCAG TCCTTCTGCT     540

CCTTCTCGTG GTCGAGGCCG GGGCCATTCT GGCCCCGGCC ACCCACGCTT GTCGAGCGAA     600

TGGGGCAATA CTTCCTCACA AACTGTTGCG CCCTG                                635
```

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 32 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

```
TATAATAAGC TTGCCCCGGA CCTCCCACCG AG                                    32
```

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

GCTCTAGATC GGGAACAACA ATTGGAAAG 29

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 33 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

TATAATAAGC TTCACTGGGT GCAAGCCCCA GAA 33

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

GCTCTAGAGG CGCAACAGTT TGTGAGGAA 29

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 26 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

GCTCTAGACA CTGGGTGCAA GCCCCA 26

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

TATAATAAGC TTGGCGCAAC AGTTTGTGAG 30

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

GCTCTAGAGC ACTGGTGCCG CGAGT 25

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GCTCTAGAGA GGGGGAAGCA AACCA 25

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

GCTCTAGACA TGGTGAATGT GTCGACCAC 29

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

GCTCTAGAAC AAGCGTGGGT GGCCGGGG 28

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GCTCTAGAGA CCACGAGAAG GAGCAGAAG 29

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

GCTCTAGACA TGATGGTATA GAAAAGAG 28

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 31 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

CCATAATCAT GAGGGAACAA CAATTGGAAA G 31

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

CATGCCATGG CGCAACAGTT TGTGAGGAA 29

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 34 base pairs
(B) TYPE: nucleic acid

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

TATAATAAAG CTTCACTCCC CTGTGAGGAA CTAC 34

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH:35 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

GTATTGCGTC ATGATGGTTT TTCTTTGGGG TTTAG 35

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 31 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

TATAATAAGC TTGCCGCGAG TTGAAGAGCA C 31

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 33 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

```
CCATAATCAT GAGCCCCCGG ACCTCCCACC GAG                                33
```

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 9493 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

```
CGTGGGAGTC CGGGGCCCCG GACCTCCCAC CGAGGTGGGG GGAAAGGGGC CCTGGACCGG      60
CCGGGTGGAA GGCCCGGAAC CGGTCCATCT TCCTCAAGGT TGAGGAAGGG GTACGTCTAT     120
CGGTCCGGTC GGTCCGAAAG GCGTCTGGAT GCCTAGTGTT AGGGTTCGTA GGTGGTAAAT     180
CCCAGCTAGG CGTGAAAGCG CTATAGGATA GGCTTATCCC GGTGACCGCT GCCCCGGAAC     240
CAGCCCCGCG GKTCTTTGGA CACGGTCCAC AGGTTGGGGG TACCGGTGTG AATAACCCCC     300
CGACTGAAGC GTCAGTCGTT AAACGGAGAC GGTCTCCTGA GATCGCAACG ACGCCCCACG     360
TACGGGAACG CCGCCAAAAC CTTCGGGACA GCTATGCGGG TTGACAATCC CAGTGGGGGG     420
CCGGGGACCA GCTGATTACT TGTCCTGCGA GTTCCTCTTG AGACTGGCCG AAAGGCAGCC     480
ACGGGGCCAC CAAGGCGGCG CAGCGCTGCA TGCGGCAAGG GGAAAAATCC TTCGGGTGAC     540
CCCTGGTGGC AATCCCTTCC CTTAGGAGCA TGAGTGTGGT CGACACATTC ACCATGGCTT     600
GGCTGTGGTT GCTGGTTTGC TTCCCCCTCG CGGGGGGGGT GCTCTTCAAC TCGCGGCACC     660
AGTGCTTCAA TGGGGACCAT TATGTGCTTT CCAATTGTTG TTCCCGAGAC GAGGTTTACT     720
TCTGTTTCGG GGACGGATGT CTGGTGGCTT ATGGCTGTAC TGTTTGCACA CAGTCTTGCT     780
GGAAGCTCTA CCGGCCTGGG GTGGCTACTC GGCCCGGGTC CGAACCAGGT GAGCTGCTGG     840
GGAGATTTGG GAGTGTAATT GGTCCGGTGT CGGCTTCGGC TTACACCGCT GGAGTCCTCG     900
GGTTGGGTGA ACCTTACAGT TTGGCCTTCT TGGGGACGTT CCTCACCAGT CGCCTCTCAC     960
GGATTCCCAA CGTCACCTGC GTGAAGGCTT GTGACCTTGA GTTTACCTAC CCAGGCTTGT    1020
```

```
CCATCGATTT TGACTGGGCG TTTACCAAGA TCTTGCAGTT GCCGGCCAAG CTGTGGCGAG    1080
GCCTAACGGC RGCWCCGGTC TTGAGCCTCC TCGTGATCCT CATGCTGGTC CTCGAGCAGC    1140
GCCTCCTGAT AGCCTTCCTA CTGCTTTTGG TAGTGGGCGA GGCTCAGAGG GGGATGTTCG    1200
ACAACTGCGT GTGTGGTTAC TGGGGGGGCA AGAGGCCCCC GTCGGTGACC CCGCTGTACC    1260
GTGGCAACGG TACTGTGGTG TGTGACTGTG ATTTTGGAAA AATGCATTGG GCCCCCCCCT    1320
TGTGTTCCGG YCTGGTGTGG CGGGACGGTC ATAGGAGGGG CACCGTGCGC GACCTCCCCC    1380
CGGTTTGCCC CCGGGAGGTT CTCGGCACGG TGACAGTCAT GTGTCAGTGG GGTTCTGCCT    1440
ACTGGATTTG GAGATTTGGG GACTGGGTTG CATTGTACGA CGAGCTACCA CGATCAGCTC    1500
TCTGTACTTT CTTCTCAGGT CATGGTCCAC AACCTAAAGA TCTCTCAGTC TTGAATCCAT    1560
CCGGGGCACC TTGTGCTTCT TGCGTCGTTG ACCAGAGGCC GCTGAAATGT GGTTCCTGCG    1620
TCCGCGACTG CTGGGAGACG GGGGGTCCTG GGTTCGATGA GTGCGGTGTC GGTACTCGGA    1680
TGACGAAGCA CCTCGAGGCC GTCCTGGTTG ATGGAGGTGT GGAGTCCAAG GTGACAACGC    1740
CCAAGGGTGA GCGCCCCAAA TACATAGGTC AGCACGGTGT GGGAACCTAC TACGGCGCTG    1800
TCCGTAGCCT CAACATCAGT TACCTAGTGA CTGAGGTGGG GGGCTATTGG CATGCGCTGA    1860
AGTGCCCGTG CGACTTTGTG CCCCGAGTGC TCCCAGAAAG AATTCCAGGT AGGCCTGTGA    1920
ATGCATGTCT AGCTGGGAAG TCTCCGCACC CGTTCGCAAG TTGGGCTCCC GGTGGGTTTT    1980
ACGCCCCCGT GTTCACCAAG TGCAACTGGC CGAAGACCTC CGGAGTGGAT GTGTGTCCTG    2040
GGTTTGCTTT CGATTTCCCT GGTGATCACA ACGGCTTCAT CCATGTTAAA GGCAACAGAC    2100
AGCAGGTTTA CAGTGGTCAG CGAAGGTCTT CGCCGGCTTG GTTGCTTACT GACATGGTCC    2160
TGGCCCTGTT GGTGGTGATG AAGTTGGCTG AGGCTAGAGT TGTCCCCCTG TTTATGCTGG    2220
CAATGTGGTG GTGGTTGAAT GGAGCATCTG CTGCCACTAT TGTCATCATA CACCCTACTG    2280
TCACGAAGTC CACTGAAAGT GTTCCATTGT GGACTCCGCC CACTGTTCCA ACTCCATCTT    2340
GCCCGAATTC TACCACCGGA GTCGCGGACT CTACCTACAA TGCTGGTTGC TACATGGTGG    2400
CAGGCCTGGC GGCCGGGGCT CAGGCGGTCT GGGGTGCTGC CAATGATGGT GCTCAGGCCG    2460
TCGTTGGTGG CATCTGGCCC GCGTGGCTCA AGCTGCGAAG CTTCGCTGCC GGTCTGGCCT    2520
GGTTGTCAAA TGTTGGGGCT TACTTGCCGG TCGTCGAGGC CGCVCTGGCT CCCGAGCTGG    2580
TGTGCACCCC GGTGGTCGGC TGGGCAGCCC AGGAGTGGTG GTTCACTGGT TGTCTGGGTG    2640
TGATGTGTGT CGTGGCGTAC CTGAATGTCC TGGGCTCTGT RAGGGCTGCC GTGCTTGTGG    2700
```

```
CGATGCACTT CGCAAGGGGT GCTCTGCCGC TGGTATTGGT GGTAGCTGCC GGGGTRACCC   2760
GGGAGCGGCA CAGCGTCTTA GGGCTTGAGG TGTGCTTCGA TCTGGATGGT GGAGACTGGC   2820
CRGACGCCAG TTGGTCTTGG GGTTTAGCAG GCGTGGTGAG CTGGGCCCTC CTGGTGGGGG   2880
GTCTGATGAC CCACGGTGGC CGATCAGCCA GAYTGACTTG GTAYGCCAGG TGGGCCGTCA   2940
ATTAYCAGAG GGTTCGYCGG TGGGTGAACA ACTCACCGGT TGGAGCYTTT GGYCGTTGGM   3000
GGCGYGCCTG GAAAGCYTGG TTRGTKGTGG CTTGGTTCTT CCCCCAGACA GTTGCCACAG   3060
TYTCCGTCAT CTTCATACTC TGTTTGAGCA GTTTAGATGT CATTGATTTC ATCTTGGARG   3120
TACTCTTGGT TAACTCACCA AATCTCGCGC GCTTGGCGCG RGTGCTGGAC TCCTTAGCTC   3180
THGCTGAGGA GCGGCTGGCC TGCTCTTGGC TGGTGGGCGT CCTGCGCAAG CGGGGCGTCC   3240
TCCTCTACGA GCACGCYGGT CACACTAGCA GGCGCGGTGC TGCCCGCTTG CGAGAGTGGG   3300
GYTTTGCGCT YGAGCCKGTT AGYATAACCA AGGAAGATTG YGCYATTGTT CGGGACTCTG   3360
CTCGTGTGTT GGGCTGTGGA CAATTGGTCC ATGGGAAACC AGTGGTCGCG AGGCGAGGCG   3420
ACGAGGTGTT GATCGGCTGT GTGAACAGTC GGTTCGACCT TCCGCCTGGC TTTGTTCCCA   3480
CTGCTCCCGT GGTSCTTCAT CARGCWGGCA ARGGRTTYTT YGGGGTTGTG AAGACMTCCA   3540
TGACAGGCAA GGACCCGTCC GAACACCACG GRAACGTGGT GGTCCTWGGG ACTTCAACAA   3600
CKCGTTCCAT GGGCTGCTGC GTGAACGGAG TAGTGTACAC RACATACCAT GGYACCAACG   3660
CCCGRCCKAT GGCGGGGCCK TTTGGKCCYG TCAAYGCTCG GTGGTGGTCW GCGAGYGACG   3720
ACGTCACGGT YTACCCGCTC CCWAATGGYG CTTCTTGCCT YCARGCWTGY AAGTGCCAAC   3780
CAACTGGGGT GTGGGTGATC CGGAATGACG GAGCTCTTTG CCATGGAACT CTCGGCAAGG   3840
TGGTGGATTT AGATATGCCC GCTGAGTTGT CAGACTTTCG CGGGTCTTCT GGATCACCAA   3900
TCTTGTGCGA TGAGGGTCAT GCTGTTGGCA TGCTGATTTC GGTGCTTCAT AGGGGGAGTA   3960
GGGTTTCCTC GGTGCGGTAT ACCAAACCTT GGGAAACTCT CCCTCGGGAG ATTGAGGCTC   4020
GATCGGAGGC CCCCCCTGTG CCAGGAACCA CTGGATACAG GGAGGCGCCA CTGTTCCTGC   4080
CCACCGGAGC TGGCAAGTCG ACGCGCGTGC CGAATGAGTA CGTCAAGGCT GGACACAARG   4140
TGCTTGTACT AAACCCATCC ATTGCCACAG TGAGGGCCAT GGGCCCTTAC ATGGAAAAGT   4200
TAACCGGCAA ACATCCGTCG GTGTACTGTG GCCATGACAC TACTGCATAT TCCAGGACTA   4260
CTGACTCATC TTTGACCTAC TGTACATACG GCAGGTTTAT GGCCAATCCC AGGAAATACT   4320
TGCGGGGGAA CGACGTCGTA ATTTGCGACG AGTTGCACGT CACCGACCCG ACCTCAATTT   4380
TGGGGATGGG TCGGGCGAGG TTACTCGCTC GCGAGTGCGG CGTACGCCTC CTGCTTTTCG   4440
```

```
CTACGGCGAC CCCACCGGTC TCTCCGATGG CGAAGCATGA ATCTATTCAT GAGGAGATGT    4500
TGGGCAGTGA GGGGGAGGTC CCCTTCTATT GCCAATTCCT CCCACTGAGT AGGTATGCTA    4560
CTGGGAGACA CCTGCTGTTT TGTCATTCCA AGGTAGARTG CACTAGGTTA TCCTCAGCTT    4620
TGGCCAGCTT TGGTGTCAAC ACCGTTGTGT ACTTCAGAGG CAAAGAAACT GACATTCCAA    4680
CTGGTGACGT GTGCGTTTGC GCCACAGACG CACTTTCCAC TGGTTACACT GGCAATTTTG    4740
ACACCGTAAC AGACTGTGGT TTAATGGTTG AGGAGGTAGT GGAAGTGACC CTGGACCCGA    4800
CCATCACTAT CGGTGTGAAG ACCGTCCCGG CCCCTGCCGA ACTGAGGGCT CAGAGGCGTG    4860
GTAGGTGTGG CCGTGGGAAA GCGGGCACTT ACTATCAGGC ATTGATGTCT TCGGCGCCGG    4920
CGGGAACSGT TCGGTCTGGG GCTCTCTGGG CAGCTGTTGA GGCTGGHGTC TCGTGGTATG    4980
GCCTAGAGCC CGATGCTATT GGAGACCTGC TTAGGGCCTA CGACTCGTGT CCTTATACTG    5040
CTGCCATCAG TGCGTCCATC GGAGAGGCCA TTGCCTTTTT TACTGGYCTA GTGCCAATGA    5100
GGAATTATCC TCAGGTGGTT TGGGCCAAGC AGAAGGGRCA CAACTGGCCA CTCTTGGTGG    5160
GTGTGCAGAG GCACATGTGT GAGGACGCGG GCTGTGGTCC KCCCGCTAAT GGTCCCGAAT    5220
GGAGCGGCAT CAGGGGAAAA GGGCCTGTTC CCCTGTTGTG CCGATGGGGT GGTGACTTGC    5280
CTGAGTCGGT GGCTCCGCAT CACTGGGTTG ATGACCTACA GGCCCGGCTC GGTGTGGCCG    5340
AGGGTTACAC TCCCTGCATT GCTGGACCGG TGCTTTTGGT CGGTTTGGCG ATGGCGGGGG    5400
GGGCTATCCT GGCACACTGG ACGGGGTCTC TGGTTGTAGT GACCAGTTGG GTTGTCAATG    5460
GGAACGGTAA CCCGCTGATA CAAAGCGCCT CTAGGGGCGT GGCKACYAGC GGTCCATACC    5520
CAGTACCCCC AGATGGTGGT GAACGGTACC CATCAGACAT CAAGCCAATY ACTGAGGCTG    5580
TGACCACCCT TGAGACTGCG TGCGGYTGGG GCCCAGCCGC GGCBAGTCTG GCTTATGTGA    5640
AGGCCTGTGA AACTGGAACC ATGTTGGCTG ACAARGCGAG TGCTGCGTGG CAGGCTTGGG    5700
CTGCAAACAA CTTTGTGCCT CCACCAGCAT CACACTCAAC TTCCTTGTTR CAGAGCTTGG    5760
AYGCTGCGTT CACTTCAGCT TGGGATAGCG TGTTCACTCA CGGCCGTTCC TTGCTTGTTG    5820
GGTTCACAGC TGCTTACGGC GCTCGGCGGA ACCCACCGCT GGGCGTCGGA GCCTCTTTCT    5880
TGCTGGGCAT GTCATCGAGC CACYTRACTC ACGTCAGACT TGCTGCTGCG TTGCTCCTCG    5940
GCGTCGGGGG TACCGTCCTA GGCACGCCTG CTACTGGGCT TGCTATGGCG GGTGCCTACT    6000
TCGCKGGGGG CAGCGTTACC GCTAACTGGC TGAGTATCAT TGTGGCTCTA ATCGGAGGCT    6060
GGGAGGGGGC RGTKAACGCA GCCTCACTCA CCTTCGAYCT CCTGGCKGGG AAGTTACAAG    6120
```

```
CKAGYGAYGC TTGGTGCCTR GTCAGYTGCY TGGCCTCTCC GGGGGCTTCG GTGGCYGGTG    6180
TGGCDCTVGG YCTDYTGCTV TGGTCTGTCA ARAAGGGTGT GGGWCARGAY TGGGTTAACA    6240
GAYTGTTGAC GATGATGCCA CGCAGTTCGG TGATGCCTGA CGATTTCTTC CTCAAAGATG    6300
AGTTCGTCAC CAAGGTGTCT ACTGTCCTGC GAAAGTTGTC ATTGTCAAGA TGGATCATGA    6360
CTCTTGTGGA CAAGCGGGAG ATGGAGATGG AGACMCCCGC TTCTCAGATT GTTTGGGACT    6420
TGCTTGACTG GTGCATCCGG CTRGGTCGGT TCCTGTACAA TAAACTYATG TTTGCTCTCC    6480
CTAGGTTGCG CCTGCCGCTT ATCGGTTGCA GTACCGGTTG GGGTGGCCCG TGGGAGGGCA    6540
ATGGTCATTT GGAAACAAGG TGTACTTGTG GCTGTGTGAT TACCGGTGAT ATTCACGATG    6600
GTATATTGCA CGACCTACAT TATACCTCCC TACTGTGCAG ACATTACTAC AAGAGGACAG    6660
TGCCTGTTGG CGTCATGGGC AATGCTGAGG GAGCAGTCCC CCTTGTGCCT ACTGGCGGTG    6720
GAATCAGGAC TTACCAAATT GGGACTTCTG ACTGGTTTGA GGCTGTGGTC GTGCATGGGA    6780
CAATCACGGT GCACGCCACC AGTTGCTATG AGTTGAAAGC TGCTGACGTT CGGAGGGCGG    6840
TGCGAGCCGG CCCGACTTAC GTTGGTGGCG TACCTTGCAG CTGGAGCGCG CCGTGTACTG    6900
CGCCTGCGCT CGTTTACAGG CTAGGCCAGG GCATCAAAAT CGATGGAGCG CGCCGACTGT    6960
TGCCCTGTGA CTTAGCACAG GGAGCGCGCC ACCCCCCGGT ATCTGGCAGT GTTGCCGGTA    7020
GTGGTTGGAC AGATGAGGAC GAGAGGGACT TGGTGGAAAC CAAGGCTGCC GCCATCGAGG    7080
CCATTGGGGC GGCCTTGCAC CTCCCTTCAC CGGAGGCTGC TCAGGCCGCT CTAGAGGCTT    7140
TGGAGGAGGC TGCCGTGTCC CTGTTGCCCC ATGTGCCCGT CATTATGGGT GATGACTGTT    7200
CATGCCGGGA TGAGGCGTTC CAAGGCCACT TCATCCCAGA ACCCAATGTG ACAGAGGTAC    7260
CCATTGAGCC CACGGTCGGA GACGTGGAGG CACTCAAGCT GCGGGCTGCA GACCTGACCG    7320
CCAGGTTGCA AGACTTGGAG GCCATGGCTC TCGCCCGCGC TGAGTCAATC GAGGATGCTC    7380
GCGCAGCTTC GATGCCTTCG CTCACCGAGG TGGACTCAAT GCCATCATTG GAGTCGAGCC    7440
CTTGCTCCTC CTTTGAACAA ATCTCTTTAA CTGAAAGTGA CCCTGAGACT GTCGTCGAGG    7500
CTGGCTTACC CTTGGAGTTC GTGAACTCCA ACACCGGGCC GTCTCCGGCT CGGAGGATTG    7560
TCAGAATCCG ACAGGCTTGC TGTTGTGACA GATCCACAAT GAAGGCCATG CCGTTGTCGT    7620
TCACTGTCGG GGAGTGCCTC TTCGTTACTC GCTATGACCC GGACGGTCAC CAACTGTTTG    7680
ACGAGCGAGG TCCGATAGAG GTATCTACTC CTATATGTGA AGTGATTGGG GACATCAGGC    7740
TTCAGTGTGA CCAAATTGAG GAAACTCCAA CATCTTACTC TTACATCTGG TCAGGGGCGC    7800
CCTTGGGTAC TGGGAGAAGT GTCCCCCAAC CCATGACGCG CCCTATAGGG ACCCATCTGA    7860
```

```
CTTGTGACAC TACCAAAGTT TATGTTACTG ACCCTGATCG GGCCGCTGAG CGGGCCGAGA      7920
AGGTTACAAT CTGGAGGGGT GATAGGAAGT ATGACAAGCA TTATGAGGCT GTCGTTGAGG      7980
CTGTCCTGAA AAAGGCAGCC GCGACGAAGT CTCATGGCTG GACCTATTCC CAGGCTATAG      8040
CTAAAGTTAG GCGCCGAGCA GCCGCTGGAT ACGGCAGCAA GGTGACCGCC TCCACATTGG      8100
CCACTGGTTG GCCTCACGTG GAGGAGATGC TGGACAAAAT AGCCAGGGGA CAGGAAGTTC      8160
CTTTCACTTT TGTGACCAAG CGAGAGGTTT TCTTCTCCAA AACTACCCGT AAGCCCCCAA      8220
GATTCATAGT TTTCCCACCT TTGGACTTCA GGATAGCTGA AAAGATGATT CTGGGTGACC      8280
CCGGCATCGT TGCAAAGTCA ATTCTGGGTG ACGCTTATCT GTTCCAGTAC ACGCCCAATC      8340
AGAGGGTCAA AGCTCTGGTT AAGGCGTGGG AGGGGAAGTT GCATCCCGCT GCGATCACTG      8400
TGGACGCCAC TTGTTTCGAC TCATCGATTG ATGAGCACGA CATGCAGGTG GAGGCTTCGG      8460
TGTTTGCGGC GGCTAGTGAC AACCCCTCAA TGGTACATGC TTTGTGCAAG TACTACTCTG      8520
GTGGCCCTAT GGTTTCCCCA GATGGGGTTC CCTTGGGGTA CCGCCAGTGT AGGTCGTCGG      8580
GCGTGTTAAC AACTAGCTCG GCGAACAGCA TCACTTGTTA CATTAAGGTC AGCGCGGCCT      8640
GCAGGCGGGT GGGGATTAAG GCACCATCAT TCTTTATAGC TGGAGATGAT TGCTTGATCA      8700
TCTATGAAAA TGATGGAACT GATCCCTGCC CTGCTCTTAA GGCTGCCCTG GCCAACTATG      8760
GATACAGGTG TGAACCAACA AAGCATGCTT CACTGGACAC AGCTGAGTGT TGCTCGGCCT      8820
ACTTGGCTGA GTGCGTAGCT GGGGGTGCCA AGCGCTGGTG GTTGAGCACG GACATGAGGA      8880
AGCCGCTCGC AAGGGCGTCT TCCGAATATT CGGACCCAAT CGGCAGTGCT TTAGGGACCA      8940
TCTTGATGTA TCCCCGGCAT CCAATCGTGC GGTATGTTCT AATACCACAC GTACTAATAA      9000
TGGCTTACAG GAGTGGCAGC ACACCGGATG AGTTGGTTAT GTGTCAGGTT CAGGGAAATC      9060
ATTACTCTTT CCCGCTGCGG CTGCTGCCTC GCGTCTTGGT CTCTCTACAT GGTCCGTGGT      9120
GCCTACAAGT CACCACGGAC AGTACGAAGA CTAGGATGGA GGCAGGCTCA GCSTTGCGGG      9180
ATTTAGGAAT GAAATCCCTA GCCTGGCACC GCCGACGTGC CGGAAATGTG CGCACTCGCC      9240
TCCTGAGGGG AGGCAAGGAG TGGGGGCACC TGGCCAGAGC CCTCCTCTGG CAYCCAGGKT      9300
TGAAGGAGCA YCCCCCRCCC ATAAATTCAC TTCCAGGTTT TCAGCTGGCG ACGCCTTACG      9360
AACACCATGA AGAGGTCTTG ATCTCGATCA AGAGTCGACC ACCTTGGATA AGGTGGATTC      9420
TTGGTGCTTG TCTCTCGTTG CTGGCCGCCT TGCTGTGAAT TCGCTCCAGG CAGTAGGACC      9480
TTCGGGTCGG GGG                                                         9493
```

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 31 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

CCATAATCAT GAGCCGCGAG TTGAAGAGCA C 31

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

GCCAAGCCAT GGTGAATGTG 20

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 38 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

GTATTGCGCC ATGGCTCGAC AAGCGTGGGT GGCCGGGG 38

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

GGACTGCCAT GGTGGTATAG AAAAGAG 27

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 36 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

TATAATAAGC TTCTCGACAA GCGTGGGTGG CCGGGG 36

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 34 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

GTATTGCGCC ATGGCACTGG GTGCAAGCCC AGAA 34

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 46 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

```
Met Ser Val Val Asp Thr Phe Thr Met Ala Trp Leu Trp Leu Leu Val
                5                   10                  15

Cys Phe Pro Leu Ala Gly Gly Val Leu Phe Asn Ser Arg His Gln Cys
            20                  25                  30

Phe Asn Gly Asp His Tyr Val Leu Ser Asn Cys Cys Ser Arg

                35                  40                  45
```

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 67 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

```
Met Gly Pro Pro Ser Ser Ala Ala Ala Cys Ser Arg Gly Ser Pro Arg
                5                   10                  15

Ile Leu Arg Val Arg Ala Gly Gly Ile Ser Leu Phe Tyr Thr Ile Met
            20                  25                  30

Ala Val Leu Leu Leu Leu Leu Val Val Glu Ala Gly Ala Ile Leu Ala
        35                  40                  45

Pro Ala Thr His Ala Cys Arg Ala Asn Gly Gln Tyr Phe Leu Thr Asn
    50                  55                  60

Cys Cys Ala
65
```

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 9143 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

```
ACCACAAACA CTCCAGTTTG TTACACTCCG CTAGGAATGC TCCTGGAGCA CCCCCCCTAG     60
CAGGGCGTGG GGGATTTCCC CTGCCCGTCT GCAGAAGGGT GGAGCCAACC ACCTTAGTAT    120
GTAGGCGGCG GGACTCATGA CGCTCGCGTG ATGACAAGCG CCAAGCTTGA CTTGGATGGC    180
CCTGATGGGC GTTCATGGGT TCGGTGGTGG TGGCGCTTTA GGCAGCCTCC ACGCCCACCA    240
CCTCCCAGAT AGAGCGGCGG CACTGTAGGG AAGACCGGGG ACCGGTCACT ACCAAGGACG    300
CAGACCTCTT TTTGAGTATC ACGCCTCCGG AAGTAGTTGG GCAAGCCCAC CTATATGTGT    360
TGGGATGGTT GGGGTTAGCC ATCCATACCG TACTGCCTGA TAGGGTCCTT GCGAGGGGAT    420
CTGGGAGTCT CGTAGACCGT AGCACATGCC TGTTATTTCT ACTCAAACAA GTCCTGTACC    480
TGCGCCCAGA ACGCGCAAGA ACAAGCAGAC GCAGGCTTCA TATCCTGTGT CCATTAAAAC    540
ATCTGTTGAA AGGGGACAAC GAGCAAAGCG CAAAGTCCAG CGCGATGCTC GGCCTCGTAA    600
TTACAAAATT GCTGGTATCC ATGATGGCTT GCAGACATTG GCTCAGGCTG CTTTGCCAGC    660
TCATGGTTGG GGACGCCAAG ACCCTCGCCA TAAGTCTCGC AATCTTGGAA TCCTTCTGGA    720
TTACCCTTTG GGGTGGATTG GTGATGTTAC AACTCACACA CCTCTAGTAG GCCCGCTGGT    780
GGCAGGAGCG GTCGTTCGAC CAGTCTGCCA GATAGTACGC TTGCTGGAGG ATGGAGTCAA    840
CTGGGCTACT GGTTGGTTCG GTGTCCACCT TTTTGTGGTA TGTCTGCTAT CTTTGGCCTG    900
TCCCTGTAGT GGGGCGCGGG TCACTGACCC AGACACAAAT ACCACAATCC TGACCAATTG    960
CTGCCAGCGT AATCAGGTTA TCTATTGTTC TCCTTCCACT TGCCTACACG AGCCTGGTTG   1020
TGTGATCTGC GCGGACGAGT GCTGGGTTCC CGCCAATCCG TACATCTCAC ACCCTTCCAA   1080
TTGGACTGGC ACGGACTCCT TCTTGGCTGA CCACATTGAT TTTGTTATGG GCGCTCTTGT   1140
GACCTGTGAC GCCCTTGACA TTGGTGAGTT GTGTGGTGCG TGTGTATTAG TCGGTGACTG   1200
GCTTGTCAGG CACTGGCTTA TTCACATAGA CCTCAATGAA ACTGGTACTT GTTACCTGGA   1260
AGTGCCCACT GGAATAGATC CTGGGTTCCT AGGGTTTATC GGGTGGATGG CCGGCAAGGT   1320
CGAGGCTGTC ATCTTCTTGA CCAAACTGGC TTCACAAGTA CCATACGCTA TTGCGACTAT   1380
GTTTAGCAGT GTACACTACC TGGCGGTTGG CGCTCTGATC TACTATGCCT CTCGGGGCAA   1440
GTGGTATCAG TTGCTCCTAG CGCTTATGCT TTACATAGAA GCGACCTCTG GAAACCCTAT   1500
CAGGGTGCCC ACTGGATGCT CAATAGCTGA GTTTTGCTCG CCTTTGATGA TACCATGTCC   1560
TTGCCACTCT TATTTGAGTG AGAATGTGTC AGAAGTCATT TGTTACAGTC CAAAGTGGAC   1620
```

CAGGCCTGTC ACTCTAGAGT ATAACAACTC CATATCTTGG TACCCCTATA CAATCCCTGG 1680

TGCGAGGGGA TGTATGGTTA AATTCAAAAA TAACACATGG GGTTGCTGCC GTATTCGCAA 1740

TGTGCCATCG TACTGCACTA TGGGCACTGA TGCAGTGTGG AACGACACTC GCAACACTTA 1800

CGAAGCATGC GGTGTAACAC CATGGCTAAC AACCGCATGG CACAACGGCT CAGCCCTGAA 1860

ATTGGCTATA TTACAATACC CTGGGTCTAA AGAAATGTTT AAACCTCATA ATTGGATGTC 1920

AGGCCATTTG TATTTTGAGG GATCAGATAC CCCTATAGTT TACTTTTATG ACCCTGTGAA 1980

TTCCACTCTC CTACCACCGG AGAGGTGGGC TAGGTTGCCC GGTACCCCAC CTGTGGTACG 2040

TGGTTCTTGG TTACAGGTTC CGCAAGGGTT TTACAGTGAT GTGAAAGACC TAGCCACAGG 2100

ATTGATCACC AAAGACAAAG CCTGGAAAAA TTATCAGGTC TTATATTCCG CCACGGGTGC 2160

TTTGTCTCTT ACGGGAGTTA CCACCAAGGC CGTGGTGCTA ATTCTGTTGG GGTTGTGTGG 2220

CAGCAAGTAT CTTATTTTAG CCTACCTCTG TTACTTGTCC CTTTGTTTTG GGCGCGCTTC 2280

TGGTTACCCT TTGCGTCCTG TGCTCCCATC CCAGTCGTAT CTCCAAGCTG GCTGGGATGT 2340

TTTGTCTAAA GCTCAAGTAG CTCCTTTTGC TTTGATTTTC TTCATCTGTT GCTATCTCCG 2400

CTGCAGGCTA CGTTATGCTG CCCTTTTAGG GTTTGTGCCC ATGGCTGCGG GCTTGCCCCT 2460

AACTTTCTTT GTTGCAGCAG CTGCTGCCCA ACCAGATTAT GACTGGTGGG TGCGACTGCT 2520

AGTGGCAGGG TTAGTTTTGT GGGCCGGCCG TGACCGTGGT CCACGTATAG CTCTGCTTGT 2580

AGGTCCTTGG CCTCTGGTAG CGCTTTTAAC CCTCTTGCAT TTGGCTACGC CTGCTTCAGC 2640

TTTTGACACC GAGATAATTG GAGGGCTGAC AATACCACCT GTAGTAGCAT TAGTTGTCAT 2700

GTCTCGTTTT GGCTTCTTTG CTCACTTGTT ACCTCGCTGT GCTTTAGTTA ACTCCTATCT 2760

TTGGCAACGT TGGGAGAATT GGTTTTGGAA CGTTACACTA AGACCGGAGA GGTTTCTCCT 2820

TGTGCTGGTT TGTTTCCCCG GTGCGACATA TGACACGCTG GTGACTTTCT GTGTGTGTCA 2880

CGTAGCTCTT CTATGTTTAA CATCCAGTGC AGCATCGTTC TTTGGGACTG ACTCTAGGGT 2940

TAGGGCCCAT AGAATGTTGG TGCGTCTCGG AAAGTGTCAT GCTTGGTATT CTCATTATGT 3000

TCTTAAGTTT TTCCTCTTAG TGTTTGGTGA GAATGGTGTG TTTTTCTATA AGCACTTGCA 3060

TGGTGATGTC TTGCCTAATG ATTTTGCCTC GAAACTACCA TTGCAAGAGC CATTTTTCCC 3120

TTTTGAAGGC AAGGCAAGGG TCTATAGGAA TGAAGGAAGA CGCTTGGCGT GTGGGGACAC 3180

GGTTGATGGT TTGCCCGTTG TTGCGCGTCT CGGCGACCTT GTTTTCGCAG GGTTAGCTAT 3240

GCCGCCAGAT GGGTGGGCCA TTACCGCACC TTTTACGCTG CAGTGTCTCT CTGAACGTGG 3300

CACGCTGTCA GCGATGGCAG TGGTCATGAC TGGTATAGAC CCCCGAACTT GGACTGGAAC 3360

```
TATCTTCAGA TTAGGATCTC TGGCCACTAG CTACATGGGA TTTGTTTGTG ACAACGTGTT    3420
GTATACTGCT CACCATGGCA GCAAGGGGCG CCGGTTGGCT CATCCCACAG GCTCCATACA    3480
CCCAATAACC GTTGACGCGG CTAATGACCA GGACATCTAT CAACCACCAT GTGGAGCTGG    3540
GTCCCTTACT CGGTGCTCTT GCGGGGAGAC CAAGGGGTAT CTGGTAACAC GACTGGGGTC    3600
ATTGGTTGAG GTCAACAAAT CCGATGACCC TTATTGGTGT GTGTGCGGGG CCCTTCCCAT    3660
GGCTGTTGCC AAGGGTTCTT CAGGTGCCCC GATTCTGTGC TCCTCCGGGC ATGTTATTGG    3720
GATGTTCACC GCTGCTAGAA ATTCTGGCGG TTCAGTCAGC CAGATTAGGG TTAGGCCGTT    3780
GGTGTGTGCT GGATACCATC CCCAGTACAC AGCACATGCC ACTCTTGATA CAAAACCTAC    3840
TGTGCCTAAC GAGTATTCAG TGCAAATTTT AATTGCCCCC ACTGGCAGCG GCAAGTCAAC    3900
CAAATTACCA CTTTCTTACA TGCAGGAGAA GTATGAGGTC TTGGTCCTAA ATCCCAGTGT    3960
GGCTACAACA GCATCAATGC CAAAGTACAT GCACGCGACG TACGGCGTGA ATCCAAATTG    4020
CTATTTTAAT GGCAAATGTA CCAACACAGG GGCTTCACTT ACGTACAGCA CATATGGCAT    4080
GTACCTGACC GGAGCATGTT CCCGGAACTA TGACGTCATC ATTTGTGACG AATGCCATGC    4140
TACCGATGCA ACCACCGTGT TGGGCATTGG AAAGGTTCTA ACCGAAGCTC CATCCAAAAA    4200
TGTTAGGCTA GTGGTTCTTG CCACGGCTAC CCCCCCTGGA GTAATCCCTA CACCACATGC    4260
CAACATAACT GAGATTCAAT TAACCGATGA AGGCACTATC CCCTTTCATG GAAAAAAGAT    4320
TAAGGAGGAA AATCTGAAGA AAGGGAGACA CCTTATCTTT GAGGCTACCA AAAAACACTG    4380
TGATGAGCTT GCTAACGAGT TAGCTCGAAA GGGAATAACA GCTGTCTCTT ACTATAGGGG    4440
ATGTGACATC TCAAAAATCC CTGAGGGCGA CTGTGTAGTA GTTGCCACTG ATGCCTTGTG    4500
TACAGGGTAC ACTGGTGACT TTGATTCCGT GTATGACTGC AGCCTCATGG TAGAAGGCAC    4560
ATGCCATGTT GACCTTGACC CTACTTTCAC CATGGGTGTT CGTGTGTGCG GGGTCTCAGC    4620
AATAGTTAAA GGCCAGCGTA GGGGCCGCAC AGGCCGTGGG AGAGCTGGCA TATACTACTA    4680
TGTAGACGGG AGTTGTACCC CTTCGGGTAT GGTTCCTGAA TGCAACATTG TTGAAGCCTT    4740
CGACGCAGCC AAGGCATGGT ATGGTTTGTC ATCAACAGAA GCTCAAACTA TTCTGGACAC    4800
CTATCGCACC CAACCTGGGT TACCTGCGAT AGGAGCAAAT TTGGACGAGT GGGCTGATCT    4860
CTTTTCTATG GTCAACCCCG AACCTTCATT TGTCAATACT GCAAAAAGAA CTGCTGACAA    4920
TTATGTTTTG TTGACTGCAG CCCAACTACA ACTGTGTCAT CAGTATGGCT ATGCTGCTCC    4980
CAATGACGCA CCACGGTGGC AGGGAGCCCG GCTTGGGAAA AAACCTTGTG GGGTTCTGTG    5040
```

```
GCGCTTGGAC GGCGCTGACG CCTGTCCTGG CCCAGAGCCC AGCGAGGTGA CCAGATACCA   5100
AATGTGCTTC ACTGAAGTCA ATACTTCTGG GACAGCCGCA CTCGCTGTTG GCGTTGGAGT   5160
GGCTATGGCT TATCTAGCCA TTGACACTTT TGGCGCCACT TGTGTGCGGC GTTGCTGGTC   5220
TATTACATCA GTCCCTACCG GTGCTACTGT CGCCCCAGTG GTTGACGAAG AAGAAATCGT   5280
GGAGGAGTGT GCATCATTCA TTCCCTTGGA GGCCATGGTT GCTGCAATCG ATAAGCTGAA   5340
GAGTACAATA ACCACAACTA GTCCTTTCAC ATTGGAAACC GCCCTTGAAA AACTTAACAC   5400
CTTTCTTGGG CCTCATGCAG CTACAATCCT TGCTATCATA GAGTATTGCT GTGGCTTAGT   5460
CACTTTACCT GACAATCCCT TTGCATCATG CGTGTTTGCT TTCATTGCGG GTATTACTAC   5520
CCCACTACCT CACAAGATCA AAATGTTCCT GTCATTATTT GGAGGCGCAA TTGCGTCCAA   5580
GCTTACAGAC GCTAGAGGCG CACTGGCGTT CATGATGGCC GGGGCTGCGG GAACAGCTCT   5640
TGGTACATGG ACATCGGTGG GTTTTGTCTT TGACATGCTA GGCGGCTATG CTGCCGCCTC   5700
ATCCACTGCT TGCTTGACAT TTAAATGCTT GATGGGTGAG TGGCCCACTA TGGATCAGCT   5760
TGCTGGTTTA GTCTACTCCG CGTTCAATCC GGCCGCAGGA GTTGTGGGCG TCTTGTCAGC   5820
TTGTGCAATG TTTGCTTTGA CAACAGCAGG GCCAGATCAC TGGCCCAACA GACTTCTTAC   5880
TATGCTTGCT AGGAGCAACA CTGTATGTAA TGAGTACTTT ATTGCCACTC GTGACATCCG   5940
CAGGAAGATA CTGGGCATTC TGGAGGCATC TACCCCCTGG AGTGTCATAT CAGCTTGCAT   6000
CCGTTGGCTC CACACCCCGA CGGAGGATGA TTGCGGCCTC ATTGCTTGGG GTCTAGAGAT   6060
TTGGCAGTAT GTGTGCAATT TCTTTGTGAT TTGCTTTAAT GTCCTTAAAG CTGGAGTTCA   6120
GAGCATGGTT AACATTCCTG GTTGTCCTTT CTACAGCTGC CAGAAGGGGT ACAAGGGCCC   6180
CTGGATTGGA TCAGGTATGC TCCAAGCACG CTGTCCATGC GGTGCTGAAC TCATCTTTTC   6240
TGTTGAGAAT GGTTTTGCAA AACTTTACAA AGGACCCAGA ACTTGTTCAA ATTACTGGAG   6300
AGGGGCTGTT CCAGTCAACG CTAGGCTGTG TGGGTCGGCT AGACCGGACC CAACTGATTG   6360
GACTAGTCTT GTCGTCAATT ATGGCGTTAG GGACTACTGT AAATATGAGA AATTGGGAGA   6420
TCACATTTTT GTTACAGCAG TATCCTCTCC AAATGTCTGT TTCACCCAGG TGCCCCCAAC   6480
CTTGAGAGCT GCAGTGGCCG TGGACGGCGT ACAGGTTCAG TGTTATCTAG GTGAGCCCAA   6540
AACTCCTTGG ACGACATCTG CTTGCTGTTA CGGTCCGGAC GGTAAGGGTA AAACTGTTAA   6600
GCTTCCCTTC CGCGTTGACG GTCACACACC TGGTGTGCGC ATGCAACTTA ATTTGCGTGA   6660
TGCACTTGAG ACAAATGACT GTAATTCCAT AAACAACACT CCTAGTGATG AAGCCGCAGT   6720
GTCCGCTCTT GTTTTCAAAC AGGAGTTGCG GCGTACAAAC CAATTGCTTG AGGCAATTTC   6780
```

```
AGCTGGCGTT GACACCACCA AACTGCCAGC CCCCTCCATC GAAGAGGTAG TGGTAAGAAA   6840
GCGCCAGTTC CGGGCAAGAA CTGGTTCGCT TACCTTGCCT CCCCCTCCGA GATCCGTCCC   6900
AGGAGTGTCA TGTCCTGAAA GCCTGCAACG AAGTGACCCG TTAGAAGGTC CTTCAAACCT   6960
CCCTTCTTCA CCACCTGTTC TACAGTTGGC CATGCCGATG CCCCTGTTGG GAGCAGGTGA   7020
GTGTAACCCT TTCACTGCAA TTGGATGTGC AATGACCGAA ACAGGCGGAG GCCCTGATGA   7080
TTTACCCAGT TACCCTCCCA AAAAGGAGGT CTCTGAATGG TCAGACGGAA GTTGGTCAAC   7140
GACTACAACC GCTTCCAGCT ACGTTACTGG CCCCCCGTAC CCTAAGATAC GGGGAAAGGA   7200
TTCCACTCAG TCAGCCCCCG CCAAACGGCC TACAAAAAAG AAGTTGGGAA AGAGTGAGTT   7260
TTCGTGCAGC ATGAGCTACA CTTGGACCGA CGTGATTAGC TTCAAAACTG CTTCTAAAGT   7320
TCTGTCTGCA ACTCGGGCCA TCACTAGTGG TTTCCTCAAA CAAAGATCAT TGGTGTATGT   7380
GACTGAGCCG CGGGATGCGG AGCTTAGAAA ACAAAAAGTC ACTATTAATA GACAACCTCT   7440
GTTCCCCCCA TCATACCACA AGCAAGTGAG ATTGGCTAAG GAAAAAGCTT CAAAAGTTGT   7500
CGGTGTCATG TGGGACTATG ATGAAGTAGC AGCTCACACG CCCTCTAAGT CTGCTAAGTC   7560
CCACATCACT GGCCTTCGGG GCACTGATGT TCGTTCTGGA GCAGCCCGCA AGGCTGTTCT   7620
GGACTTGCAG AAGTGTGTCG AGGCAGGTGA GATACCGAGT CATTATCGGC AAACTGTGAT   7680
AGTTCCAAAG GAGGAGGTCT TCGTGAAGAC CCCCCAGAAA CCAACAAAGA AACCCCCAAG   7740
GCTTATCTCG TACCCCCACC TTGAAATGAG ATGTGTTGAG AAGATGTACT ACGGTCAGGT   7800
TGCTCCTGAC GTAGTTAAAG CTGTCATGGG AGATGCGTAC GGGTTTGTCG ACCCACGTAC   7860
CCGTGTCAAG CGTCTGTTGT CGATGTGGTC ACCCGATGCA GTCGGAGCCA CATGCGATAC   7920
AGTGTGTTTT GACAGTACCA TCACACCCGA GGATATCATG GTGGAGACAG ACATCTACTC   7980
AGCAGCTAAA CTCAGTGACC AACACCGAGC TGGCATTCAC ACCATTGCGA GGCAGTTATA   8040
CGCTGGAGGA CCGATGATCG CTTATGATGG CCGAGAGATC GGATATCGTA GGTGTAGGTC   8100
TTCCGGCGTC TATACTACCT CAAGTTCCAA CAGTTTGACC TGCTGGCTGA AGGTAAATGC   8160
TGCAGCCGAA CAGGCTGGCA TGAAGAACCC TCGCTTCCTT ATTTGCGGCG ATGATTGCAC   8220
CGTAATTTGG AAGAGCGCCG GAGCAGATGC AGACAAACAA GCAATGCGTG TCTTTGCTAG   8280
CTGGATGAAG GTGATGGGTG CACCACAAGA TTGTGTGCCT CAACCCAAAT ACAGTTTGGA   8340
AGAATTAACA TCATGCTCAT CAAATGTTAC CTCTGGAATT ACCAAAAGTG GCAAGCCTTA   8400
CTACTTTCTT ACAAGAGATC CTCGTATCCC CCTTGGCAGG TGCTCTGCCG AGGGTCTGGG   8460
```

```
ATACAACCCC AGTGCTGCGT GGATTGGGTA TCTAATACAT CACTACCCAT GTTTGTGGGT   8520
TAGCCGTGTG TTGGCTGTCC ATTTCATGGA GCAGATGCTC TTTGAGGACA AACTTCCCGA   8580
GACTGTGACC TTTGACTGGT ATGGGAAAAA TTATACGGTG CCTGTAGAAG ATCTGCCCAG   8640
CATCATTGCT GGTGTGCACG GTATTGAGGC TTTCTCGGTG GTGCGCTACA CCAACGCTGA   8700
GATCCTCAGA GTTTCCCAAT CACTAACAGA CATGACCATG CCCCCCCTGC GAGCCTGGCG   8760
AAAGAAAGCC AGGGCGGTCC TCGCCAGCGC CAAGAGGCGT GGCGGAGCAC ACGCAAAATT   8820
GGCTCGCTTC CTTCTCTGGC ATGCTACATC TAGACCTCTA CCAGATTTGG ATAAGACGAG   8880
CGTGGCTCGG TACACCACTT TCAATTATTG TGATGTTTAC TCCCCGGAGG GGGATGTGTT   8940
TGTTACACCA CAGAGAAGAT TGCAGAAGTT TCTTGTGAAG TATTTGGCTG TCATTGTTTT   9000
TGCCCTAGGG CTCATTGCTG TTGGACTAGC CATCAGCTGA ACCCCCAAAT TCAAAATTAA   9060
TTAACAGTTT TTTTTTTTTT TTTTTTTTTT TTTTAGGGCA GCGGCAACAG GGGAGACCCC   9120
GGGCTTAACG ACCCCGCGAT GTG                                           9143
```

WHAT IS CLAIMED IS:

1. A method for controlling the translation of HGBV nucleic acids to HGBV proteins, comprising

a. contacting a first non-naturally occurring nucleic acid sequence with HGBV nucleic acid sequence under conditions which permit hybridization of the first nucleic acid sequence and the HGBV nucleic acid sequence, and

b. altering the level of translation of the HGBV nucleic acid.

2. The method of claim 1 wherein said first nucleic acid sequence is an antisense nucleic acid sequence which is substantially complementary to a sequence of the sense strand within the 5' NTR region of the HGBV nucleic acid sequence.

3. The method of claim 1 wherein said first nucleic acid is a nucleic acid analog.

4. The method of claim 3 wherein said nucleic acid analog is selected from the group consisting of a morpholino compound, a peptide nucleic analog and a phosphorothioate nucleic acid analog.

5. A method of enhancing the translation of a nucleic acid comprising operably linking a nucleic acid with a nucleic acid having a sequence corresponding to the sequence of GBV-A, -B or -C 5' region, to form a combined nucleic acid capable of being translated.

6. A composition for enhancing the translation of a nucleic acid, which composition comprises a nucleic acid having a sequence corresponding to the squence of GBV-A, -B, or -C 5' region, for operable linkage to nucleic acid to be translated.

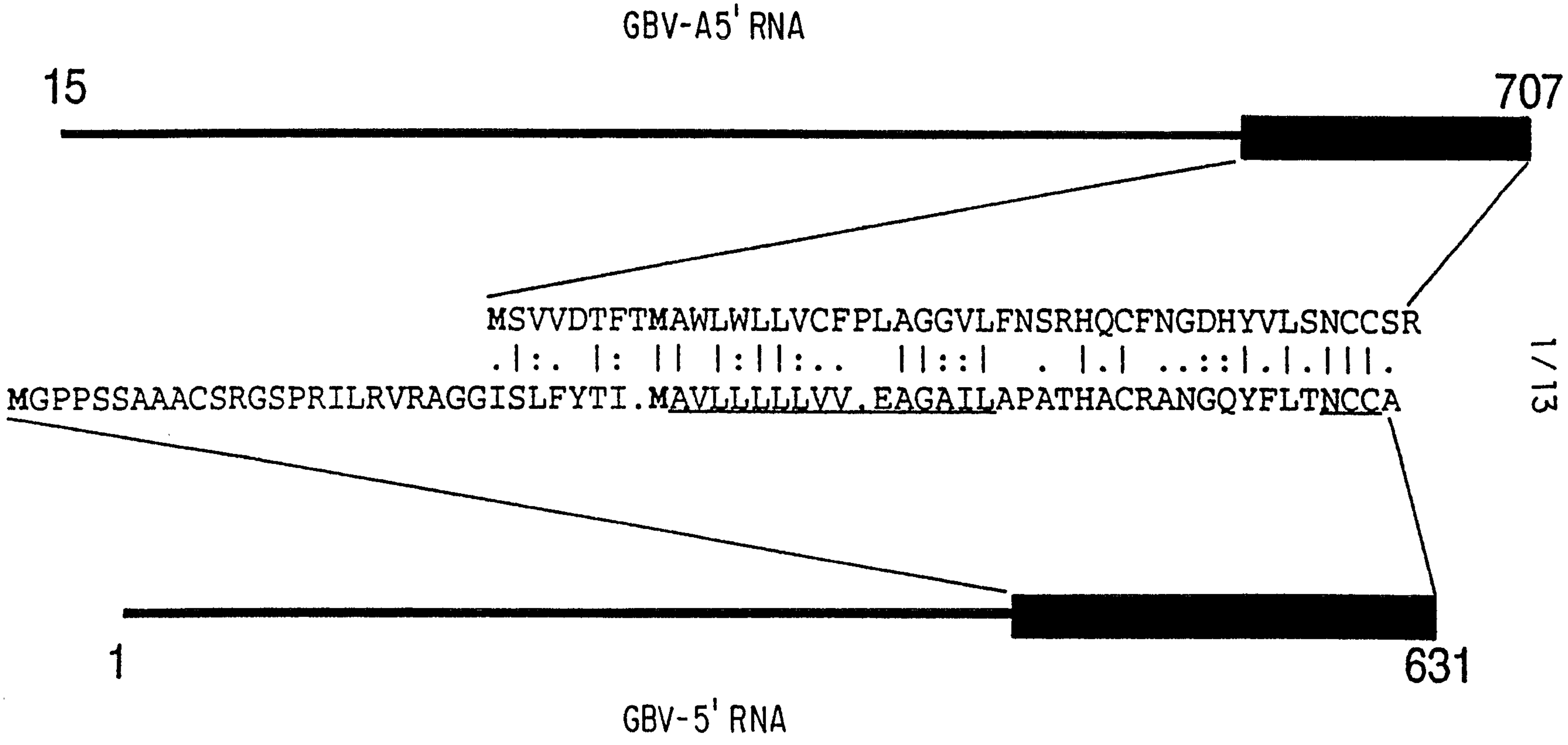
7. A composition for controlling translation of hepatitis GB virus -A, -B, or -C from GBvirus -A, -B or -C nucleic acid, which comprises a first non-naturally occurring nucleic acid having a sequence complementary to, or capable of being transcribed to form, a nucleic acid having a sequence complementary to, a sequence of the sense strand within the 5'-NTR region of HGBV-A, -B, or -C, wherein said first nucleic acid comprises a sequence selected from the 5' NTR

region of GBV-A, -B, or -C, and a cleavage area at which the full length GBV-A, -B, or -C RNA is cleaved to form a subgenomic HGBV-A, -B, or -C RNA.

8. The composition of claim 7 wherein said first nucleic acid is be a nucleic acid analog.

9. The composition of claim 8 wherein said nucleic acid analog is selected from the group consisting of morpholino compounds, peptide nucleic analogs and a phosphorothioate nucleic acid analog

10. The composition of claim 7 wherein said first nucleic acid is linked to a cholesteryl moiety at the 3' end.



```
                       MSVVDTFTMAWLWLLVCFPLAGGVLFNSRHQCFNGDHYVLSNCCSR
                       .|:. |: || |:||:..  ||::|  . |.| ..::|.|.|||.
MGPPSSAAACSRGSPRILRVRAGGISLFYTI.MAVLLLLLVV.EAGAILAPATHACRANGQYFLTNCCA
```

FIG. 1

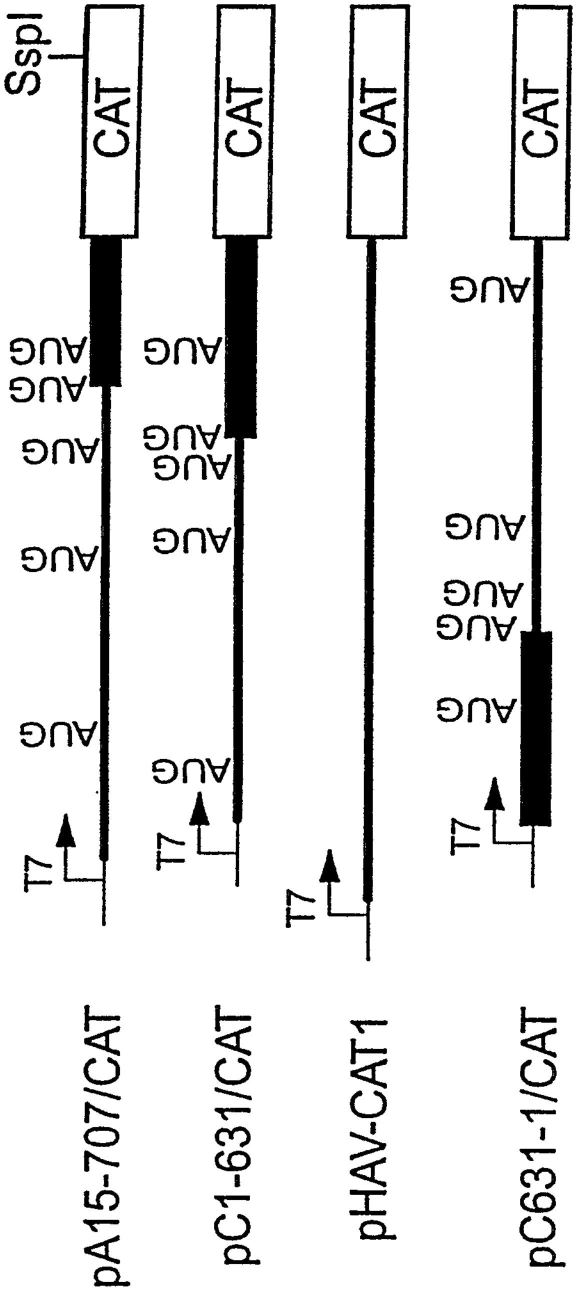

pA15-707/CAT
T7
AUG
AUG
AUG
AUG
AUG
CAT
SspI
pC1-631/CAT
T7
AUG
AUG
AUG
AUG
AUG
CAT
pHAV-CAT1
T7
CAT
pC631-1/CAT
T7
AUG
AUG
AUG
AUG
AUG
CAT

FIG.2A

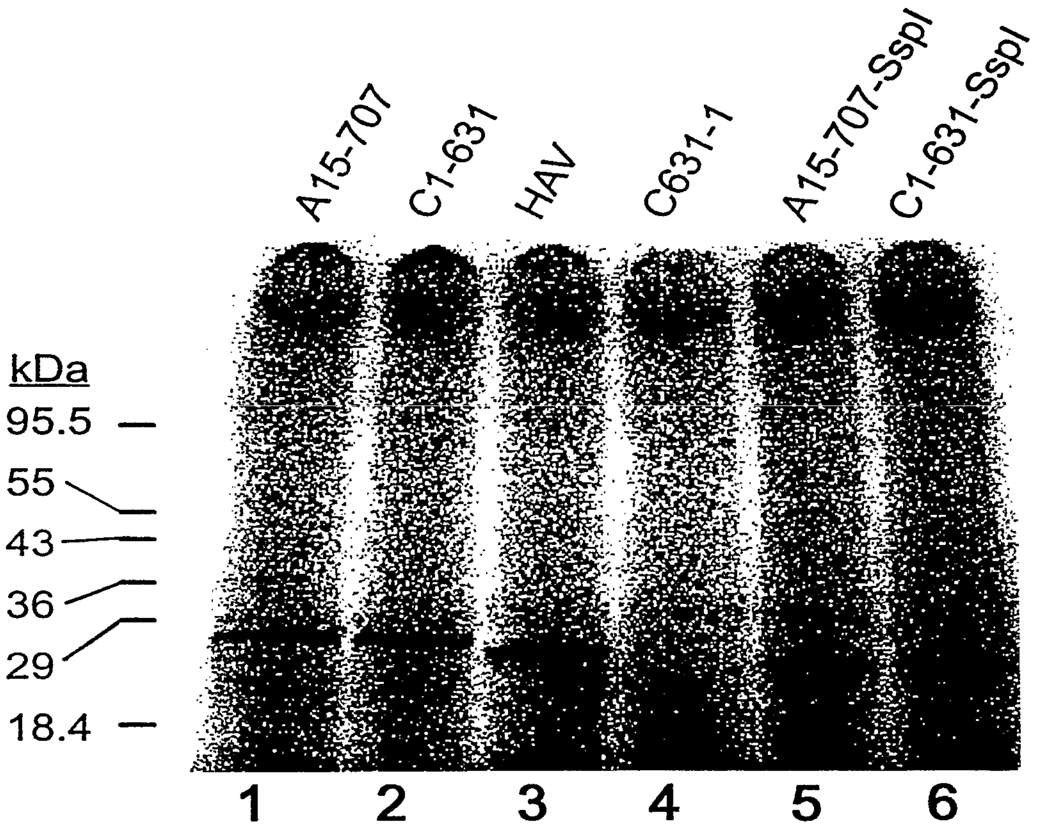

A15-707
C1-631
HAV
C631-1
A15-707-SspI
C1-631-SspI
kDa
95.5
55
43
36
29
18.4
1
2
3
4
5
6

FIG.2B

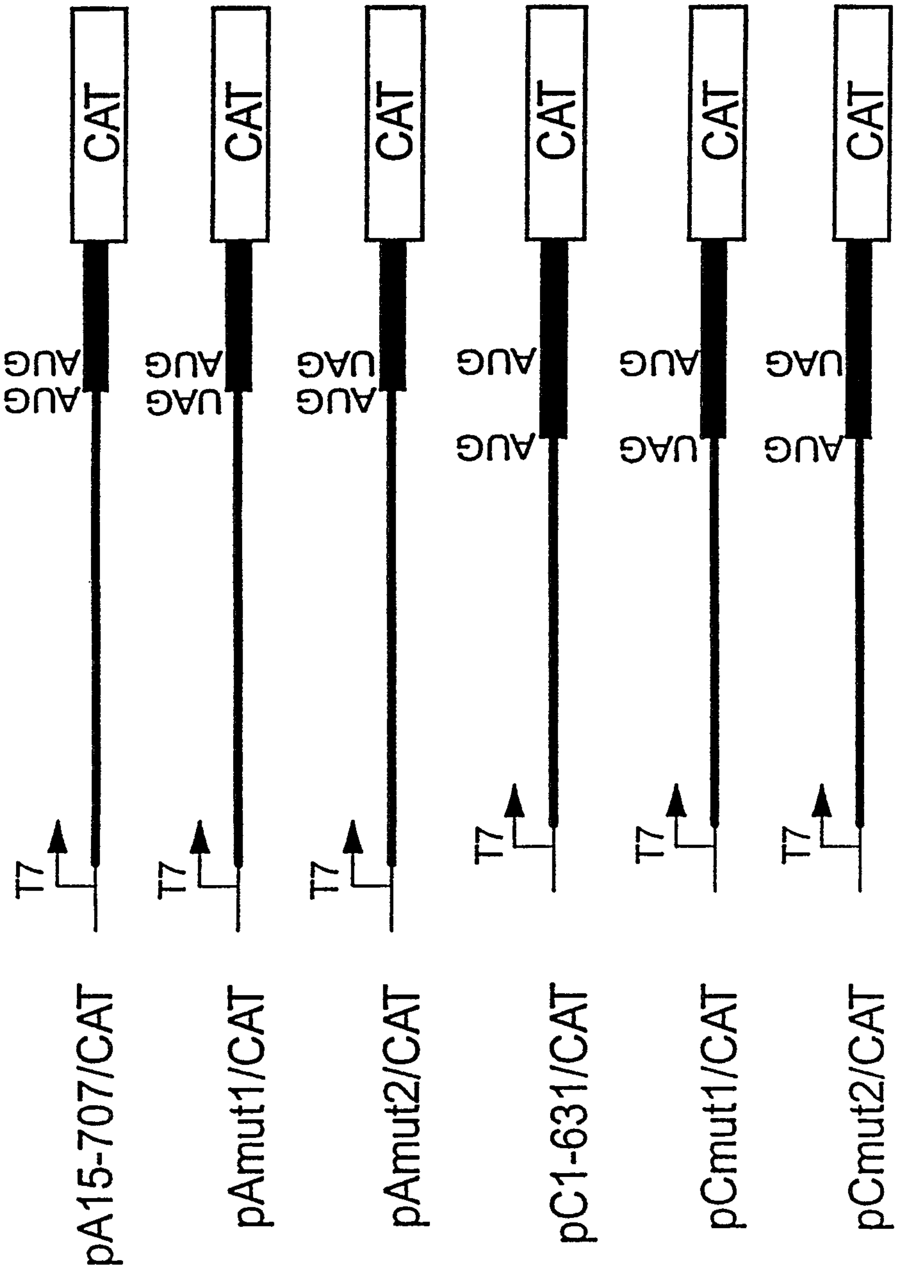

pA15-707/CAT
T7
AUG
AUG
CAT
pAmut1/CAT
T7
UAG
AUG
CAT
pAmut2/CAT
T7
AUG
UAG
CAT
pC1-631/CAT
T7
AUG
AUG
CAT
pCmut1/CAT
T7
UAG
AUG
CAT
pCmut2/CAT
T7
AUG
UAG
CAT

FIG. 3A

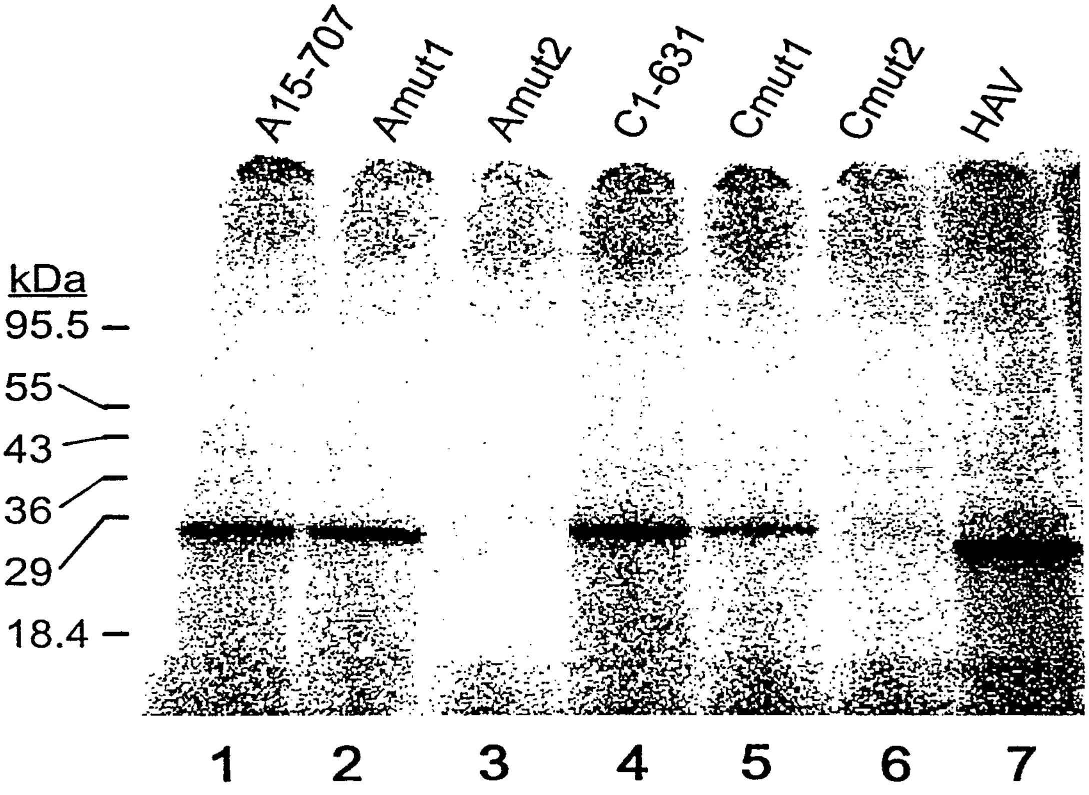

A15-707
Amut1
Amut2
C1-631
Cmut1
Cmut2
HAV
kDa
95.5
55
43
36
29
18.4
1
2
3
4
5
6
7

FIG.3B

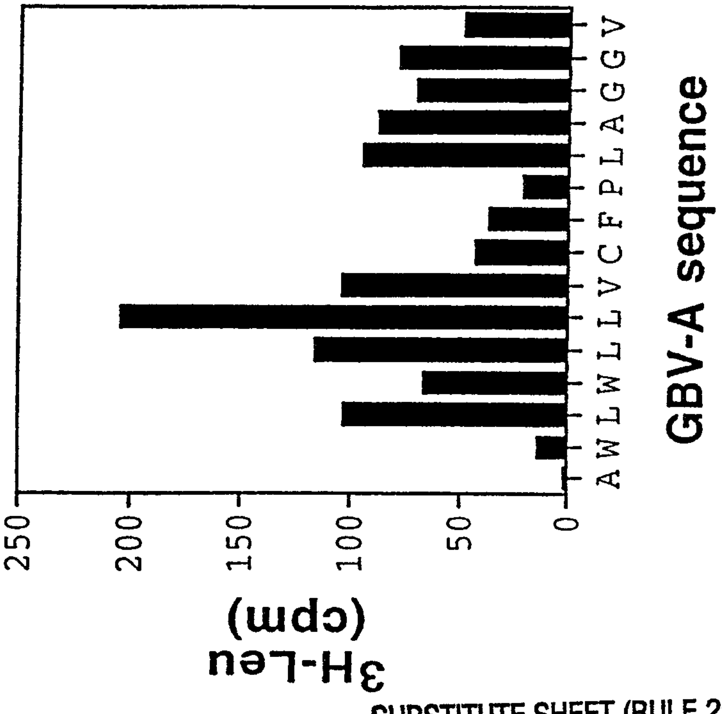

3H-Leu (cpm)
0
50
100
150
200
250
A W L W L L V C F P L A G G V
GBV-A sequence

FIG.4A

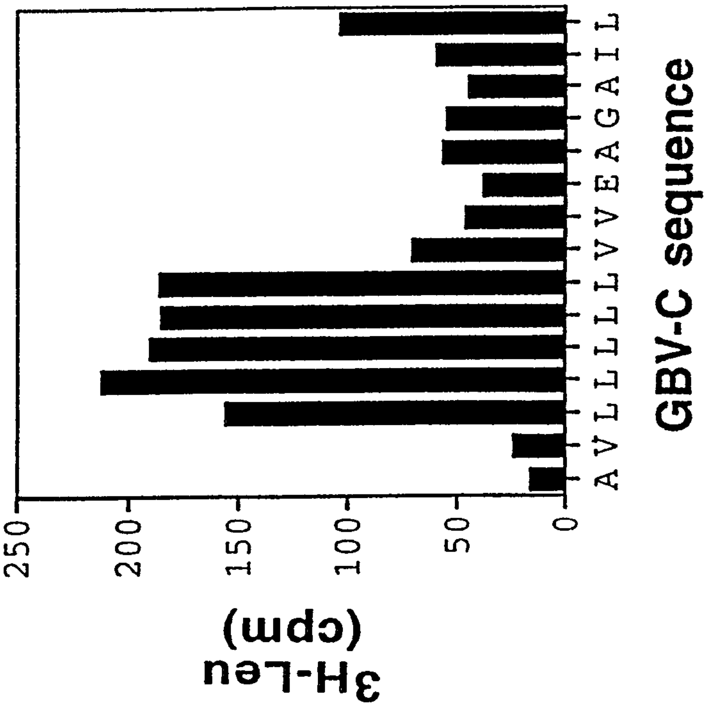

3H-Leu (cpm)
0
50
100
150
200
250
A V L L L L L V V E A G A I L
GBV-C sequence

FIG. 4B

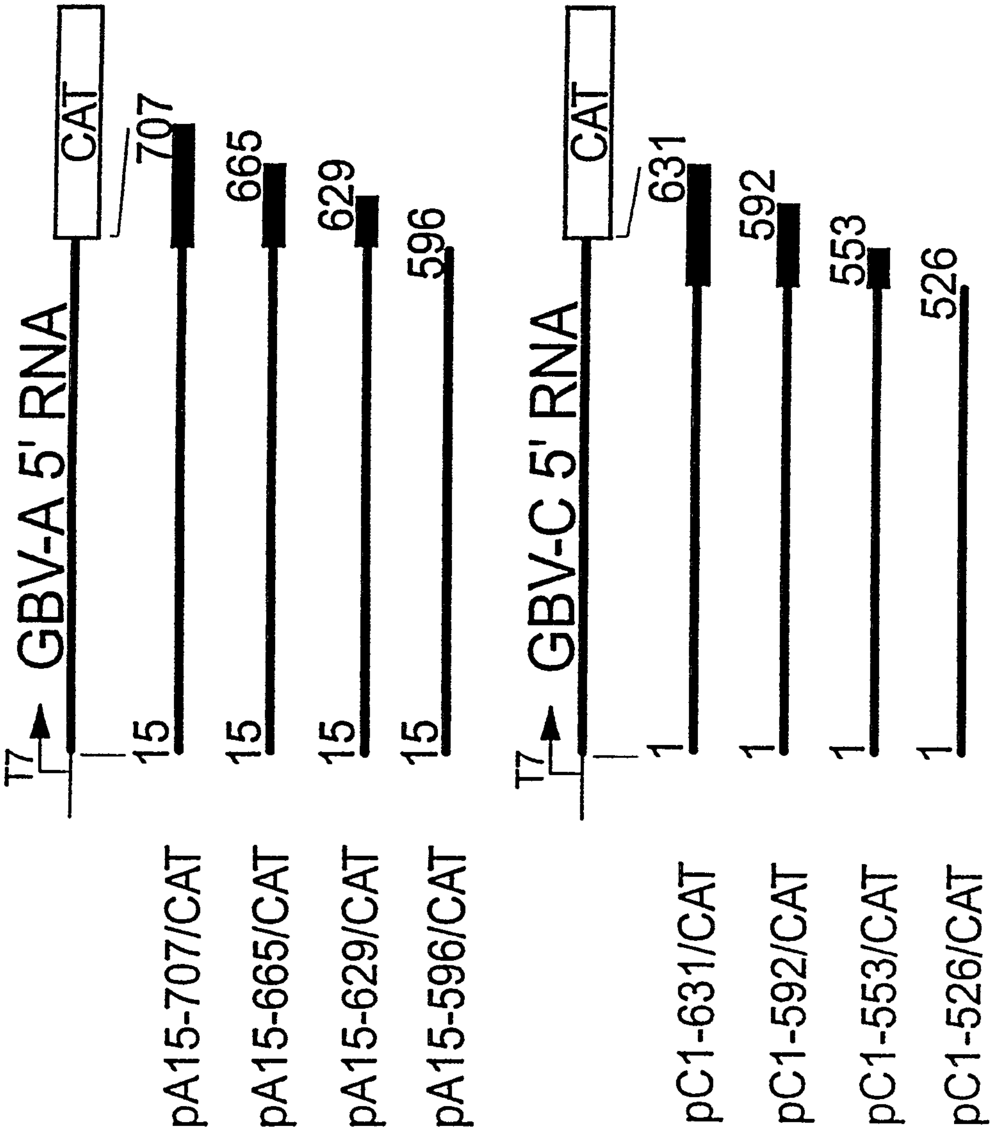

GBV-A 5' RNA
T7
CAT
15
707
665
629
596
pA15-707/CAT
pA15-665/CAT
pA15-629/CAT
pA15-596/CAT
GBV-C 5' RNA
T7
CAT
1
631
592
553
526
pC1-631/CAT
pC1-592/CAT
pC1-553/CAT
pC1-526/CAT

FIG. 5A

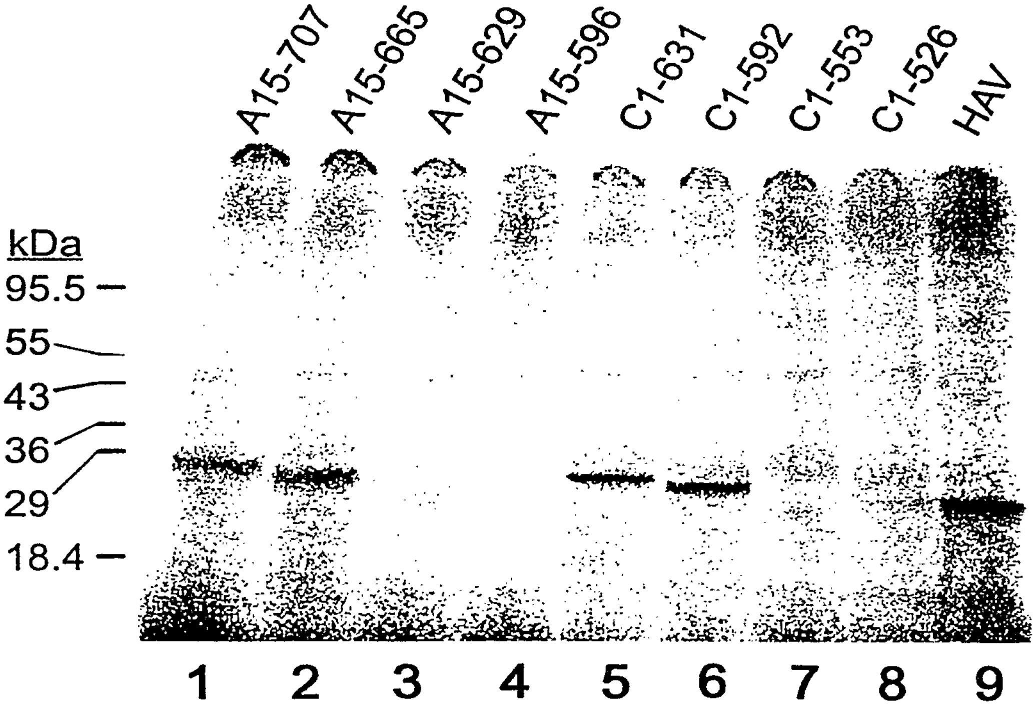

A15-707
A15-665
A15-629
A15-596
C1-631
C1-592
C1-553
C1-526
HAV
kDa
95.5
55
43
36
29
18.4
1
2
3
4
5
6
7
8
9

FIG.5B

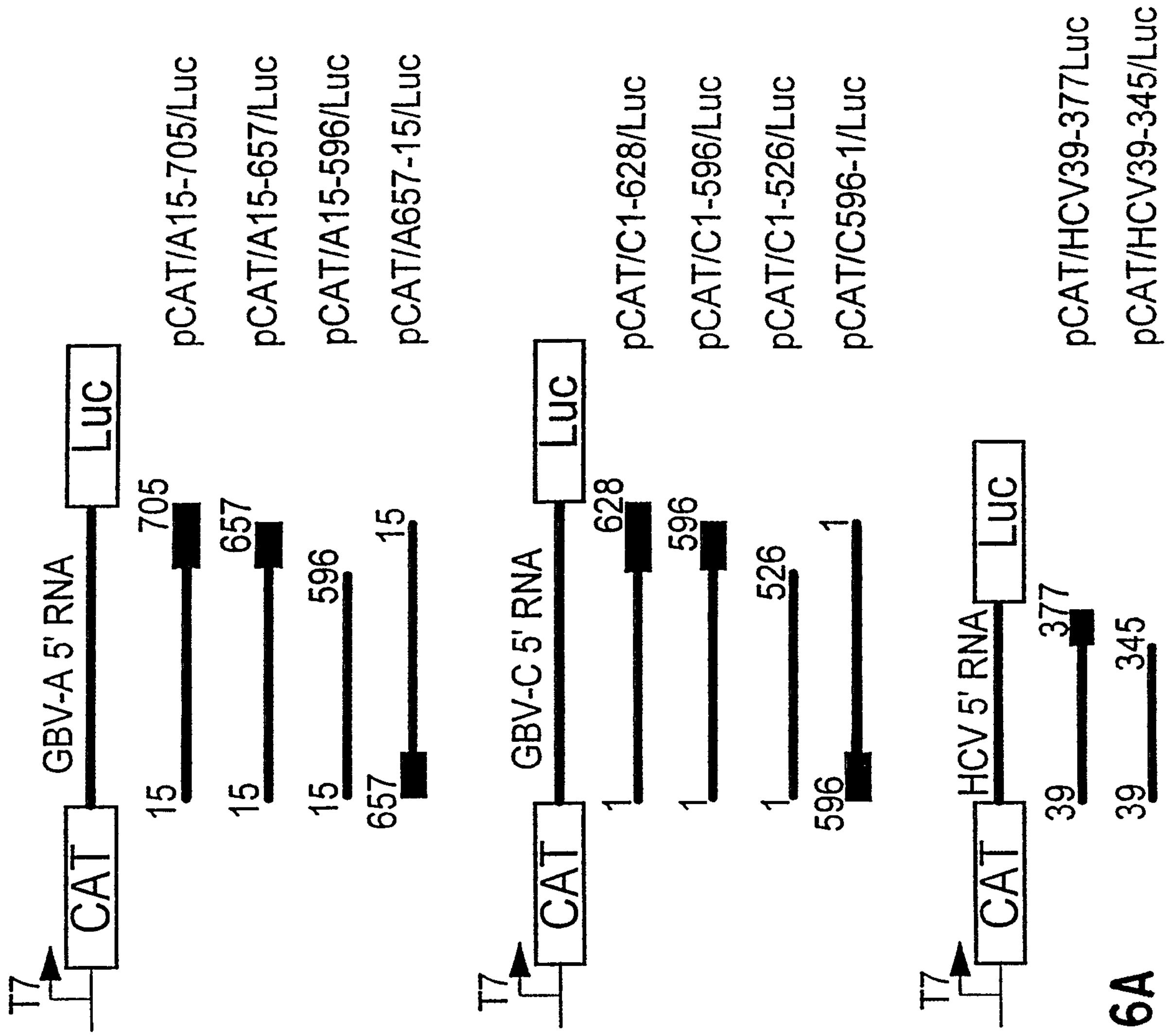

T7
CAT
GBV-A 5' RNA
Luc
15
705
pCAT/A15-705/Luc
15
657
pCAT/A15-657/Luc
15
596
pCAT/A15-596/Luc
657
15
pCAT/A657-15/Luc
T7
CAT
GBV-C 5' RNA
Luc
1
628
pCAT/C1-628/Luc
1
596
pCAT/C1-596/Luc
1
526
pCAT/C1-526/Luc
596
1
pCAT/C596-1/Luc
T7
CAT
HCV 5' RNA
Luc
39
377
pCAT/HCV39-377/Luc
39
345
pCAT/HCV39-345/Luc

FIG. 6A

pCAT/--/Luc	Luc-A	Luc-P
A15-705	10.2	25.2
A15-657	1.13	14.8
A15-596	10.3	13.8
A657-15	0.13	6.37
C1-629	4.09	15.4
C1-596	0.80	12.4
C1-526	13.5	14.7
C596-1	2.02	5.75
HCV39-377	546	403
HCV39-345	270	205

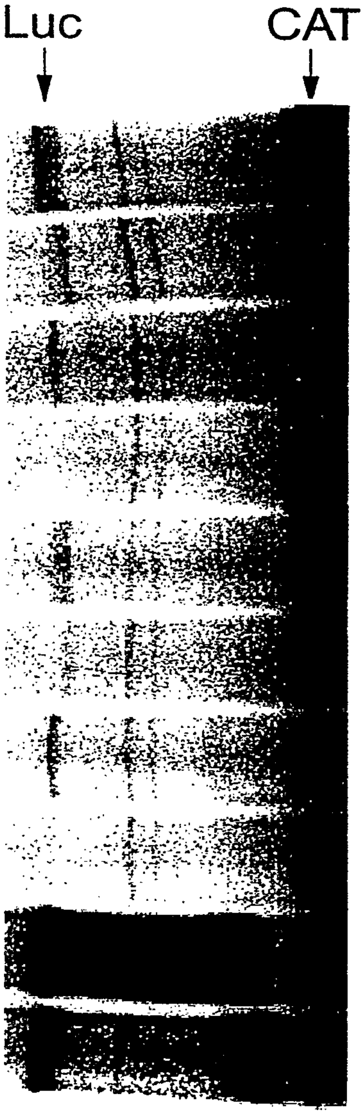
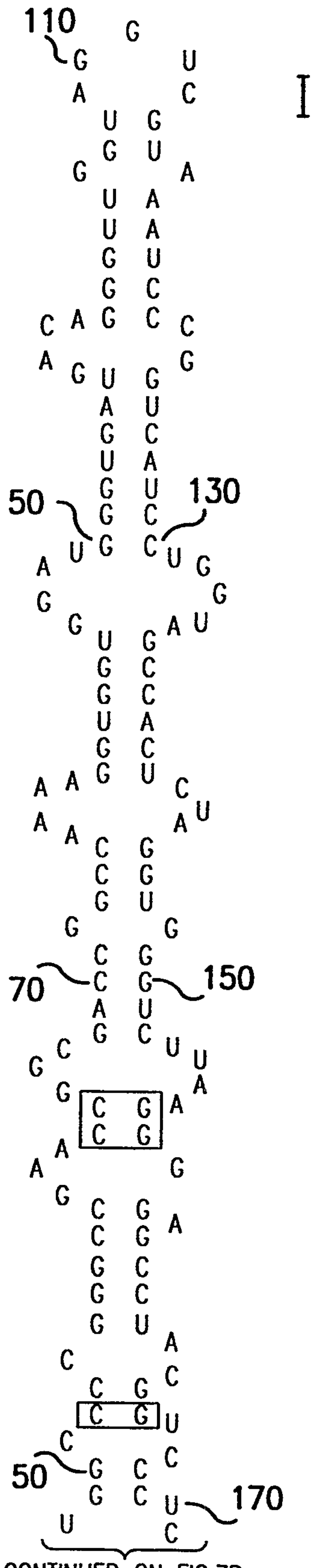
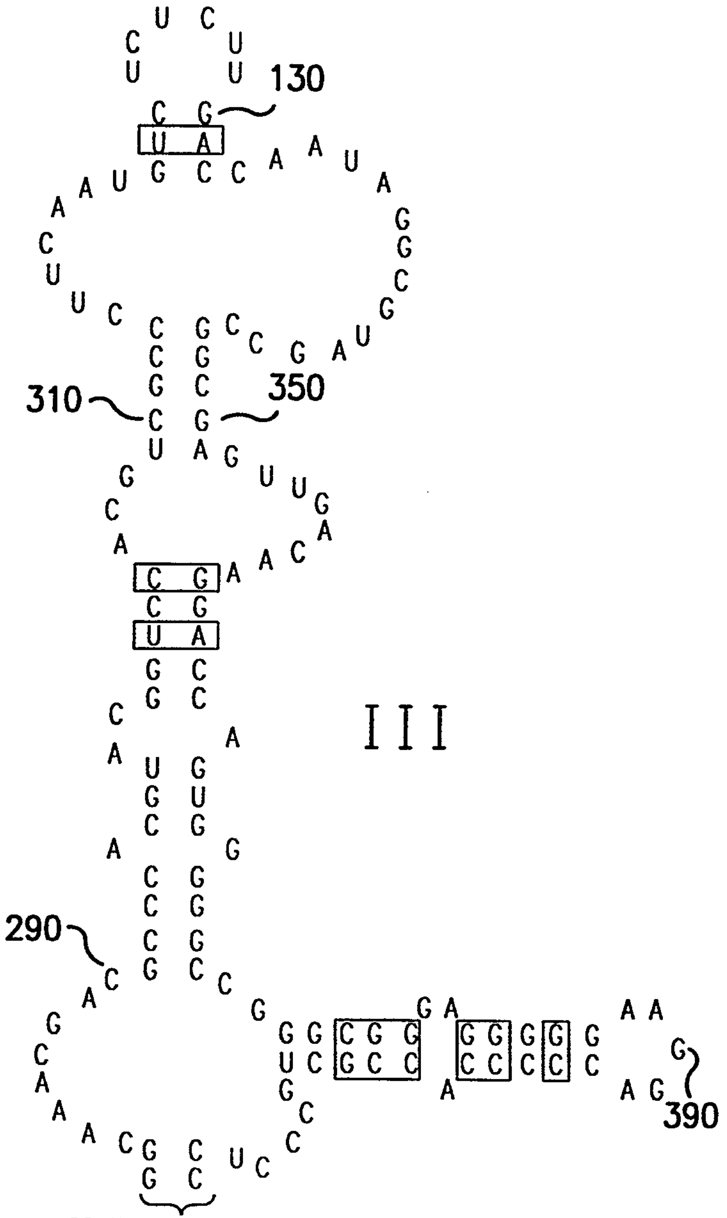


FIG.6B

FIG.7A



CONTINUED ON FIG.7B



CONTINUED ON FIG.7B

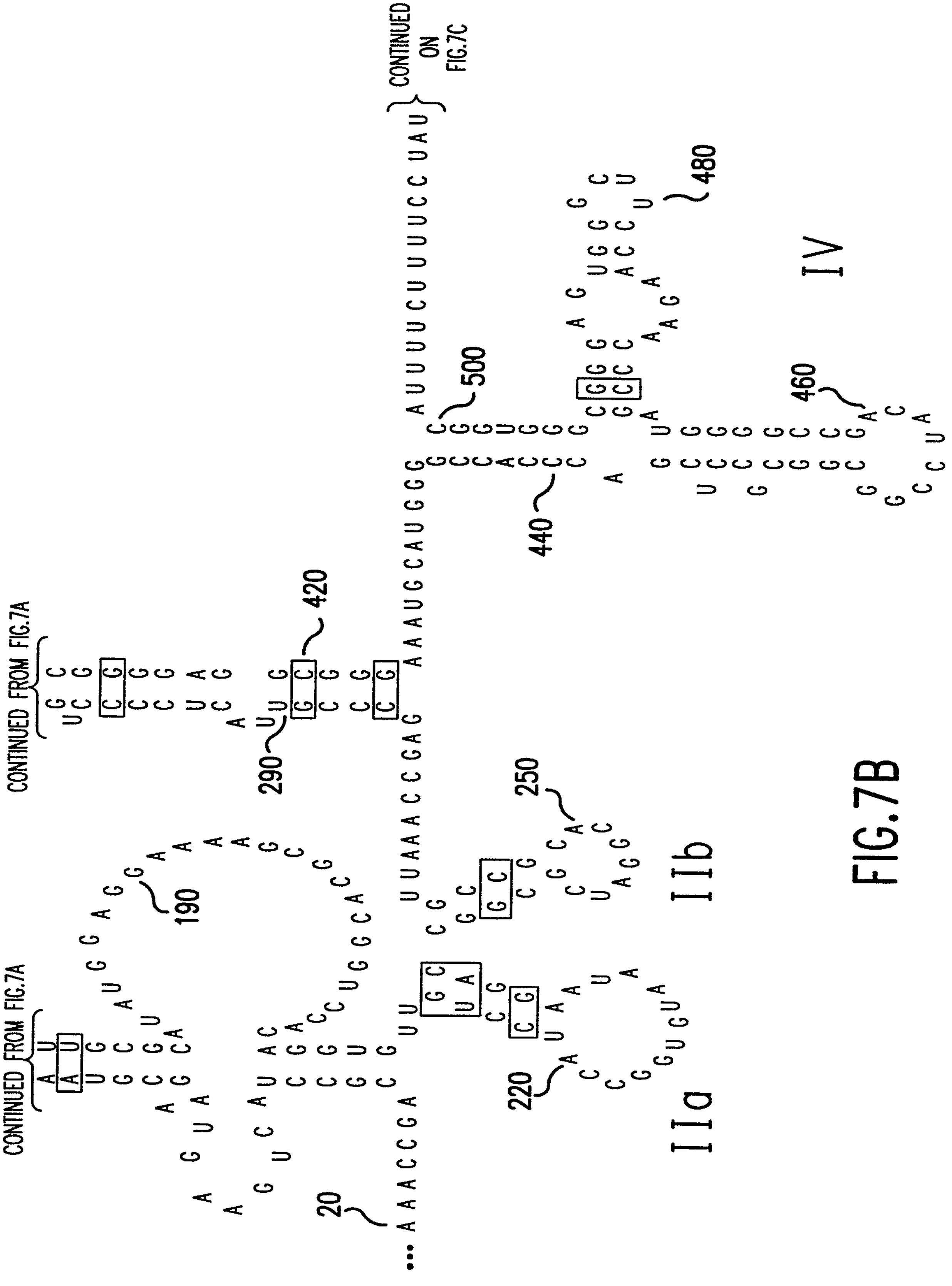

CONTINUED ON FIG.7C
CONTINUED FROM FIG.7A
CONTINUED FROM FIG.7A
20
190
220
250
290
420
440
460
480
500
IIa
IIb
IV

FIG.7B

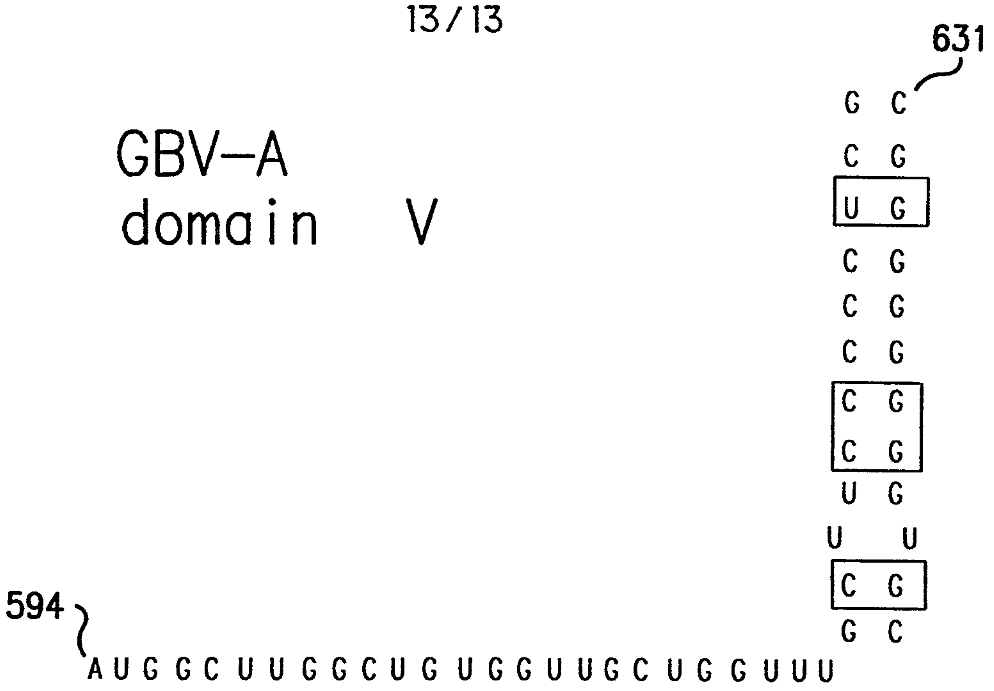

GBV-A
domain V
631
G C
C G
U G
C G
C G
C G
C G
C G
U G
U U
C G
G C
594
A U G G C U U G G C U G U G G U U G C U G G U U U

FIG.7D

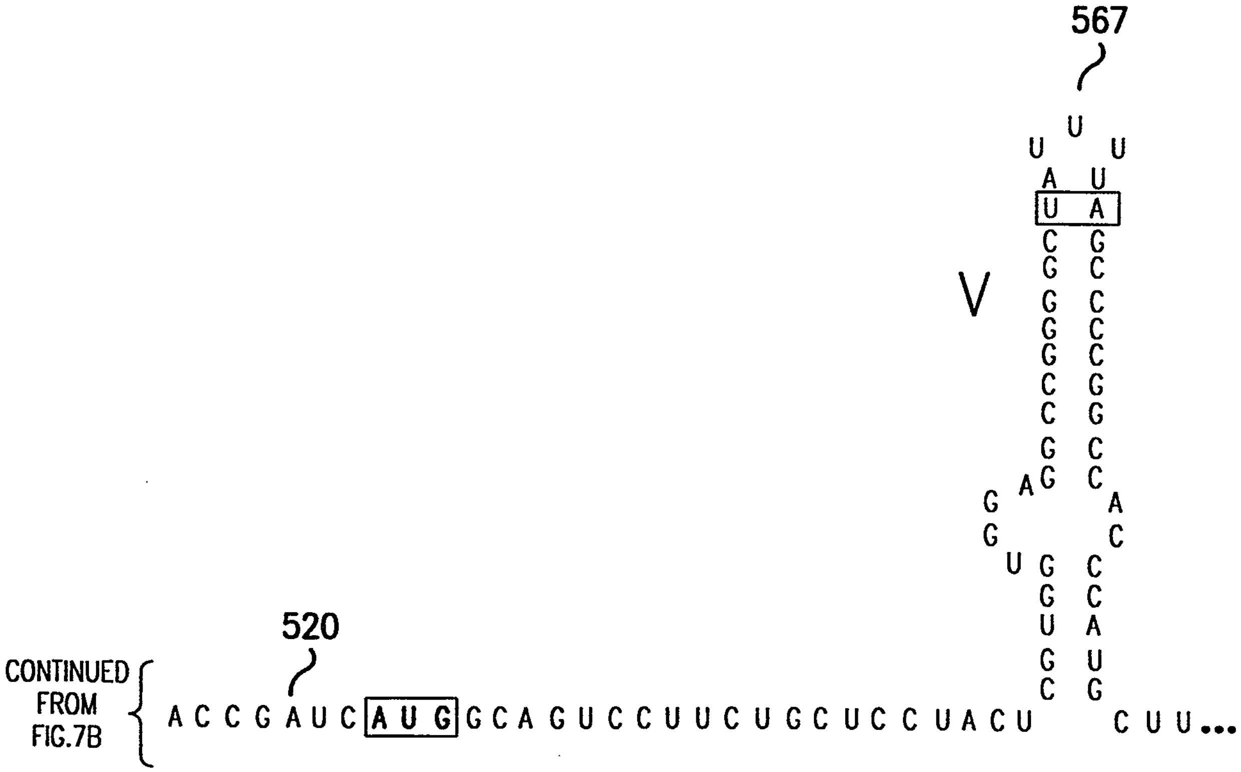

567
U
U U
A U
U A
C G
G C
G C
G C
G C
C G
C G
G C
A G C A
G C
G C
U G C
G C
U A
G U
C G
V
CONTINUED FROM FIG.7B
520
A C C G A U C A U G G C A G U C C U U C U G C U C C U A C U
C U U...

FIG.7C