A Basis for New Approaches to the Chemotherapy of AIDS: Novel Genes in HIV-1 Potentially Encode Selenoproteins Expressed by Ribosomal Frameshifting and Termination Suppression

E. W. Taylor,* C. S. Ramanathan, R. K. Jalluri, and R. G. Nadimpalli

Computational Center for Molecular Structure and Design and Department of Medicinal Chemistry, The University of Georgia, Athens, Georgia 30602-2352

Received April 25, 1994®

Several previously unnoticed genes in the human immunodeficiency virus type 1 (HIV-1), potentially encoding selenoproteins, have been discovered by analyzing the genomic RNA structure and its relation to novel open reading frames. We have found a number of new potential RNA pseudoknots, including one in the long terminal repeat, several that coincide with highly conserved enzyme active site sequences in the *pol* coding region, and one in the *env* coding region. These pseudoknots can potentially direct the synthesis of selenocysteine (SeC) containing −1 frameshift fusion proteins. This is possible because we have found potential SeC insertion sequences (SECIS) in the RNA of HIV and other retroviruses; such structures are known to be necessary and sufficient for the incorporation of SeC at UGA "stop" codons anywhere in a eukaryotic mRNA. In several locations, UGA codons in the −1 reading frame are highly conserved across a broad spectrum of primate immunodeficiency viruses. Due to the degeneracy of the genetic code, this conservation cannot be explained by evolutionary selection of the *pol* gene protein sequence alone. Such observations, combined with the conservation of the associated reading frames, strongly suggest that these are real genes, and thus that the pseudoknots are also real. A protease pseudoknot-directed −1 frameshift fusion protein contains a highly conserved SeC codon and has significant similarities to a number of DNA binding proteins, including papillomavirus E2 proteins, suggesting it may be a virally encoded repressor of HIV transcription when cleaved by protease from the rest of the *gag–pol* gene product. A reverse transcriptase (RT) frameshift fusion protein replaces the RT active site with a highly conserved SeC-containing module. An integrase frameshift fusion protein contains the N-terminal integrase DNA-binding domain and a potential ATP-binding "GKS" motif; it has significant similarities to several helicases, but no SeC codons. A potential frameshift fusion protein from *env* has one SeC codon, but not in a highly conserved position. SeC incorporation could extend the *nef* gene product by 33 residues through the C-terminal UGA codon without frameshifting, potentially leading to substantial SeC utilization in infected cells. Significantly, a characteristic decline in plasma Se has been observed in ARC and AIDS patients; this can contribute to the pathology of AIDS, because Se is a constituent of glutathione peroxidase and one of the enzymes involved in thyroid T3 hormone synthesis. Our results suggest the possibility that the symptoms of progressive Se depletion in AIDS (e.g., impairment of antioxidant status) are not only due to malabsorption of Se, but also due to sequestration of Se in viral proteins, particularly within infected cells. We have also found potential SECIS elements encoded in the mRNA of other viruses, e.g., polio and coxsackie B, which is inhibited by Se compounds *in vivo*. Independent of questions raised regarding the possible roles of Se, these potential novel gene products should offer various opportunities for new approaches to anti-HIV therapy.

Retroviruses utilize a number of elaborate transcriptional and translational control mechanisms in order to influence not only the timing of gene expression, but also to precisely balance the relative quantities of their various gene products. The latter is necessary because structural proteins (products of the retroviral *gag* and *env* genes) are needed in much greater quantity than the products of the *pol* gene, which encodes the viral enzymes protease, reverse transcriptase (RT), and integrase.

Since the *pol* gene lacks an independent start codon (Figure 1A), the *pol* gene products can only be synthesized as *gag–pol* fusion proteins, the formation of which is controlled at the translational level.[1] The two alternative mechanisms that are used in controlling the relative amounts of *gag* and *pol* gene products are *ribosomal frameshifting*, when the genes are partially overlapping in different reading frames,[2] or *termination suppression*, when the two genes are in the same reading frame, separated by a stop codon; the latter is less common.[3-5] Since both of these processes are inefficient, these fusion proteins are formed only on the relatively infrequent occasions (approximately 1–10% of the time) when there is either a readthrough of the *gag* stop codon (when *gag* and *pol* are in the same reading frame) or when ribosomal collision with a downstream stem or pseudoknot structure[6,7] kicks the message into the −1 reading frame. This may involve a simple recoil mechanism and requires a slippery

* Author to whom correspondence should be addressed. Email: wtaylor@rx.uga.edu

® Abstract published in *Advance ACS Abstracts*, July 15, 1994.

0022-2623/94/1837-2637$04.50/0

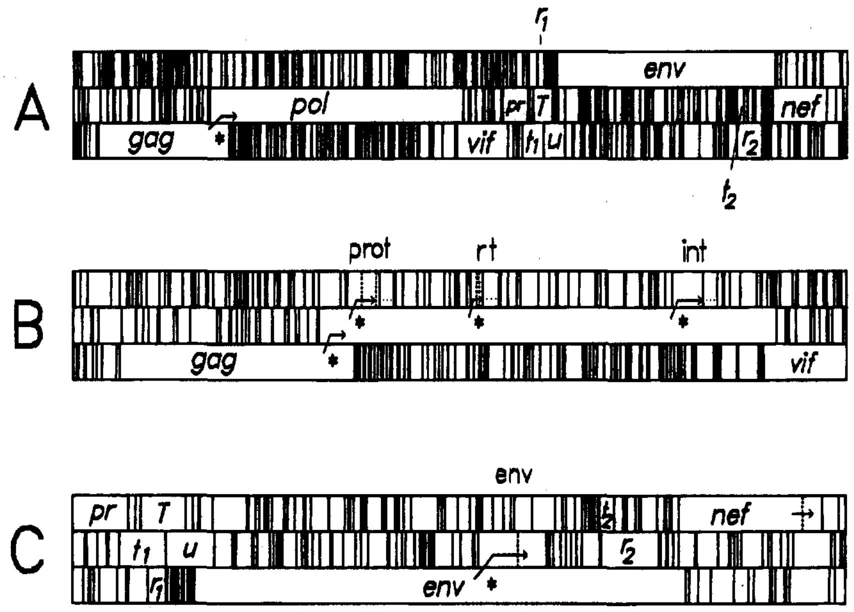

Figure 1. Open reading frames in HIV-1. Known and potential pseudoknots are shown by an asterisk, with the bent arrow symbol indicating potential for pseudoknot-directed frameshifts into open regions of the −1 reading frame. Vertical lines represent stop codons, which appear as thick bars when densely packed in the same reading frame. Selected UGA codons, potentially encoding selenocysteine, are shown as dashed vertical lines. Dotted horizontal lines indicate the potential for extension by termination suppression of amber or ochre stop codons (see Figure 12) (A) Entire sequence, showing known genes, and known *gag–pol* frameshift site. In addition to *gag*, *pol*, *env*, *vif*, and *nef*, *r1* and *r2* are first and second exons of *rev*, *t1* and *t2* are first and second exons of *tat*; *pr* is *vpr* and *u* is *vpu*; *T* is the T open reading frame.[11] (B) Expansion of the 5′ half, showing three additional potential pseudoknots and −1 frameshift sites in the protease coding region (PROT), the RT coding region (RT), and the integrase coding region (INT). Each of these sites has the potential to direct the synthesis of a fusion protein beginning in the indicated gene regions. (C) Expansion of the 3′ half, with the *env* reading frame rotated to the bottom, showing a potential pseudoknot and −1 frameshift site in the *env* coding region, and a potential readthrough of a UGA codon at the end of *nef* that could extend the *nef* gene product by 33 amino acids. These figures were generated from the output of the program ORFwriter.[91]

sequence in the message, usually located about 10−15 bases upstream from the base of the 5′ stem of the pseudoknot. This is known as ribosomal frameshifting.[2,8,9] It is also noteworthy that pseudoknots have recently been implicated in termination suppression at the *gag–pol* junction in certain type C retroviruses.[10]

The possibility that similar mechanisms might be used to control the synthesis of regulatory proteins required only in very small amounts has been previously pointed out by Cohen *et al.* in their paper on the T open reading frame (ORF) of HIV-1, which is potentially expressed by a −1 frameshift from the *tat* gene.[11] Consistent with this idea, given the inherently low level of *pol* gene expression relative to *gag*, it is probable that a second −1 frameshift from *pol* would yield potentially minute quantities of product, depending on the efficiency of the second frameshift event. This would not be unprecedented; in at least one retrovirus, two successive frameshifts are necessary for complete translation of *pol*; these are known to occur at surprisingly high efficiency.[8] Detailed molecular analysis has shown that for *highly efficient* frameshifting to occur,[12] the slippery sequence ideally should be of the form X XXY YYZ, where triplets represent codons in the zero reading frame; however, deviations from this pattern are known, such as that at the *pro–pol* frameshift site in the mouse mammary tumor virus (MMTV), which is G GAU UUU.[9] It is also worth considering that a low-efficiency frameshift, and thus a less than ideal slippery sequence, could be desirable in some situations, e.g., where very low levels of a potent regulatory protein might be required.

In this paper, we will make a case for the probable existence of three new HIV-1 fusion proteins that can only be formed by −1 frameshifting from the *pol* gene, under the direction of potential pseudoknots in the protease, RT, and integrase coding regions (Figure 1B). We will also point out the existence of an ORF that lacks a start codon but could be expressed as a −1 frameshift fusion protein from *env* (Figure 1C); this ORF is well-conserved in primate retroviruses. In addition, the potential for an extended form of the *nef* gene product will be suggested.

Three of these four potential novel −1 frameshift fusion proteins may contain one or more selenocysteine residues, encoded by opal stop codons (UGA), which are now known to potentially code for selenocysteine in both eukaryotic and prokaryotic genomes, if appropriate signals are present in the mRNA, specifically SECIS elements (selenocysteine insertion sequences) in eukaryotic cells;[13,14] a somewhat different but related mechanism is used in bacteria.[15] We will also demonstrate the presence of *multiple* potential SECIS-like elements in the HIV-1 mRNA and, indeed, in the RNA of many other retroviruses and retroelements that we have examined. The presence of a SECIS element in a eukaryotic mRNA has been shown to be both necessary and sufficient for the incorporation of selenocysteine at in-frame UGA "stop" codons anywhere in the mRNA where it is located.

Although the existence of these potential genes and the RNA structures that would make their expression possible is at present purely theoretical, we will argue that there is a high probability that these are real genes due to (1) the conservation of sequence in a −1 reading frame with respect to known coding regions, (2) conservation of the reading frames associated with these

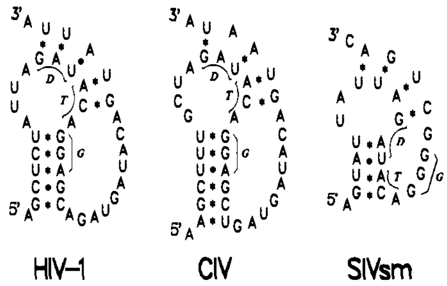

Figure 2. Novel potential pseudoknots in the protease coding region, which are directly associated with the highly conserved DTG consensus sequence of retroviral proteases, placing most of the bases of the DTG codons into base pairs. This D is the catalytic aspartate of the protease. CIV is chimpanzee immunodeficiency virus; SIVsm is simian immunodeficiency virus, sooty mangabey.

potential genes, (3) at least in several cases, the existence of significant sequence similarities to proteins that are known to occur in other viruses, and (4) the consistency between the implications of the theoretical model and a large body of experimental and clinical data related to selenium, AIDS, and viruses in general.

Novel Pseudoknots in HIV-1 Protease and RT Coding Regions Are Associated with Codons for Active-Site Consensus Sequences and Catalytic Aspartate Residues. Schinazi *et al.*[16] recently reported a correlation between sites of drug resistance mutations in the HIV-1 RT coding region and features of the predicted RNA structure. The results suggested that unpaired bases, or paired bases immediately adjacent to unpaired regions, are more mutation prone, whereas codons corresponding to highly conserved regions are associated with stacked paired bases located in more extended helical structures. These findings are also consistent with the results of a previous study that correlated envelope protein hypervariable regions with nonhelical RNA structural regions in the *env* coding region of HIV-1.[17]

Consistent with this hypothesis, we have discovered several potential RNA pseudoknot structures (see the Experimental Section) associated with highly conserved coding sequences. One is precisely associated with codons of the highly conserved DTG consensus sequence of retroviral proteases, including the catalytic Asp25 of HIV-1 protease (Figure 2) and another with the highly conserved YMDD (motif C) region of RT, which includes two of the three aspartates of the RT catalytic triad (Figure 3). The major stem of this RT pseudoknot was predicted using the Zuker FOLD program,[18] which is, however, unable to predict potential pseudoknots, because the empirical energy parameters for their relative thermodynamic stability are not accurately known. Nonetheless, the 5′ stem of the potential pseudoknot was sufficiently stable to be predicted in a global fold of the RT-coding RNA.[16] It may also be significant that the RT motif C pseudoknot has several topological similarities to the known pseudoknot[2,7] at the *gag–pol* frameshift site (Figure 3). Both the RT and *gag–pol* pseudoknots have the zipper-like property of being able to exist in an equilibrium between two extreme conformational states that are both pseudoknots; they also have the same size 5′ loop and 3′ stem in one of those conformations.

Both of these new potential pseudoknots in HIV-1 place almost every base of each codon of the respective consensus sequences into a base pair (8 of 9 in protease, 11 of 12 in RT). Based on the statistic reported by Schinazi *et al.* for the RT coding region (about 57% of total bases were paired in the global minimum energy structure), the probabilities of having that many bases paired in the two consensus sequences purely by chance are $P < 0.05$ and $P < 0.02$, respectively. The probabilities of having the number of contiguous stacked base pairs observed in the pseudoknots are substantially lower, $P < 0.01$ and $P < 0.0005$, respectively. This appears to be further evidence in support of the hypothesis that evolution may select RNA structures that place codons for critical conserved sequences in helical regions, suggesting that such structures are less prone to mutation.[16]

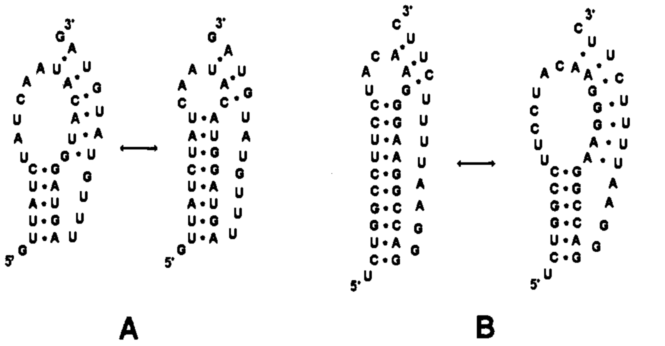

Figure 3. Comparison of the known frameshift-directing pseudoknot (B) in the HIV-1 *gag* region with a novel potential pseudoknot in the RT coding region (A), which is directly associated with the highly conserved YMDD consensus sequence of retroviral RTs, placing most of the bases of the YMDD codons into base pairs. The YMDD motif sequence begins at UAC (Y) immediately following the upper (5′) loop. This motif includes two of the three aspartates of the RT catalytic triad. The potential for a zipper-like change in conformational state, shared by both of these pseudoknots, is indicated by arrows.

A comparative analysis of potential RNA structures at these locations (for which a few examples are shown in Figures 2 and 4) for various related primate retroviral sequences shows that pseudoknot type structures, or at least relatively stable stem structures (for other sequences not shown), are located at these positions in many instances; however, the details of these structures vary considerably. Thus, it is the structural *theme* of a pseudoknot that is conserved rather than an identical pseudoknot structure being observed in the various cases.

Potential Novel Genes Encoding Selenoproteins Can Only Be Expressed by −1 Ribosomal Frameshifting from *pol* Induced by the Protease and RT Pseudoknots. Since one of the few well-established functions of pseudoknots is the direction of ribosomal frameshifting events,[1,7,19] it was obviously necessary to examine the possibility that potential fusion proteins might be synthesized by −1 frameshifts from the *pol* gene at these points. A cursory initial examination by translation of the relevant regions of the reading frame that is −1 with respect to *pol* was not encouraging, since there were stop codons in this reading frame within about 10 amino acids of any potential frameshift sites. However, working under the assumption that termination suppression of some kind might conceivably be involved, the translated sequences between the first and second stop codons in the −1 reading frame (only 29

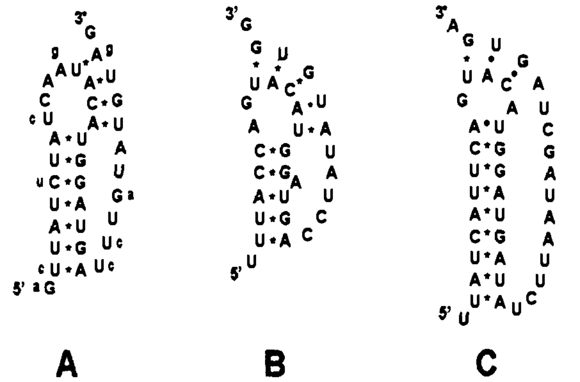

Figure 4. Comparison of the RT YMDD pseudoknot from several primate lentiviruses. The YMDD motif sequence begins at UAC (Y) immediately following the upper (5') loop. (A) HIV-1 variants (upper case) and SIV-chimpanzee (CIV) with alterations from HIV-1 shown in lower case, showing that six of nine mutated bases in CIV are on loop regions and that the changes in helical regions are consistent with maintaining base pairing, except in the case of one broken base pair, which only shortens the 5' stem by one base pair (UA to CA). (B) An alternative conformation for the CIV pseudoknot, intermediate between the HIV-1 and HIV-2 structures (A and C). (C) Potential YMDD pseudoknot for HIV-2, showing that the base pairings have totally shifted with respect to those seen in HIV-1 (A), but that the structural theme of a pseudoknot is maintained.

and 13 amino acids long for the protease and RT frameshifts, respectively) were scanned for potential matches against the entire GenBank database, using the FASTDB program (see the Experimental Section). This produced a number of intriguing matches at high levels of significance, particularly with various DNA binding and finger proteins for the short RT −1 frameshift sequence (at 5 to 7 SD significance relative to the database average), and with a number of viral proteins, class I MHC antigens, ribosomal and DNA binding proteins (at 4.5 to 7 SD) for the protease −1 frameshift sequence. This prompted a more detailed literature search for possible mechanisms by which there might be a readthrough of some of the stop codons (i.e., termination suppression), because this was already known to be a mechanism for the synthesis of *pol* gene products in some retroviruses,[4,20,21] as discussed above.

Selenocysteine and the Opal Codon. The UGA stop codon, known to geneticists as the opal codon, has in recent years been realized to be unique in the genetic code, due to its ability to code for the 21st amino acid, selenocysteine, as well as to function as a stop codon in conventional contexts. As stated in a recent review: "...UGA may originally have been a sense codon for selenocysteine, the use of which was counterselected for by the introduction of oxygen into the earth's atmosphere. This excluded the use of this highly oxidizable amino acid except to anaerobic or well-protected chemical environments. Indeed, the finding that UGA encodes this amino acid in both prokaryotes and eukaryotes indicates that it developed before the two lineages separated."[13]

Since viruses (and RNA viruses in particular) have been described as "molecular fossils"[22] (a description which more likely pertains only to a portion of their molecular machinery), the possibility that they might still depend upon primordial chemistry certainly seemed worth investigating. Thus, possible protease and RT fusion proteins that could be expressed under the direction of these pseudoknots, with the insertion of selenocysteine at UGA codons, were translated and studied by database scanning, alignment with similar potential gene products from other primate retroviruses, etc. However, before presenting and discussing these results, evidence for the existence of potential SECIS-like elements in HIV-1 and other retroviral genomes will be presented, since UGA could not function as a selenocysteine codon without them.

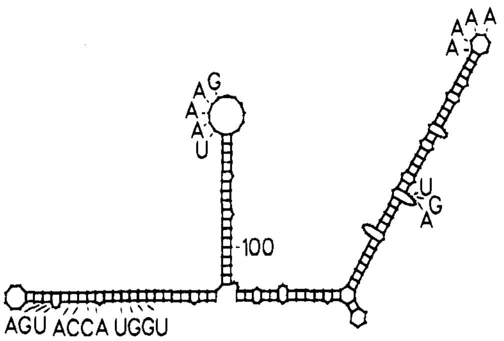

Figure 5. Global minimum energy conformation for the first 250 bases of the 5' end of the HIV-1 mRNA, showing the TAR element at left, which places an E2-like palindromic sequence (UGGUUAGACCA) in the helical stem of TAR. A UGA codon is near the end loop of TAR where Tat protein is believed to bind. The next major structure (middle stem with UAAAG SECIS motif on loop) is strikingly similar to the 5'-DI SECIS loop structure, but there is no UGA codon displayed on a projecting loop (see Figure 6). Both these structures are in the LTR repeat region, so a second copy is also located at the 3' end of the HIV-1 mRNA. The third structure (right stem) is in the 5' untranslated region only, and is a potential SECIS stem. As well as being the predicted global minimum ($\Delta G = -78.9$ kcal/mol), this entire structure (all three stems) is significant at 2.4 SD ($P < 0.01$) relative to global minima for random oligonucleotides of the same size and base composition. The TAR stem conformation shown here is identical to that reported previously by other authors.[23,32]

Multiple Potential SECIS Elements in the HIV-1 RNA. Using a simple systematic sequence scanning technique, followed by RNA folding of the regions of interest (see the Experimental Section), a number of potential SECIS structures were found in the HIV-1 RNA.

An essential feature of SECIS elements is a stem−loop structure displaying a conserved triad of unpaired adenine bases on the loop, e.g., in the context of the sequence UAAAG in the iodothyronine deiodinase (5'-DI) gene.[14] Further down the stem, there is a UGA codon protruding on a bulge, usually with at least two bases of the UGA codon unpaired in the isolated RNA. This UGA apparently combines with the anticodon of a selenocysteine tRNA; the aminoacyl acceptor arm and bound selenocysteine are recognized by the joint action of a special protein translation factor and/or the SECIS loop AAA sequence. In association with this protein factor (probably equivalent to the bacterial *selB* protein[15]), which may bind to the ribosome, the RNA SECIS structure essentially acts as an attractor for selenocysteine tRNAs, making them available for insertion at any in-frame UGA codon in the mRNA, leading to termination suppression of opal stop codons.[13,14] This complex may also inhibit binding of the termination factor.

To date, we have discovered six such potential SECIS stems in the HIV-1 RNA (Figure 5−7); there may be more. Two are identical, being in the R repeat region

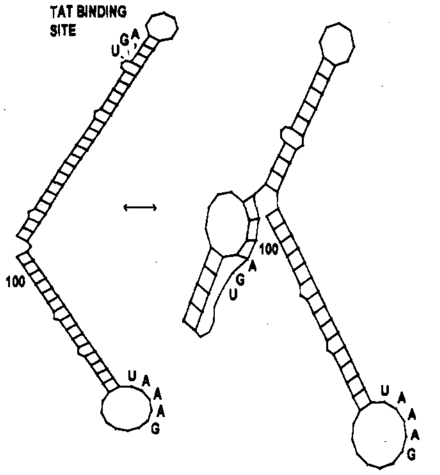

Figure 6. A possible alternative conformation for the TAR stem of Figure 5, which places the E2-like palindromic sequence in a pseudoknot and displays the UGA codon on the 3′ loop of the pseudoknot, where it can potentially work with the SECIS-like structure (UAAAG motif) to serve as a SECIS element at both ends of the HIV-1 mRNA. Since Tat protein is known to bind on the TAR stem (left) precisely where the UGA codon is located, Tat binding may stabilize that conformation, and thus inhibit SECIS function. Proteins that stabilize pseudoknots might facilitate SECIS function. The significance of the structure at right, excluding the pseudoknot (which the FOLD program is unable to predict), is 2.0 SD ($p < 0.01$). The free energies of these two structures cannot be directly compared since the free energy contribution of the pseudoknot is unknown.

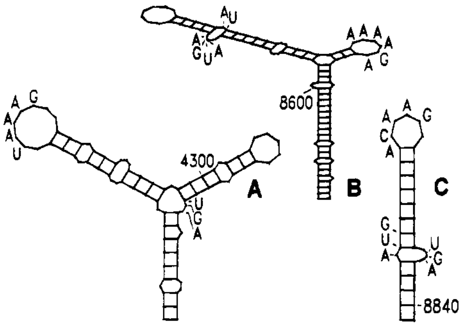

Figure 7. Other potential SECIS elements in the HIV-1 mRNA, located as follows: (A) in the integrase coding region, 3.6 SD, $p < 0.001$; (B) at the very 3′ end of the *env* coding region, 2.7 SD, $p < 0.004$; and (C) just inside the 3′ untranslated region, 4.8 SD, $P < 0.0001$.

of the long terminal repeat (LTR), which is the only region of the LTR that would place one at each end of the processed HIV-1 RNA; this structure may also overlap with the "TAR element" [23] (Figure 5 and 6). In its 3′ location, the adenine-rich potential SECIS motif also serves as the polyadenylation signal. Another potential SECIS element is in the U5 region of the LTR, and thus in a 5′ untranslated region (Figure 5, right), and two others are in or near the 3′ untranslated U3 region (Figure 7), in addition to the one in the 3′ untranslated R region. Another is in the integrase coding region (Figure 7A). Several of these have the entire UAAAG SECIS motif observed in the mRNA of the cellular 5′-DI gene.[14] The HIV-1 SECIS-like elements are more complex in that several of them have one or more additional stems inserted as side arms off the main SECIS stem, and the one in the LTR repeat region may display the UGA codon on unpaired bases on the 3′ loop of a potential pseudoknot 5′ to the SECIS stem (Figures 6 and 13A). This pseudoknot will be discussed later in the context of a potentially critical DNA sequence that it is associated with. These potential SECIS-like RNA structures (Figures 5−7) are exceptionally stable in some cases, as demonstrated by a comparison of their free energies to those of randomized sequences,[16,24] generally giving probability levels of 10^{-2} to 10^{-3} or less. These results demonstrate that there are a number of RNA regions in HIV-1 with the potential to function as SECIS elements; if at least one of them actually functions, then HIV-1 would have the ability to insert selenocysteine at in-frame UGA codons.

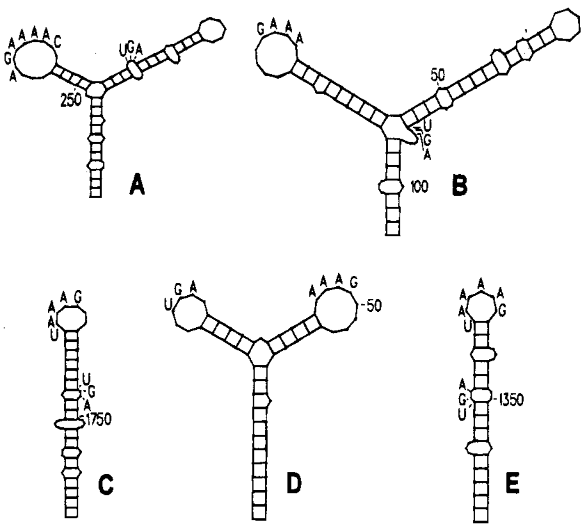

Figure 8. Potential SECIS elements in other RNA viruses (see the Experimental Section for a listing of the virus strains and GenBank accession numbers). (A) Coxsackie B5, 0.8 kcal/mol from global minimum, so only 0.8 SD, $p < 0.25$; (B) Visna retrovirus, 2.4 SD, $p < 0.01$; (C) Polio type 1 Mahoney, 1.3 SD, $p < 0.2$; (D) Equine infectious anemia virus, 2.9 SD, $p < 0.003$; (E) Coxsackie B5, 4.25 SD, $p < 0.0001$.

Since only one such structure is required in an mRNA for selenocysteine insertion, the existence of multiple potential SECIS elements in HIV-1 suggests that the RNA structure of the viral genome may be optimized to an exceptional degree for the efficient recruitment of selenocysteine tRNAs.

It is also of interest that the preferred cellular tRNA that is packaged in HIV-1 virions and used as a primer for RT is $tRNA^{Lys}$. In *Escherichia coli*, when Se is abundant, up to 50% of the pool of $tRNA^{Lys}$ molecules may contain the modified base 5-[(methylamino)-methyl]-2-selenouridine as the wobble base.[25] As is the case for selenocysteine, the Se in this unusual tRNA comes from the active reduced Se product of the *selD* protein.[15] Evidence for the existence of a similar modified $tRNA^{Lys}$ has been demonstrated in a mouse leukemia cell line.[69] If this is also true of human cells, it may be more than coincidental that HIV prefers to package a tRNA that can potentially contain Se.

We have also found potential SECIS-like elements in a number of other retroviruses, and in eukaryotic retrotransposons (e.g., Copia), suggesting that Se biochemistry may have been used by retroelements throughout evolution and indeed may be a hallmark of retroviruses. This observation also extends to a number of other RNA viruses like polio and coxsackie viruses (Figure 8) and even to hepatitis B virus, which encodes a reverse transcriptase, and thus is a member of the retroid family.

Termination Suppression as a General Mechanism in Retroviruses. As outlined in the introduction, it must be emphasized that termination suppression *independent* of selenocysteine insertion at UGA codons is a well-documented retroviral strategy for translation of certain gene products in relatively low amounts (as recently reviewed[5,26]).

As an alternative to frameshifting, termination suppression of an amber stop codon (UAG) is essential for *pol* gene expression in several type C retroviruses, including murine leukemia virus.[3,4] There are several primate retroviral sequences in GenBank that have an in-frame stop codon in the middle of the *pol* gene, even when *gag* is in a different reading frame; these viruses could not replicate unless termination suppression is routine.

The presence of a high density of multiple in-frame stop codons (seen as thick black lines in Figure 1A) at the 3′ end of coding regions that are under the direction of strong initiation codons (e.g., the *gag, nef, rev*, and *env* reading frames in HIV-1) suggests that termination by a single stop codon may be relatively inefficient, explaining why readthrough of a lone stop codon may actually be an effective strategy for the expression of gene products that are only required in small amounts.

This process does not even have to depend upon unusual suppressor tRNAs, since, for example, acceptance of a single less-than-ideal but widely permitted G–U base pair in a codon–anticodon interaction would permit a glutamine tRNA with a UUG anticodon to pair with an ochre stop codon (UAA), permitting termination suppression with Gln insertion.

It has been shown that mammalian cells contain tRNAs capable of suppressing all three stop codons, that this process does not depend upon retroviral modification of cellular tRNAs, and that in appropriate sequence contexts, Gln can be inserted at UAA or UAG, and Arg, Cys, or Trp at UGA.[20,21] Recently, Feng *et al.*,[10] demonstrated that in the well-studied murine leukemia virus *gag–pol* readthrough, a downstream pseudoknot is involved. This suggests that in this case, at least, there may be distinct parallels between the mechanisms of termination suppression and frameshifting.

Consistent with these observations of Feng *et al.*, we have found pseudoknots (Figure 12) immediately downstream from non-UGA stop codons at the 3′ ends of several of the novel potential fusion proteins reported here. These are in locations where protein sequence analysis and other considerations suggest that termination suppression may occur (see below).

Finally it must be noted that in bacteria, the *selB*-mediated selenocysteine incorportion system that is analogous to the SECIS system in eukaryotes has been shown to permit incorporation of selenocysteine not only at UGA, but also at UAA and UAG codons when these are placed by mutagenesis into the same RNA structural context required for insertion at UGA.[27] Whether the presence of SECIS elements in eukaryotic or viral mRNAs permits the same latitude in some cases remains to be established.

Given the existence of these multiple mechanisms for termination suppression, both with and without selenocysteine insertion, and the certainty that retroviruses exploit at least some of them,[3–5,26] clearly one must keep an open mind about precisely where a potential retroviral gene product may actually terminate.

A Conserved Potential DNA-Binding Selenoprotein Can Be Expressed as a Protease −1 Frameshift Fusion Protein. Realizing that the UGA codons in the −1 reading frames downstream from the pseudoknots could be sense codons for selenocysteine, we decoded the sequence for a potential fusion protein (Table 1), beginning as a *gag–pol* product but switching to the *env* reading frame (−1 to *pol* in HIV-1) shortly before the protease pseudoknot (Figure 1B). A potential slippery sequence (A AAG GAA in the *pol* reading frame) was found immediately upstream from the base of the 5′ stem of the pseudoknot; this not an ideal heptameric slippery sequence, but it is as close to the ideal as the MMTV *pro–pol* heptameric sequence mentioned previously.

Although detailed mechanistic studies of *gag–pol* frameshifting have suggested slippage while two tRNAs are bound on the ribosome,[9] we believe another possible scenario here is a frameshift based on slippage while a single lysine tRNA is bound, exploiting the A AAG portion of the slippery sequence listed above. This is because the A AAG GAA sequence is very close to the base of the pseudoknot, which might induce a frameshift by recoil off the ribosome before a glutamate tRNA could bind to the GAA codon.

It has also been proposed that the role of pseudoknots in frameshifting may be more of an active one than has been previously thought,[28] suggesting that the mechanism may not be extremely dependent upon optimally slippery sequences in all cases. Consistent with the nonideal heptameric sequence, this frameshift event may be deliberately programmed to be less efficient than that at *gag–pol*, particularly if the gene product is involved in negative feedback of gene expression, as its similarity to various DNA binding proteins (see below) and the E2 papillomavirus DNA binding domain suggests (Figure 9).

Searches against the entire PIR database using the complete 69 residue sequence of this hypothetical protease −1 frameshift fusion protein (up to the first non-UGA stop codon) as a probe produced the most significant hits on HIV and related retroviruses, due to the identity of the first 20 residues with HIV-1 protease. Among the next highest ranked hits are a nonstructural protein of unknown function from ainovirus (a Simbu serogroup bunyavirus; >8 SD), and a number of other viral proteins, a *drosophila* gene suppressor and several transforming proteins, and the bovine papillomavirus E2 DNA binding protein, all at >5 SD. Numerous other DNA-binding proteins of various types, viral early proteins, and transcription factors are highly ranked (>4 SD).

Since the retroviral protease is known to cleave *gag–pol* gene products at a conserved protease cleavage site at the N-terminal of the protease domain, after a moderate amount of protease had accumulated in the host cell, the primary product of this *gag–pol*-frameshift fusion protein would very likely be a product with approximately the first 20 of its N-terminal residues identical to protease, and the rest of its sequence corresponding to a 49 amino acid domain containing two selenocysteine residues, the first of which is in a position

Table 1. Potential Fusion Proteins in HIV-1[a]

NO.	FUSION PROTEIN N-TERMINAL	PREDICTED PRIMARY STRUCTURE
1	PROTEASE	PQITLWQRPL VTIKIGGQL$\overset{*}{K}$ GSSIRYRSR$\underline{C}$ YSIRRNEFAR KMETKNDRGN WRFYQSKTV$\underline{C}$ SDTHRNLWT$\hat{Q}$ SYRYSISRTY TCQHNWKKSV DSDWLHFKFS H
2	RT	PISPIETVPV KLKPGMDGPK VKQWPLTEEK IKALVEICTE MEKEGKISKI GPENPYNTPV FAIKKKDSTK WRKLVDFREL NKRTQDFWEV QLGIPHPAGL KKKKSVTVLD VGDAYFSVPL DEDFRKYTAF TIPSINNETP GIRYQYNVLP QGWKGSPAIF QSSMTKILEP FR$\overset{*}{K}$TKSRHSY LSIHG$\underline{C}$FVCR I$\underline{C}$LRNRAA$\hat{Q}$N KNRGAETTSV EVGTYHTRQK TSERTSIPLD GL$\underline{C}$TPS$\underline{C}$
3	INTEGRASE	FLDGIDKAQD EHEKYHSNWR AMASDFNLPP VVAKEIVASC DKCQLKGEAM HGQVDCSPGI WQLDCTHLEG $\overset{*}{K}$SYPGSSSCS QWIYRSRSYS SRNRAGNSIL SFKISRKMAS KNNTYRQWQQ FHQYYG$\hat{Q}$GRL LVGGNQAGIW NSLQSPKSRS SRIYE
4	ENVELOPE	*RRVVQRE*$\overset{*}{K}$KS SGNR$\underline{SFVPWV}$ $\underline{L}$GSSRKHYGR TVNDADGTGQ TIIVWYSAAA EQFAEGY$\underline{C}$GA TASVATHSLG HQAAPGKNPG CGKIPKGSTA PGDLGLLWKT HLHHCCALEC
5	NEF	*PEYFKNC*$\tilde{\underline{C}}$HR ACYKGLSAGD FPGRRGLGGT GEWRALRCCI

[a] The predicted frameshift sites, all at lysine codons (K), are indicated by asterisks. Sequences at the frameshift sites have been translated according to a hypothetical single lysine tRNA slippage mechanism (see text). The sequence after the asterisk is encoded in the −1 reading frame; the sequence up to the asterisk is identical to the known protein sequence in the zero reading frame. The underlined "$\underline{C}$"s are potential selenocysteines, encoded by UGA codons. The ^ symbol indicates sites of possible termination suppression, consistent with the presence of potential pseudoknots 3′ to the termination codons (Figure 12). In the envelope fusion protein, the underlined region ($\underline{SFVPWVL}$) is a possible protease cleavage site. The *nef* fusion protein can be formed without frameshifting by a readthrough of the terminal UGA codon, indicated by $\tilde{\underline{C}}$. For *env* and *nef*, the entire fusion protein sequences are not shown; only the C-terminal ends of the "known" protein sequences immediately preceding the frameshift or readthrough sites are shown (italicized).

that is highly conserved in the primate retroviruses (Figure 10).

This first selenocysteine falls precisely at a point which can be aligned with the highly conserved cysteine of the papillomavirus E2 DNA binding domain,[29] which is in the center of the DNA recognition helix. This cysteine is at the heart of the interaction of this DNA binding protein with its DNA target, which have recently been cocrystallized.[30] The hypothetical HIV-1 DNA binding protein conforms well to a multiple alignment of various E2 protein DNA-binding domain sequences (Figure 9) and contains the most essential residues in the correct positions, assuming selenocysteine can substitute for cysteine. The essential tryptophan that is critical for dimerization of this domain[29,30] is correctly placed and is conserved in HIV-1 variants and in chimpanzee immunodeficiency virus (Figure 9).

Consistent with the requirements observed by Feng *et al.*,[10] the possibility for termination suppression of a UAA codon at the end of the 69 residue region discussed above is supported by the existence of another potential pseudoknot (Figure 12A) immediately 3′ to the UAA codon. With this readthrough suppression, this gene product could be extended another 32 amino acids (Table 1: lower panel of Figure 9), making it very close to the exact length of the aligned E2 DNA binding domains. While the most essential domains for DNA binding and the critical dimerization residues are contained within the upper panel of the alignment shown in Figure 9, it seems likely that maximal DNA binding activity would not be obtained without readthrough of this UAA codon. This is reasonable if the molecule functions primarily as a repressor, since it would be another check against overexpression, which could be undersirable for a potent transcriptional regulator. Addition of this readthrough extension region also increases the significance scores of alignments to several other DNA binding proteins, for which pairwise sequence alignments at significance levels of 4 to 5 SD can be produced using the entire sequence of the hypothetical HIV-1 protease fusion protein (data not shown). This, combined with the results of the database

```
                            *                             *
HIV-1     TIKIGGQLKGSSIRYRSRCYSIRRNEFARKMETKNDRGNWRFYQSKTVCS.DTHRN.LWT
DPV       CPCLLGTISGNG..NQVKCYSFRVKRWHDRDKY.HHTTTWWAVGGQGSERPGDATV.IVT
BPV-1     SCFA..LISGTA..NQVKCYRFRVKKNHRHRYE.NCTTTWFTVADNGAERQGQAQI.LIT
CRPV      PPVI..CLKGGH..NQLKCLRYRLKSKHSSLFD.CISTTWSWVDTTSTCRLGSGRM.LIK
HPV-8     PPVI..LVRGGA..NTLKCFRNRARVRYRGLFK.YFSTTWSWVAGDSTERLGRSRM.LIL
HPV-1     PPVV..CVKGGA..NQLKCLRYRLKASTQVDFD.SISTTWHWTDRKNTERIGSARM.LVK
HPV-11    TPIV..QLQGDS..NCLKCFRYRLNDKYKHLFE.LASSTWHWASPEAP.....HKNAIVT
HPV-6     TPIV..QFQGES..NCLKCFRYRLNRDHRHLFD.LISSTWHWASSKAP.....HKHAIVT
HPV-18    TPII..HLKGDR..NSLKCLRYRLRKHSDH.YR.DISSTWHWTGAGN......EKTGILT
HPV-16    TPIV..HLKGDA..NTLKCLRYRFKK.HCTLYT.AVSSTWHWTGHNYK.....HKSAIVT
HPV-33    APIV..HLKGES..NSLKCLRYRLKP.YNELYS.SMSSTWHWTSDNKN.....SKNGIVT
                   <--Recog.Helix-->


                         ^
HIV                 .....QSYRYSISGTYTCNHNWKKSVDSDWLHFKFSH
DPV                 ..FKDQSQRSHFLQQVPLPPGMSAHGVTMTVDF
BPV-1               ..FGSPSQRQDFLKHVPLPPGMNISGFTASLDF
CRPV                ..FADSEQRDKFLSRVPLPSTTQVFLGNFYGL
HPV-8               ..FTSAGQREKPDETVKYPKGVDTSYGNLDSL
HPV-1               ..FIDEAQREKFLERVALPRSVSVFLGQFNGS
HPV-11              LTYSSEEQRQQFLNSVKIPPTIRHKVGFMSLHLL
HPV-6               VTYDSEEQRQQFLDVVKIPPTISHKLGFMSLHLL
HPV-18              VTYHSETQRTKFLNTVAIPDSVQILVGYMTMY
HPV-16              LTYDSEWQRDQFLSQVKIPKTITVSTGFMSI
HPV-33              VTFVTGQQQQMFLGTVKIPPTVQISTGFMTLV
```

Figure 9. Multiple alignment of the protease −1 frameshift fusion protein (shown complete except for the first eleven residues at the N-terminal, PQITLWQRPLV, which are identical to HIV-1 protease) with a set of papillomavirus E2 protein DNA binding domain sequences. Note similarities (boldfaced) in the DNA recognition helix region, with selenocysteine (shown by C plus an asterisk) encoded by UGA aligning with the conserved Cys (underlined) of the E2 proteins. The conserved tryptophan (W, underlined) is important for dimerization and is conserved in HIV-1 variants, CIV and some other primate retroviruses. The most critical DNA binding domains are in the upper panel. Readthrough suppression of a UAA codon by glutamine insertion, at the first Q in the lower panel ($\hat{Q}$), aligned with boldfaced glutamates (E), could extend the protein to the length shown (see Figure 12); this region also conforms to the E2 alignment (e.g., conserved R near the beginning of the lower panel).

```
                  * *
SIVMAND        SSTRYRSCCYHL
SIVAGM3        SIIRYGGRCYHY
SIVAGM         VLLGYRGCCFYC
SIVSOOTY       SIIRYRGCRFNC
SIVMAC         SIIGYRGCCFYC
HIV2ROD        SLVRHRGCRLNS
HIV1BRU        SSIRYRSRCYSI
HIVYU          SSIRYRSRCYSI
HIV2           SFTRHRGCRLNS
RESIVXX        GTARHRGRCHHN
SIVACUTE       SIIRYRGCCFNC
SIV1AGM        SLVRYRSRCHYN
SIVGAA         SIIRYRGCCFNC
SIVSYKES       NVSRYRGRCYYN
HIVMAL         SSIRHRSRCYSI
CIV            SFARYRSCCYSN
```

Figure 10. Multiple alignment of the putative protease fusion protein DNA recognition helix from a series of primate retroviruses, showing conservation of UGA selenocysteine codons (shown as C̲) and other residues. In some of the SIV sequences, two UGA codons are found; all sequences with only one C̲ have arginine (R) in the other position. Conservation of sequence (e.g., the R residues indicated by an asterisk) is improbable unless this is a real gene, because, e.g., the conserved T of the protease DTG consensus sequence should permit any base in its third codon position, but only A or C at that point will yield an R in the −1 reading frame. Similar arguments apply to the conservation of a UGA codon in the −1 reading frame (see text).

searches summarized above, and the fitting of the HIV sequence into the E2 protein multiple alignment, is compelling evidence in favor of the hypothesis that this gene encodes a DNA binding protein.

As detailed in the legend to Figure 10, in the putative DNA-recognition helix region conservation of sequence in the −1 reading frame suggests that the probability of this *not* being a real gene is vanishingly small (also see Figure 11).

It is also perhaps significant that both retroviral proteases and E2-type DNA binding proteins function as dimers. The arrangement of these genes in HIV-1 suggests that perhaps a common dimerization motif (the protease N-terminal region) is being attached to two different functional domains (protease and DNA binding activities) encoded by the same nucleic acid sequence in two different reading frames.

An E2-like Dyad-Symmetric Sequence at the very 5′ End of the HIV-1 Transcribed Region. The papillomavirus E2 DNA binding proteins bind as dimers to a region in DNA with partial dyad symmetry, with the general consensus sequence $\underline{\mathrm{ACCG}}\mathrm{NNNN}\underline{\mathrm{CGGT}}$, where N can be any base.[29,30] In keeping with common practice in the E2 literature, we will call this partial inverted repeat a palindrome. The underlined bases are those that are invariantly conserved in E2 recognition sites for all papillomaviruses. Consistent with the suggested similarity between the potential HIV DNA binding protein and E2 proteins, we have found a thematically similar but distinct palindromic sequence in the HIV-1 LTR, having the sequence TGGTTA-GACCA. Here, the symmetry is in TGGT pairing with ACCA, with a three-base spacer, as compared to ACCG with CGGT and a four-base spacer in the E2 sequence. In the HIV-1 palindrome, the GG pair is at the 5′ end, which is the reverse of the E2 sequence. Thus the sequences are similar in general design but quite distinct in their details. The HIV-1 E2-like palindrome is located immediately downstream of the TATA box and proposed initiator protein binding site, at the +9 to +19 positions in the LTR, which is a perfect position to block an initiation complex. This region is also contained within the binding site spanned by the so-called leader binding protein, LBP-1, which is a cellular repressor of HIV expression that inhibits binding of the TATA

	* SeC
HIV1BRU	aucuaucaauacauggaUgau
HIVJSRF	aucuaucaauacauggaUgau
HIVYU	aucuaucaguacauggaUgau
HIVNL43	aucuaucaauacauggaUgau
HIVU455A	aucuaucaauacauggaUgac
HIVNY5CG	aucuaucaauacauggaUgau
COPIA	guauuauuauauguagaUgau
HIVMAL	auauaccaauacauggaUgau
HIVCAM1	aucuaucaauacauggaUgau
HIVMN	aucuaucaauacauggaUgau
HIVOYI	aucuaucaauacauggaUgau
HIV2	aucguucaguacauggaUgau
HIV2ROD	aucauucaguacauggaUgau
HIV2GH1	cucauccaauacauggaUgau
HTLV	auucuucaauacauggaUgac
CIVCG	auuuaccaguacauggaUgac
SIVAGM1	uuaguccaguauauggaUgac
SIVAGM2	auuguccaauacauggaCgau
SIVAGM3	auugugcaauacauggaUgac
SIVAGM4	auuguacaauacauggaUgau
SIVMAND	uuauaucaauacauggaUgau
SIVSOOTY	cugauccaauacauggaUgac
SIVMAC	uuaguccaguauauggaUgac
SIVACUTE	cugauccaauacauggaUgac
SIVSYKES	cuaauacaguacauggaUgac
SIVGAA	cugauccaauacauggaUgac
FIV	auuuaccaauauauggaUgac
PANTHER	auauaucaauauauggaUgau
PUMA	guauaucaauauauggaUgau
BIV	uuguaucaauauauggaUgau
EIAVCG	uuguaucaauauauggaUgau
MMLV	cugcuacaguacguggaUgac
VISNA	uuuggaauauacauggaUgau
	Y M D D

Figure 11. Conservation of the UGA codon in the −1 reading frame of the RT YMDD region in primate and other closely related retroviruses. Due to the degeneracy of the genetic code, this conservation cannot be explained by evolutionary selection of the *pol* gene protein sequence alone, since either GAU or GAC codes for Asp (D). The U of UGA is probably being conserved due to a combined requirement for the UGA codon in the −1 reading frame, assisted by conservation of secondary structure in some specific cases (see text). In the one case where the UGA is not conserved (SIVagm2), there is evidence of a compensatory mutation: a UGA codon, unique to SIVagm2, is found in the −1 reading frame slightly downstream from the region shown.

binding protein.[31] Possible monomer binding versions of this sequence, ACCA or TGGT, are found in several locations in the LTR (around −305, −320, −360, and −400); this is considered to be a potential negative regulatory region of the HIV-1 enhancer.

Its position in the repeat region of the LTR also places an RNA copy of this palindromic sequence in the TAR element[23,32] at the very 5′ end of the HIV-1 RNA, as well as near the 3′ end. Consistent with the established pattern that critical or conserved sequences can be associated with helical RNA structures,[16,17] this E2-like palindromic sequence in the HIV-1 LTR falls precisely on the stem of the TAR element in the viral RNA (Figure 5), for which we have found an alternative conformation that is a potential pseudoknot (Figure 6), which can display an unpaired UGA codon on its 3′ loop. This might serve as a binding site for the anticodon loops of selenocysteine tRNAs interacting with an unmistakable SECIS stem−loop structure just downstream from the pseudoknot (Figure 6), since there is no other nearby potentially unpaired UGA codon. Since the Tat protein[23] is known to bind precisely to a bulge on the stem of the TAR element where this UGA codon is located, it is possible that Tat may inhibit the formation of this pseudoknot, preventing the SECIS from functioning (Figure 6). Since it has been suggested that there may be protein factors that recognize and stabilize RNA pseudoknots, e.g., during frameshifting,[28] it is possible that there could be competition between Tat and any such protein factors. There is precedent for pseudoknots being involved in switching between conformational states as a regulatory mechanism.[19]

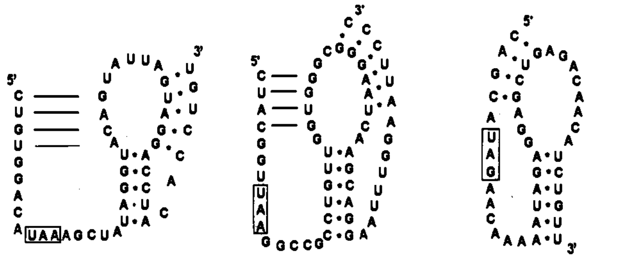

Figure 12. Potential pseudoknots associated with non-UGA stop codons (boxed) where termination suppression is suspected (see text), consistent with the demonstrated involvement of pseudoknots in retroviral readthrough suppression.[5,10] Similar to what was reported for the type C retroviruses, all three of these pseudoknots have a purine-rich consensus sequence (Pu-Pu-C-N-Pu) immediately following the stop codon. The possibility of a third stem region is indicated for A and B by lines connecting complementary bases; presumably these could not be formed at the same time as the 3′ stem shown, so a conformational change could be involved. (A) At the 3′ end of the protease fusion protein, following the ochre stop codon translated as Q in the lower panel of Figure 9. (B) At the 3′ end of the integrase fusion protein, following an ochre stop codon translated as Q and underlined in Figure 14. (C) At the 3′ end of the RT fusion protein, involving an amber stop codon. This pseudoknot lacks the extreme 3′ stem seen in the two pseudoknots associated with ochre codons.

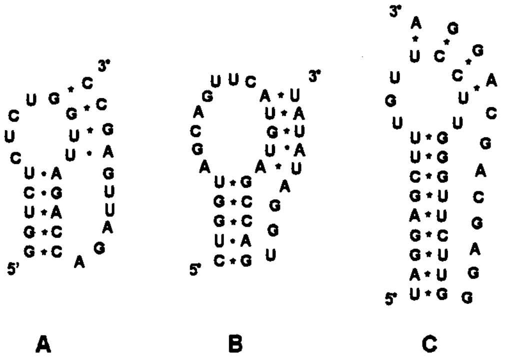

Figure 13. Additional potential pseudoknots in HIV-1: (A) Details of the pseudoknot shown schematically in Figure 6, containing the E2-like palindromic sequence (UGGUUA-GACCA), in the repeat region of the LTR. (B) In the integrase region of *pol*, potentially directing integrase fusion protein synthesis. (C) In the *env* coding region, potentially directing synthesis of a gp120 fusion protein.

Given the apparent association between codons for highly conserved sequence motifs and pseudoknots (Figures 2 and 4), it may be significant that this HIV-1 E2-like sequence also falls precisely on a potential pseudoknot (Figure 13A).

Finally, although among viruses this palindromic sequence is quite unique to HIV and its close relatives (where it is usually not precisely identical, and is found in varied locations), it is also found in noncoding (intron) regions of several human cytokine-related genes, including the tumor necrosis factor (TNF) receptor gene. This suggests that if this indeed is a recognition sequence for a virally encoded repressor, HIV may be attempting to resist simulation by cytokines like TNF, as well as modulate its own expression. These observations will be reported in detail in a separate publication.

Although the existence of this E2-like palindromic sequence in the TAR region of the HIV-1 LTR is indisputable, it remains to be determined if this is a DNA binding site for a virally encoded protein, or a cellular factor, or indeed if it has any biological significance at all.

A Conserved Potential Selenoprotein Can Be Expressed as a Reverse Transcriptase −1 Frameshift Fusion Protein. Like the pseudoknot in the protease region, the potential pseudoknot at the RT active site coding region (Figure 3) also has the potential to direct a frameshift and the synthesis of a fusion protein. The most probable slippery sequence is A AAA CAA in the *pol* reading frame, which is about 15 bases before the 5′ stem of the pseudoknot. Again, this is not an ideal heptamer, but it is a candidate for the single lysine tRNA slippage mechanism proposed previously. The fusion protein (Table 1) is by necessity a structural analog of RT, identical up to the end of the first major helix of the RT palm domain (helix E),[33] but with the active site YMDD β-hairpin region replaced by a selenocysteine containing module.

As shown in Figure 11, this selenocysteine codon is highly conserved in primate retroviruses, which strongly suggests that this must be a real gene. The U of the UGA codon is the third position of an Asp codon in the *pol* reading frame, which, particularly given the high mutation rate of retroviruses, should be a mixture of GAC and GAU (the two possibilities for Asp in the genetic code), *if* there was no evolutionary selection for U instead of C in the third position. By comparison, the third position of the immediately following second Asp codon is an even mixture of U and C (in the last column of the alignment). Clearly, what is being conserved in the first case is the UGA codon in the −1 reading frame, and possibly some structural feature in the RNA.

If the only criterion for evolutionary selection was a requirement for conservation of the Asp of the RT sequence in the *pol* reading frame, assuming a 50% chance of finding either U or C at this position, the probability of the distribution observed in Figure 11 arising by chance (U in 33 of 34 sequences) is less than 10^{-8}. Thus, there must be selection based either upon a stringent requirement for RNA structure, or a requirement for conservation of the sequence in the −1 reading frame (we can rule out the +1 reading frame since it has a high density of stop codons in this region). In fact, these two factors are interrelated, since a protein sequence in the −1 reading frame could not be conserved unless an RNA structure were present to facilitate the frameshift necessary for the expression of that protein.

The key point here is that, although both a UGA codon and an RNA structure are probably being conserved, the codon is more highly conserved than the structure. As discussed previously, the details of the stem−loop or pseudoknot structure found in this location for various retroviruses are highly variable, with those in Figure 4 being only a closely related sample. When a large set of sequences are compared (Figure 11), the A base that pairs with the U of UGA in the HIV-1 RT pseudoknot, which is the first base in the alignment (indicated by an asterisk), is seen to be not as well-conserved as the U; this is consistent with the previously mentioned diversity of RNA structures associated with this region of the RT gene. The observation that bases in other positions may pair with the U in various other retroviruses shows that it is the RNA structure that is shifting to accommodate the unchanging UGA codon, and not vice versa. In some structures, e.g., HTLV-2 and BIV (not shown), the U of UGA pairs with a G; in such cases, the stability of the structure would actually be enhanced by a transition of the U to a C, and yet that is almost never observed. In the one case where a CGA *is* observed rather than a UGA (in SIVagm2, Figure 11), there appears to be a compensatory mutation: slightly downstream from the CGA, an AGA codon seen in other SIVagm variants has mutated to a UGA codon in SIVagm2, potentially encoding a SeC residue eight residues past the CGA codon.

The situation is much clearer in regard to the conserved UGA codons in the protease fusion protein (Figure 10) since in this case the UGA is not in a stem in a majority of the structures we have examined. It is on the 3′ loop of the HIV-1 and CIV protease pseudoknots (Figure 2). Thus, one cannot invoke conservation of RNA structure to explain the conserved UGA codons in the case of the protease fusion protein.

On the basis of these considerations, combined with the conservation of the ORFs associated with these genes, we believe that the conservation of the UGA codons strongly suggests that these are real genes, and that the UGA codons are critical for the encoded proteins.

Database searches using only the −1 frameshift portion of the potential RT fusion protein did not yield an unequivocal picture, as there was no outstanding match to a unique class of proteins. However, there were a number of highly ranked matches to different actual and hypothetical viral proteins, including the most significant match (cytomegalovirus hypothetical UL63 protein, >6 SD), and a protein from vacciniavirus (4.5 SD) that appears to be a thymidine kinase (TK), based on its sequence similarity to other viral kinases. There were matches to several other types of kinases at >3.5 SD and to other ATP utilizing enzymes. Using the entire potential RT fusion protein as a probe, and eliminating hits on RT and viral *pol* gene sequences (obtained since the N-terminal is identical to HIV-1 RT), several viral TKs were ranked in the top 15 hits (out of >64 000 protein sequences; 4.4 SD).

As with the potential DNA binding protein, extension of the product by readthrough suppression, in this case of a UAG stop codon, is suggested by the existence of a potential pseudoknot that was found in the immediate region of the stop codon (Figure 12C). Addition of this region also gave improved scores (relative to the terminated fusion protein) in pairwise alignments with several TK sequences (data not shown).

The absence of a distinct ATP binding motif (e.g., of the GXXGXGK type seen in some viral TKs) in the RT fusion protein argues against a potential role as a kinase. However, it must be remembered that polymerases like RT also lack such motifs but nonetheless bind nucleotide triphosphates; this fusion protein contains the entire N-terminal half of RT, which is known to contain the dNTP binding region around lysine-73.[34]

```
 10 DKAQDEHEKYHSNWRAMASDFNLPPVVAKEIVASCDKCQLKGEAMHGQVD 59
    ||||||.:.  ...|.  |    | .||.: ||| . ...|| .   : |
402 dkaqderdwvlnefrtgks....pimvatd.vasrg.idvkgithvfnyd 445

 60 CSPGIWQLDCTHLEGKSYPGSSSCSQWIYRSRSYSSRNRAGNSILSF... 106
      ||  . |..|  |:: .:....:...| :.. ....|.  ||||
446 f.pgnte.dyvhrigrtgragakgtaytyftsdnakqarelvsilseakq 493

107 KISRKM........ASKNNTYRQWQQFHQYYGQGRLLVG.GNQA.GIWNS 146
    .|..|:        :.:...|| ...: .  |. |  .| ||.. |. .|
494 didpkleemaryssggrggnyr.rggygr..ggfrrgggygnrnrgftgs 540

147 LQSPKSRSSR 156
      .| .||..
541 nsaplarsrw 550
```

Figure 14. Alignment between the putative integrase fusion protein (upper sequence) and the C-terminal end of an *E. coli* RNA helicase (lower sequence), produced using the GAP program with the default normalized Dayhoff matrix;[89] significance, 3.7 SD. Gap weight is 1.5; length weight, 0.3; 26% identity. A very similar alignment (not shown) was produced using a PAM 120 matrix, which gave a significance score of 4.2 SD. The GKS sequence in the second panel, aligned with GRT in the helicase, is a known ATP binding motif.[38]

Although HIV is not a DNA virus, its dependence on DNA synthesis for proviral production by RT may justify the need for a TK, particularly when infecting quiescent cells where nucleotide pools may be decreased. However, this protein could equally well be some other enzyme involved in nucleotide metabolism or even an unusual nuclease or polymerase type of enzyme. Since many of the known selenoproteins are involved in redox type reactions, one possibility that would be reasonable is a ribonucleotide reductase, which is found in some viruses; however, no matches to reductases were observed at >3 SD in these searches.

One notable and intriguing alternative match that did come up in the database search was to glutathione synthase (4.4 SD), which has one cysteine rich region (**GNGVCRICGR**) with a similarity to the selenocysteine-rich region of the fusion protein (**GCFVCRICLR**, where **C** is selenocysteine). The frameshift fusion domain also has similarities (up to 4.5 SD in database searches) to various versions of CD44, the lymphocyte homing receptor. Since it is structurally a variant of reverse transcriptase, we think it improbable that this fusion protein could function as a surface receptor. Nonetheless, the observed sequence similarities between its frameshift fusion domain and CD44 could be significant in terms of proposed autoimmune aspects of AIDS.[35,36]

An Integrase Fusion Protein Potentially Expressed by a Pseudoknot-Directed −1 Frameshift from *pol*. A systematic search of the HIV-1 genome for other novel genes that could be expressed by frameshifting revealed a potential 64 residue ORF, well-conserved in primate retroviruses, in the −1 reading frame of the integrase coding region of *pol* (Figure 1B). As expected, a potential pseudoknot was found near the beginning of this region (Figure 13B). A possible single tRNA slippery sequence based on a lysine codon (A AAA) is appropriately placed. The potential fusion protein (Table 1) consists of the entire N-terminal integrase putative zinc finger domain connected to a module in the −1 reading frame that may be extended by readthrough of a UAA stop codon. This possibility is strongly supported by the presence of another pseudoknot about seven bases downstream of the UAA codon (Figure 12B). This is the only one of the potential new genes that does not contain UGA codons.

This fusion protein contains a GKS sequence, which is a known ATP binding motif found in helicases and other ATP-binding enzymes.[37,38] The protein also contains a DE sequence,[37] a motif which is often found associated with the GKS motif in ATP binding enzymes (see Figure 14). Thus, this fusion protein may have the potential to bind ATP.

Database searches with the translated peptide sequence from the −1 reading frame demonstrated significant similarities to certain viral proteins (e.g., hepatitis delta antigen, >5 SD), various kinases and ATPases (>4 SD), which generally showed the greatest similarity to a region just downstream of the GKS motif. Various proteins that interact with DNA and RNA, including several histones (>5 SD) and helicases (3.8 SD) were also highly ranked in the search.

Pairwise alignments of at least borderline significance (>3 SD) can be made between the sequence of this putative fusion protein and helicases of various types, e.g., from coxsackie B virus. One of these, to the C-terminal domain of an RNA helicase from fission yeast, is shown as Figure 14, which shows that the sequence similarity extends through the UAA stop codon (translated here as Gln[21]), extending the fusion protein by an additional 29 amino acids, ending almost precisely at the C-terminal of the yeast helicase. This is significant at 3.7 SD ($P < 0.001$) relative to alignments of randomized versions of the same sequences.

Since a number of viruses encode ATP-dependent helicases, e.g., picornaviruses, herpesviruses and papillomaviruses, it would not be unprecedented to find a virally encoded helicase in HIV. Since HIV-1 integrase is known to multimerize, this enzyme, whatever its function, may work in association with integrase, since their first 75 residues are identical. Alternatively, it may utilize the potential zinc finger module of integrase, which is a DNA targeting device, to target the molecule encoded in the −1 reading frame, which is likely to use ATP or GTP, whatever its function. The most probable function of this fusion protein is as a helicase that might work in tandem with integrase, or possibly play a role in transcription.

This example illustrates the elegance of this frameshift mechanism as a modular approach for encoding different proteins that share a common domain. The only difference in principle between this and RNA splicing of alternate 3′ exons is that the alternate 3′ coding domains are in different reading frames of the same piece of RNA.

A >100 Residue Potential Selenoprotein Domain Can Be Expressed by a Pseudoknot-Directed −1 Frameshift from *env*. The search for potential novel genes that could be expressed by frameshifting also revealed a 125 residue ORF, containing one UGA codon near the middle, in the −1 reading frame in the *env* coding region (Figure 1C); a major portion of this is well-conserved in primate retroviruses (see Figure 1 in ref 39). As in the case of the integrase frameshift site, a potential pseudoknot was found near the beginning of this region (Figure 13C). A slippery sequence, again involving lysine codons (A AAA AGA in the *env* reading frame) is appropriately placed (again, this is not an ideal heptameric slippery sequence, but the proposed single tRNA mechanism could apply). A potential fusion protein beginning in *env* would consist of almost all of gp120 fused to a 103 amino acid domain potentially containing a selenocysteine (Table 1).

Before the pseudoknot was discovered, which revealed the site of a potential frameshift, protein database searches were undertaken with the entire 125 residue fragment. The results were slightly ambiguous, with the most outstanding similarities to various small nuclear ribonucleoprotein particle (snRNP) proteins, which comprised 8 of the top 13 hits (>4 SD) out of the entire PIR database. Various ribosomal and other nucleic acid binding proteins, such as transcription factors, nucleocapsid, and homeotic proteins, were also highly ranked. However, potentially significant matches to several surface antigens were also found (>3.5 SD), which seems inconsistent with nucleic acid binding character, but nonetheless reasonable if the gene product is a modified version of the envelope glycoprotein gp120.

The latter possibility can probably be discounted. In this case, there may be no functional relationship between the protein modules encoded in the different reading frames of the fusion protein. Immediately after the frameshift site, in the −1 reading frame, there is a region with potential as an HIV-1 protease cleavage site, RSFVPWLG. In a set of six aligned primate retroviral sequences for this hypothetical protein module (not shown), which has a number of well-conserved regions, the sequence of ARVLG or ARVPR is conserved in five of the six sequences at this point. This is very similar to the HIV-1 protease cleavage site at the C-terminal of the capsid protein, which is ARVLA. If there is a protease cleavage here, attachment during synthesis to gp120 would not be functionally significant. In this case, based on the database search results, the possibility that the C-terminal protease cleavage product could a virally encoded RNA splicing factor may be worthy of consideration.

A role as a potential splicing factor would be consistent with the known action of the HIV-1 Rev protein, which facilitates the delivery of mature, full-length transcripts to the cytoplasm, and thus is critical for the late expression of structural genes and virion assembly.[1] Recent work has shown that Rev binding begins on a bulge in an RNA structure in the *env* coding region called the Rev-responsive element (RRE), followed by the cooperative binding of multiple Rev protein molecules along the base stem of the RRE.[40] The potential pseudoknot that we have found in the *env* region (Figure 13C), which would be necessary for the expression of this novel gene, is an alternative conformation for the proposed base stem (stem 1) of the RRE, involving the bases at the 5′ end just up to the point where a large circular structure begins, from which the various Rev stem−loop structures project (see Figure 4 in ref 1). Because cooperative Rev protein binding apparently stabilizes that stem, it could inhibit formation of the pseudoknot and thus prevent the synthesis of the hypothetical splicing protein. This would be a reasonable mechanism to explain the action of Rev, because by inhibiting the splicing of HIV-1 RNA, more full-length transcripts would be produced. This concept of a pseudoknot as one of several conformational states of a region of RNA and the potential role of a conformational change between those states as a regulatory mechanism has been previously proposed for several other systems where pseudoknots are believed to be involved.[19]

snRNPs and snRNP proteins have been described as "molecular fossils" of the RNP world,[22] a description that has been applied to viruses as well (a point which will be discussed further in the concluding section); thus, it would be very exciting to find them associated. Certainly, RNA splicing and the control thereof plays a critical role in the life cycle of HIV.[1] Because this hypothetical protein may also contain selenocysteine, which is believed to have played a more prominent role early in evolution, it could be more representative of the earliest splicing-associated proteins than those now found in higher organisms. However, at this point this is only a preliminary guess for the possible function of this potential gene product; it may turn out to have some other function entirely. We also noted almost equally significant sequence similarities to several transcription factors; however, these are generally much larger proteins than the *env* frameshift product. Whatever the function of its gene product, conservation of this reading frame and certain regions of its sequence in primate retroviruses strongly suggest it is a real gene.

HIV-1 *nef* Terminates in a UGA Codon: Its Gene Product Can Be Extended by 33 Residues following Selenocysteine Incorporation. The *nef* gene product in retroviruses, named as a putative negative factor, has been the subject of considerable controversy, conflicting hypotheses, and contradictory experimental results (reviewed in ref 39). Since the *nef* ORF ends in a UGA codon, readthrough of which could extend the protein by 33 amino acids, a possible explanation for this controversy is simply that artificial constructs of *nef* terminated at the UGA codon may not give a fully active molecule, potentially leading to incomplete or inaccurate assessments of its biological activity. Studies that have focused on modifications of the intact retrovirus, or deletions of the entire *nef* gene, may be more definitive. It is interesting that, as if in anticipation of possible truncation due to a shortage of selenocysteine tRNAs, the unextended gene ends in a cysteine residue, certainly the most acceptable alternative.

One point that should be made is that termination at a UGA codon, as seen in HIV-1 *nef*, is fairly atypical for primate retroviral *nef* genes. For example, in the less pathogenic HIV-2, *nef* does not terminate at a UGA codon. Thus, if selenocysteine incorporation can occur in these retroviruses, selenocysteine utilization is likely to be more pronounced in HIV-1 than in HIV-2, particularly since a frameshift is not required for the expression of this potential selenoprotein.

Selenium as a Factor in AIDS, Cancer, and Viral Pathology. If HIV utilizes selenocysteine in any of its essential genes, one would expect to find evidence of a possible involvement of Se in some aspect of the clinical picture of AIDS. Evidence *for* such involvement would be consistent with the possible existence of viral selenoproteins but would not prove their existence. Conversely, the lack of any connection between AIDS and Se would suggest that the theory is probably wrong, i.e., the virus does not encode selenoproteins.

The Evidence: A MedLine search revealed significant documentation of a characteristic decline in plasma Se levels in ARC and AIDS patients (refs 41–49 and references therein), which has been widely assumed to be solely a consequence of malabsorption, because gastrointestinal function is often impaired in AIDS patients.

Se is known to be an essential micronutrient, of particular importance in maintaining glutathione-dependent antioxidant status, because glutathione peroxidase is a selenoenzyme, essential and abundant in the liver and in blood cells.[50] In addition, the type I 5′-DI enzyme involved in the synthesis of thyroid T3 hormone from T4 is a selenoenzyme. There is extensive literature on thyroid depression in AIDS patients, some of which specifically points out depression in T3 relative to T4.[51–55] Hypothyroidism is typically characterized by metabolic depression, and low T3 levels have been linked to weight loss in AIDS patients.[54] Recent anecdotal reports of the successful treatment of some AIDS patients in Mexico with human growth hormone, possibly combined with nutritional therapy, may also be relevant, as it has been demonstrated that growth hormone elevates T3 by induction of Type I 5′-DI,[56,57] one of the key human selenoenzymes.

In one study, red blood cell glutathione peroxidase levels in AIDS patients (who also had reduced Se levels) were found to average 55% of normal controls ($p < 0.0001$).[48] The impairment of glutathione function in AIDS patients[58] is so unmistakable that in recent years several research groups have developed strategies to boost glutathione levels and/or antioxidant status in AIDS patients, including direct administration of glutathione[59] and *N*-acetylcysteine, which increases glutathione synthesis.[60] It has now been demonstrated that even in uninfected cells, oxidative stress or depletion of glutathione significantly impairs the ability of lymphocytes to respond to cytokines and antigenic stimuli.[61]

In at least two brief (50–70 day) clinical trials,[43,46] Se supplementation was reported to lead to symptomatic improvements in AIDS patients. A more extensive clinical trial of Se alone (as sodium selenite), in patients not on AZT, is currently in progress in Germany.

There is a substantial body of literature on the antiviral and anticancer effects of Se, going back a number of years (documented as of 10 years ago in ref 62). There is extensive evidence that Se has anticancer and antiproliferative effects[62,63] and antiviral effects both *in vivo* (e.g., in inhibiting retrovirally induced mammary tumors,[64] in protecting against cardiac damage induced by coxsackie B virus,[65] and in chemoprevention of human hepatitis[66]) and *in vitro*, where addition of Se to culture media has been shown to inhibit the pathogenicity of several RNA viruses, including oncogenic retroviruses[67] and influenza.[68] Se compounds have also been shown to inhibit the growth of cultured human mammary tumor cells.[70] A 5-year clinical trial of Se supplementation in China, for liver cancer linked to hepatitis B infection, led to a significant decline in liver cancer in the treatment group.[71] These and other studies have been reviewed.[47,50]

Finally, a direct link between selenocysteine and viral disease is suggested by a report that identified an autoantibody in an unusual type of autoimmune hepatitis as an antibody to a tRNA^{SeC}/protein complex; the protein was suggested to be the human equivalent to the bacterial *selB* protein.[72]

Possible Interpretations. On the basis of the above data, in addition to the progressive immune deficiency and consequent opportunistic infections characteristic of AIDS, it seems probable that some of the associated pathology arises from thyroid T3 depression, impairment of glutathione function, and their consequences. These consequences include hypothyroid-like symptoms, increased susceptibility to oxidative damage, and impairment of responsiveness and membrane function in lymphocytes, which further contribute to immune deficiency.

The Se depletion potentially underlying these aspects of AIDS pathology could be entirely due to malabsorption, the conventional explanation. Whatever its origins, the implications of this depletion are sufficiently obvious that Se supplementation in AIDS is being advocated to prevent it. For example, Schrauzer and Sacher state that "supplemental selenium could be considered a *sine qua non* of the practical management of HIV infected subjects".[47] Obviously, they did not need to invoke the existence of viral selenoproteins to reach such a conclusion.

Alternatively, in the light of our results, it is possible that the symptoms of progressive Se depletion in AIDS are not only due to malabsorption, but also due to sequestration of Se in viral proteins, particularly within infected cells. Viral utilization of selenocysteine inside infected cells might impair the synthesis of cellular selenoproteins such as glutathione peroxidase to a greater extent than might be expected from measurements of extracellular (plasma) Se, because of virus–host competition for selenocysteine inside infected cells.

Although recent work has suggested that, even in asymptomatic patients, the total viral load may be much higher than previously thought (e.g., due to lymph node infection), it seems unlikely that formation of viral selenoproteins alone could completely account for the observed decreases in plasma Se; thus impaired absorption likely accounts for a significant portion, if not all, of the decrease. Nonetheless, by either mechanism, available Se levels in infected cells will be decreased relative to normal, leading to an impairment of antioxidant defenses.

Since we have found potential SECIS elements not only in HIV-1 but also in other retroviruses and in picornaviruses (e.g., Figure 8), given the evolutionary relationships between many viral genes, it is possible that there may be a common underlying mechanism for some of the observed anticancer and antiviral effects of Se, and reason to hope that HIV-1 may respond similarly. However, there are conventional explanations for these effects (e.g., involving known antioxidant effects of Se). Thus, at best we can say that these observations are consistent with, but do not prove, the kind of direct virus–Se link that we are proposing.

Discussion and Conclusions

Another essential role of Se in mammals is in the formation of sperm, where it is involved in formation of the mitochondrial capsule, which contains a selenoprotein.[73] Apparently, even during conditions of Se deficiency, Se is preferentially routed to the testes in the male.[74] This is interesting in the light of the role of semen as a primary vehicle for HIV transmission, and also because it potentially provides a new insight into the exogenous vs endogenous aspect of the retroviral life cycle. It is well-known that retroviruses can become endogenous in the germline, essentially dormant in the DNA, and be passed down for generations, or potentially even millions of years, without ever replicating through their RNA life cycle.[75,76] During this time, they are essentially in a time capsule, until some stimulus triggers their expression and/or exogenous replication, at which time their mutation rate is greatly accelerated. If retroviruses are programmed to utilize and seek out Se, they would find a rich source in sperm and thus be presented the opportunity for entering the germline. While this germline integration may be a very rare event, it has certainly happened many times over the course of mammalian evolution. Recent work has suggested that some endogenous retroviruses may have a significant (dare we say beneficial?) role in some cellular processes (such as T-cell repertoire selection) and that certain autoimmune diseases can be caused by abnormalities in the expression of endogenous retroviral genes.[77]

It must be noted that a potentially critical role for Se biochemistry in viruses is consistent with the view that we and others have been attempting to advocate in recent years that at least RT,[78-80] and possibly the entire *pol* gene of retroelements (Taylor *et al.*, paper in preparation), may be as archaic as the dawn of the DNA world. The ability of retroviruses to go into the time capsule of endogeny explains how some of them, perhaps including the ancestors of HIV, may have had less time to adapt to the modern world than some of their relatives. Se biochemistry is known to have been more prominent in the low oxygen atmosphere of the early earth several billion years ago.[15] Thus the discovery of selenoproteins in viruses, which have been characterized as "molecular fossils", is not entirely surprising:

"...some RNA viruses may be molecular fossils of the RNA or RNP world... retroviruses (whose RNA genomes replicate through a DNA intermediate) and hepadnaviruses (whose DNA genomes replicate through an RNA intermediate) may be fossils of a world in transition from RNA to DNA genomes".[22]

This "molecular fossil" view of viruses is not necessarily inconsistent with the modern view that retroviruses are escaped cellular transposable elements, if indeed the core genes of the latter have been in eukaryotes from the beginning, as postulated above.

Consistent with the "ancient origins" view of the polymerase-associated genes of viruses, we have located SECIS-like structures in some archaic retroelements like Copia, which, in an ORF −1 to its protease gene, has sequence similarities to the HIV-1 protease fusion protein. We have also found potential pseudoknots and frameshift fusion selenoproteins in the protease and replicase regions of nonattenuated polioviruses. These preliminary findings suggest that the evolutionary roots of these potential selenoprotein genes are deep.

An intriguing question is what benefit might the virus gain by using Se in some of its enzymes, and what is the nature of the chemical role that it plays? The fact that selenocysteine residues appear to be highly conserved in at least two of these potential new genes suggests that its role may be quite important. Interactions between Se and oxygen appear to be typical of Se biochemistry, and the effectiveness of selenoenzymes is dependent on the oxidative stress within the cells where they are located. Recently, several redox-sensitive transcriptional regulators have been described that are capable of inducing genes for various protective enzymes. These use the sulfur of cysteine residues in different ways to achieve their effects, which depend on the production of an oxidized molecular state to turn on their target genes. It is interesting that E2 proteins are dimers; perhaps under oxidative conditions, the two cysteines can covalently join the monomers via a disulfide bridge, which would probably eliminate their DNA binding ability. Alternatively, production of an oxidized form of sulfur (or Se in the case of selenocysteine) may modify DNA binding ability, as is the case for the OxyR control of the catalase gene,[81] where oxidation of thiols to sulfenic acid leads to activation. Such mechanisms would provide yet another level of transcriptional control, enabling the virus to respond directly to oxidative stress within the cell. This should be a promising area for further experimentation.

Regarding possible homology between the putative HIV-1 DNA binding protein and papillomavirus E2 proteins, there is some independent circumstantial evidence of a functional and possible evolutionary relationship between the transcriptional regulatory mechanisms of the two types of viruses. HIV-1 Tat has been shown to enhance E2-dependent gene expression in model systems;[82] these studies were undertaken because of a known epidemiological association between HIV-1 infection and human papillomavirus (HPV) associated anogenital disease. This association suggests that there may have been opportunities for recombination and the exchange of genes between HIV and HPV or their ancestors at some time in the past.

The papillomavirus E2 proteins function as transcriptional activators when attached to a large N-terminal domain, but as repressors of viral transcription when expressed as the DNA binding domain alone.[29] In early stages of cellular HIV infection, the very low levels of viral protease may permit the accumulation of an N-terminal extended form of the putative DNA binding protein (i.e., the unprocessed fusion protein), which could have all or part of *gag*, and the poorly understood N-terminal region of *pol*, attached to the DNA-binding domain. By analogy to the E2 proteins, this unprocessed fusion protein could be a transcriptional activator,[29] consistent with what would be required at that stage of infection. However, because the HIV protease processed form of the potential HIV-1 DNA binding protein only aligns with the DNA binding domain, by analogy to the E2 proteins there is reason to believe that this gene product would function as a repressor in any late stage of viral expression. Moreover, much as the E2 proteins use cysteine, the hypothetical HIV protein apparently contains an obligate selenocysteine, conserved throughout the primate retroviruses (Figure 10). Under conditions of selenocysteine depletion, this UGA codon could more frequently function as a stop codon and thus lead to a truncated repressor that could not bind DNA, because the first selenocysteine residue is in the center of the DNA recognition helix (Figure 9). This suggests the hypothesis that the ultimate depletion of Se within an infected cell could be a signal for unrestrained expression of viral genes and thus be a trigger for release from latency. If the virus is dependent on Se for some aspects of its biochemistry, it is logical that it should be programmed to move out in search of a new host if Se is exhausted, particularly since the host cell could be close to dying at that point, due to oxidative stress.

This hypothesis is consistent with the clinical picture of HIV latency and the progressive decline in Se levels in AIDS patients. If it is correct, it would imply that nutritional status may be a factor in some HIV-positive long-term survivors. It is perhaps significant that Haitians and intravenous drug users, two of the original three high-risk populations for HIV infection, are notoriously impoverished and probably malnourished in many instances. It is also interesting that nitrites are powerful oxidizing agents that can antagonize Se and other antioxidants like glutathione: various alkyl nitrites were widely used as sexual stimulants by male homosexuals, particularly in the early years of the epidemic. The use of nitrite inhalants has been correlated with more rapid seroconversion in HIV-infected men[83] and with the incidence of Kaposi's sarcoma.[84] The hypothesis for HIV latency proposed above would definitely predict that prolonged exposure to nitrites and other oxidants should accelerate the course of the disease. Nonetheless, it is also true that recent experimental work has shown that oxidant stress impairs lymphocyte function even in the absence of HIV infection.[61]

A link between Se and AIDS, with progressive depletion of Se in ARC and AIDS patients, and depression of thyroid T3 and glutathione/antioxidant function, which directly depend on Se availability, is well-documented, although that depletion can be explained without invoking the existence of viral selenoproteins. The potential clinical benefits of Se supplementation in AIDS patients[43,46,49] certainly need to be further explored. It should be noted that Se supplementation is already practiced by a segment of AIDS and ARC patients, especially those that have joined self-help groups utilizing holistic or natural healing methods as additive therapies. Se doses for long-term supplementation that are recommended in the patient literature are generally in the safe nutritional range, typically corresponding to 200 μg/day. Since signs of chronic Se toxicity begin to appear after the prolonged intake of 1000 μg of Se/day, any application of Se at doses above those commonly used for nutritional supplementation should not be attempted except under close medical supervision.

On the basis of a number of arguments and lines of evidence, Schrauzer and Sacher have recently proposed that it would probably be better to begin a nutritional Se supplementation immediately in HIV positive patients, rather than waiting until they have developed AIDS symptoms.[47] Our results potentially provide an independent rationale supporting their conclusions, although we must emphasize that the molecular mechanisms we have proposed are entirely theoretical at present, and there are alternative explanations for the observed antiviral effects of Se.

In conclusion, we have identified a number of potential new genes in HIV-1 and related retroviruses, demonstrated the existence of potential RNA structures that would be required for their expression, and provided a number of arguments why there is a high probability that at least several of these are real genes. Our results suggest the possible existence of up to four selenoproteins in HIV-1, which is remarkable as only five have been characterized in humans. We have attempted to identify the probable nature of these genes by comparative sequence analysis and have some degree of confidence regarding the assignments in several cases; in other cases we can only make suggestions, based in part upon the types of genes that are known to occur in other viruses. These possibilities are offered primarily as potential starting points for future experimental investigations.

It is also possible that in some cases, or in some retroviruses but not others, some of these may be inactive or vestigial genes, or genes that are fading in or out of use. Obviously, the existence and actual functions of these gene products will have to be verified experimentally before many of these questions can be resolved.

If even one of these potential new genes proves to be real, it will open up entirely new approaches to anti-HIV therapy. If one of them indeed codes for an activator/repressor protein, i.e., a gene product that can potentially turn off HIV expression, it and its binding site in the proviral DNA will be particularly promising as targets for therapeutic intervention.

Because we have also found evidence for SECIS-like elements in the RNA of other viruses, some of the considerations discussed above relating to potential mechanisms of Se involvement in viral pathology may be very broadly applicable. Perhaps most significantly, the approach we have used to discover these new genes in HIV is quite general, and we are very confident that it will be possible to discover new genes in other viruses, and perhaps even in higher organisms, by the same method.

Experimental Section

1. Comparative Sequence Analysis: (A) Database Searches. Hypothetical protein sequences of interest were used as probes to search the PIR 39 (Protein Identification Resources Data Bank) protein database using the FASTDB program (IntelliGenetics, Inc., Mountain View, CA). In both database searches and pairwise alignments, cysteine was used in place of selenocysteine in the probe sequences. The database searching algorithm used by FASTDB has been described previously.[85] The initial comparison of the query sequence to the database is performed in a manner similar to that used by the FASTP or FASTA programs, except that FASTDB uses the specified similarity matrix in the first pass of the search, rather than in the second pass. The structure–genetic matrix of McLachlan,[86] which measures a combination of chemical and genetic code similarities, was used as the similarity matrix for database searching. We have found that this combination of searching algorithm and scoring matrix has several benefits, including the finding of optimal alignments to the entire probe sequence, rather than just a small region of optimal local similarity. The values of the various parameters used in the present study were Mismatch penalty, 5; Gap penalty, 2.0; Gap size penalty, 0.26; Joining penalty, 20; K-tuple, 1; Threshold level, 83. The scores of the best match of the probe sequence against each of the target (database) sequences are averaged to give the database mean score and its standard deviation. Significance scores from the database searches are calculated as the difference from the mean divided by the standard deviation. The approach we have used in these database searches is simply to ask, out of the most significant matches, are there hits which represent types of proteins known to occur in viruses, or with biological activities similar to other known viral proteins? Alternatively, if particularly significant matches are observed to multiple examples of a given protein type, it may be worthy of investigation. One of the final criteria has to be: does it make any sense that this virus might encode such a protein?

(B) Pairwise Alignments. In some cases, the top matches from FASTDB searches were individually compared to the sequence of interest using the BESTFIT or GAP programs (as implemented in the GCG software package[90]) to generate pairwise alignments. BESTFIT makes an optimal alignment of the best segment of similarity between two sequences, using the local homology algorithm of Smith and Waterman.[87] GAP uses the algorithm of Needleman and Wunsch[88] to generate an optimal alignment of two complete sequences. Unless otherwise specified, the default similarity matrix used for the pairwise alignments was the normalized Dayhoff matrix of Gribskov.[89] The potential significance of matches was assessed by measuring the deviation of the quality score of the actual alignment from the average quality score calculated by running BESTFIT or GAP with randomized sequences of the same length and amino acid composition. The average quality score and its standard deviation, SD, were determined for 100 alignments of randomized sequences, which were amino acid sequences in this study. Reported percent identities were calculated as no. identities/(sequence length + no. insertions).

(C) Sequences Used. The HIV-1 BRU sequence, GenBank #K02013, also known as LAI, was used for RNA structure and pseudoknot predictions, analysis of ORFs (Figure 1), and translation of hypothetical protein sequences (Table 1). Figure 8 shows predicted structures from the picornaviruses coxsackie B5, #X67706, and polio type 1 Mahoney, #J02281. The following abbreviations have been used (Figures 8, 10, and 11) for other retroid sequences: VISNA, Visna lentivirus, #M10608; EIAVCG, Equine infectious anemia virus, #M16575; HIVYU, HIV-1, #M93259; HIVJSRF, HIV-1, #M38429; HIVNL43, HIV-1, #M19921; HIVU455A, HIV-1, #M62320; HIVNY5CG, HIV-1, #M38431; HIV1BRU, HIV-1, #K02013; HIVMAL, HIV-1, #K03456; HIVCAM1, HIV-1, #D10112; HIVMN, HIV-1, #M17449; HIVOYI, HIV-1, #M26727; HIV2GH1, HIV-2, #M30895; HIV2ROD, HIV-2, #M15390; HIV2, HIV-2, #A05350; HTLV, HTLV-1, #L02534; SIVAGM1, SIVagm, #Y00295; SIVAGM2, SIVagm, #M66437; SIVAGM3, SIVagm, #M30931; RESIVXX/SIVAGM4, SIVagm, #X07805; CIV, SIV chimpanzee, #X52154; SIVSYKES, SIV sykes, #L06042; SIVMAND, SIV mandrill; #X15781; SIVACUTE, SIV, #L09211; SIVGAA: SIV, #M80193; SIVMAC, SIV macaque, #M19499; SIVSOOTY, SIV sooty mangabey, #X14307; MMLV, Maloney murine leukemia virus, #J02255; Copia, Retrotransposon Copia, #S03612; Panther, Panther lentivirus, #M95476; FIV, Feline immunodeficiency virus, #M25381; BIV, Bovine immunodeficiency virus, #M32690; PUMA, Puma lentivirus, #U03982.

2. Prediction and Statistical Analysis of RNA Secondary Structure: (A) RNA Folding. The secondary structure of RNA regions was predicted using the FOLD program[18] with updated energy parameters as implemented in the GCG software package.[90] The significance of the stability of such structures can be assessed by measuring the deviation of the free energy of stabilization of the predicted structure from the average calculated for randomized structures of identical size and base composition.[24] In the present study, the average computed free energy and its standard deviation, SD, were determined from foldings of 30 randomized versions of the nucleotide sequence of interest. The randomized sequence was obtained using the GCG SHUFFLE program. A subroutine was developed which performs the shuffle and folding processes for "*n*" number of times, the value of "*n*" given by the user. The difference between the computed free energy of folding and the mean value calculated for the randomized sequences, expressed as the number of SD (z) below the average for the randomized sequences, is a useful metric for assessing the relative stability of the computed folding of the actual sequence relative to that attainable by similarly composed random sequences.[16,24]

(B) Pseudoknots and SECIS Structures. Programs like FOLD[18] and most other global RNA folding programs have not been programmed to predict pseudoknot structures, due in large part to the lack of a comprehensive set of experimental free energy parameters for their relative stability. In the present study, a systematic search for potential pseudoknots was undertaken in regions of interest by a semiautomated method. This involved the use of FOLD in a sliding window type of search for possible stem–loop structures with at least five base pairs in the stem, and loops that were no more than 20 unpaired bases in size. This was followed by a search, within regions no more than 20 bases upstream or downstream from the base of the first stem, for sequences of three or more bases that were the inverse complement to bases on the loop, no more than one base removed from the first stem. In the later stages of the study, analysis of potential open reading frames (Figure 1) was used to target the regions for the search.

Our search of the HIV-1 genome for potential pseudoknots is still in progress, and to date has focused primarily on the *pol* gene region, except for the connection domain and ribonuclease H regions, which are still in progress. We have searched less than a third of the *gag* and *env* coding regions. We have not found "simple" pseudoknots such as those reported here associated with any other highly conserved sequence motif regions of RT or integrase, although we have observed a tendency for helical (stem) regions in more complex folded RNA structures to be associated with such motifs.[16] The pseudoknots reported here are all that we have found to date in HIV-1, except for one near the 3′ end of the *nef* coding region, which will be reported in a future publication. Thus, potential pseudoknots do not appear to be common or easily found, at least in this case.

Similarly, potential SECIS stems were found by a systematic sequence scanning method that initially focused on finding occurences of the 5′-DI SECIS loop motif (UAAAG), with UGA codons either upstream or downstream of the AAA motif. In some cases regions were rapidly eliminated, often by visual inspection of the sequence, due to lack of potential for a stem–loop in the AAA region, under the requirement that no more than one of the three A bases could be paired. Candidate regions were then folded, usually with a sliding window approach, in order to determine the most stable and significant RNA structures in the surrounding region. The thermodynamic stability and statistical significance of the most stable RNA structures were then determined as described in the previous section.

Acknowledgment. This work was supported by U.S. Public Health Service Grant AI-30392 from the National Institute of Allergy and Infectious Diseases (E.W.T.). We would also like to thank Dr. James S. Gibbs for his timely provision of the ORFwriter program, which was used in generating Figure 1, and Dr. Raymond Schinazi, without whose initial encouragement we might not have become involved in analyzing the RNA structure of HIV.

References

(1) Parslow, T. G. Post-transcriptional regulation of human retroviral gene expression. In *Human Retroviruses*; Cullen, B. R., Ed.; Oxford University Press: New York, 1993; pp 101–136.

(2) Jacks, T.; Power, M. D.; Masiarz, F. R.; Luciw, P. A.; Barr, P. J.; Varmus, H. E. Characterization of ribosomal frameshifting in HIV-1 gag-pol gene expression. *Nature* **1988**, *331*, 280–283.

(3) Yoshinaka, Y.; Katoh, I.; Copeland, T. D.; Oroszlan, S. Murine leukemia virus protease is encoded by the gag-pol gene and is synthesized through suppression of an amber termination codon. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 1618–1622.

(4) Feng, Y. X.; Hatfield, D. L.; Rein, A.; Levin, J. G. Translational readthrough of the murine leukemia virus gag gene amber codon does not require virus-induced alteration of tRNA. *J. Virol.* **1989**, *63*, 2405–2410.

(5) Rein, A.; Levin, J. G. Readthrough suppression in the mammalian type C retroviruses and what it has taught us. *New Biol.* **1992**, *4*, 283–289.

(6) Puglisi, J. D.; Wyatt, J. R.; Tinoco, I., Jr. A pseudoknot RNA oligonucleotide. *Nature* **1988**, *331*, 283–286.

(7) Le, S. Y.; Shapiro, B. A.; Chen, J. H.; Nussinov, R.; Maizel, J. V. RNA pseudoknots downstream of the frameshift sites of retroviruses. *Genet. Anal. Tech. Appl.* **1991**, *8*, 191–205.

(8) Jacks, T.; Townsley, K.; Varmus, H. E.; Majors, J. Two efficient ribosomal frameshifting events are required for synthesis of mouse mammary tumor virus gag-related polyproteins. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 4298–4302.

(9) Jacks, T.; Madhani, H. D.; Masiarz, F. R.; Varmus, H. E. Signals for ribosomal frameshifting in the Rouse Sarcoma Virus gag-pol region. *Cell* **1988**, *55*, 447–458.

(10) Feng, Y. X.; Yuan, H.; Rein, A.; Levin, J. G. Bipartite signal for read-through suppression in murine leukemia virus mRNA: an eight-nucleotide purine-rich sequence immediately downstream of the gag termination codon followed by an RNA pseudoknot. *J. Virol.* **1992**, *66*, 5127–32.

(11) Cohen, E. A.; Lu, Y.; Gottlinger, H.; Dehni, G.; Jalinoos, Y.; Sodroski, J. G.; Haseltine, W. A. The T open reading frame of human immunodeficiency virus type 1. *J. AIDS* **1990**, *3*, 601–608.

(12) Chamorro, M.; Parkin, N.; Varmus, H. E. An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 713–717.

(13) Shen, Q.; Chu, F. F.; Newburger, P. E. Sequences in the 3′-untranslated region of the human cellular glutathione peroxidase gene are necessary and sufficient for selenocysteine incorporation at the UGA codon. *J. Biol. Chem.* **1993**, *268*, 11463–11469.

(14) Berry, M. J.; Banu, L.; Harney, J. W.; Larsen, P. R. Functional characterization of the eukaryotic SECIS elements which direct selenocysteine insertion at UGA codons. *EMBO J.* **1993**, *12*, 3315–3322.

(15) Bock, A.; Forchhammer, K.; Heider, J.; Leinfelder, W.; Sawers, G.; Veprek, B.; Zinoni, F. Selenocysteine: the 21st amino acid. *Mol. Microbiol.* **1991**, *5*, 515–520.

(16) Schinazi, R. F.; Lloyd, R. M., Jr.; Ramanathan, C. S.; Taylor, E. W. Antiviral drug resistance mutations in human immunodeficiency virus type 1 reverse transcriptase occur in specific RNA structural regions. *Antimicrob. Agents Chemother.* **1994**, *38*, 268–274.

(17) Le, S. Y.; Chen, J. H.; Chatterjee, D.; Maizel, J. V. Sequence divergence and open regions of RNA secondary structures in the envelope regions of the 17 human immunodeficiency virus isolates. *Nucleic Acids Res.* **1989**, *17*, 3275–3288.

(18) Zuker, M.; Steigler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxillary information. *Nucleic Acids Res.* **1981**, *9*, 133–148.

(19) ten Dam, E.; Pleij, K.; Draper, D., Structural and functional aspects of RNA pseudoknots. *Biochemistry* **1992**, *31*, 11665–11676.

(20) Feng, Y. X.; Levin, J. G.; Hatfield, D. L.; Schaefer, T. S.; Gorelick, R. J.; Rein, A. Suppression of UAA and UGA termination codons in mutant murine leukemia viruses. *J. Virol.* **1989**, *63*, 2870–3.

(21) Feng, Y. X.; Copeland, T. D.; Oroszlan, S.; Rein, A.; Levin, J. G. Identification of amino acids inserted during suppression of UAA and UGA termination codons at the gag-pol junction of Moloney murine leukemia virus. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 8860–8863.

(22) Watson, J. D.; Hopkins, N. H.; Roberts, J. W.; Steitz, J. A.; Weiner, A. M. *Molecular biology of the gene*, 4th ed.; Benjamin/Cummings: Menlo Park, 1987.

(23) Peterlin, B. M.; Adams, M.; Alonso, A.; Baur, A.; Ghosh, S.; Lu, X.; Luo, Y. Tat *trans*-activator. In *Human Retroviruses*; Cullen, B. R., Ed.; Oxford University Press: New York, 1993; pp 75–100.

(24) Le, S. Y.; Maizel, J. V. A method for assessing the statistical significance of RNA folding. *J. Theor. Biol.* **1989**, *138*, 495–510.

(25) Wittwer, A. J.; Stadtman, T. C. Biosynthesis of 5-methylaminomethyl-2-selenouridine, a naturally occuring nucleoside in Escherichia coli tRNA. *Arch. Biochem. Biophys.* **1986**, *248*, 540–550.

(26) Hatfield, D. L.; Levin, J. G.; Rein, A.; Oroszlan, S. Translational suppression in retroviral gene expression. *Adv. Virus. Res.* **1992**, *41*, 193–239.

(27) Heider, J.; Baron, C.; Bock, A. Coding from a distance: dissection of the mRNA determinants required for the incorporation of selenocysteine into protein. *EMBO J.* **1992**, *11*, 3759–3766.

(28) Tang, C. K.; Draper, D. E. Unusual mRNA pseudoknot structure is recognized by a protein translational repressor. *Cell* **1989**, *57*, 531.

(29) Ham, J.; Dostatni, N.; Gauthier, J.-M.; Yaniv, M. The papillomavirus E2 protein: a factor with many talents. *Trends Biochem. Sci.* **1991**, *16*, 440–444.

(30) Hegde, R. S.; Grossman, S. R.; Laimins, L. A.; Sigler, P. B., Crystal structure at 1.7 A of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature* **1992**, *359*, 505–512.

(31) Kato, H.; Horikoshi, M.; Roeder, R. G. Repression of HIV-1 transcription by a cellular protein. *Science* **1991**, *251*, 1476–1479.

(32) Berkhout, B. Structural features in TAR RNA of human and simian immunodeficiency viruses: a phylogenetic analysis. *Nucleic Acids Res.* **1992**, *20*, 27–31.

(33) Jacobo-Molina, A.; Ding, J.; Nanni, R. G.; Clark, A. D., Jr.; Lu, X.; Tantillo, C.; Williams, R. L.; Kamer, G.; Ferris, A. L.; Clark, P.; Hizi, A.; Hughes, S. H.; Arnold, E. Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 Å resolution shows bent DNA. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6320–6324.

(34) Cheng, N.; Merrill, B. M.; Painter, G. R.; Frick, L. W.; Furman, P. A. Identification of the nucleotide binding site of HIV-1 reverse transcriptase using dTTP as a photoaffinity label. *Biochemistry* **1993**, *32*, 7630–7634.

(35) Kennedy, J. R. AIDS--an autoimmune model. *Med. Hypotheses* **1992**, *37*, 16–19.

(36) Katz, D. H. AIDS: primarily a viral or an autoimmune disease? *Aids Res. Hum. Retrovir.* **1993**, *9*, 489–493.

(37) Gorbalenya, A. E.; Koonin, E. V.; Donchenko, A. P.; Blinov, V. M. A novel superfamily of nucleoside triphosphate-binding motif containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination. *FEBS Lett.* **1988**, *235*, 16–24.

(38) Koonin, E. V. A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif. *J. Mol. Biol.* **1993**, *229*, 1165–1174.

(39) Gibbs, J. S.; Desrosiers, R. C. Auxiliary proteins of the primate immunodeficiency viruses. In *Human Retroviruses*; Cullen, B. R., Ed.; Oxford University Press: New York, 1993; pp 137–158.

(40) Malim, M.; Cullen, B. HIV-1 structural gene expression requires the binding of multiple Rev monomers to the viral RRE: implications for HIV-1 latency. *Cell.* **1991**, *65*, 241–248.

(41) Beck, K. W.; Schramel, P.; Hedl, A.; Jager, H.; Kaboth, W. Trace element concentrations in HIV infected patients. *Onkologie* **1989**, *3*, 43–47.

(42) Dworkin, B. M.; Antonecchia, P. P.; Smith, F.; Weiss, L.; Davidian, M.; Rubin, D.; Rosenthal, W. S., Reduced cardiac selenium content in the acquired immunodeficiency syndrome. *J. Parenter. Enteral. Nutr.* **1989**, *13*, 644–647.

(43) Olmsted, L.; Schrauzer, G. N.; Flores-Arce, M.; Dowd, J. Selenium supplementation of symptomatic human immunodeficiency virus infected patients. *Biol. Trace Elem. Res.* **1989**, *20*, 59–65.

(44) Beck, K. W.; Schramel, P.; Hedl, A.; Jaeger, H.; Kaboth, W. Serum trace element levels in HIV-infected subjects. *Biol. Trace Elem. Res.* **1990**, *25*, 89–96.

(45) Allavena, C.; Dousset, B.; May, T.; Amiel, C.; Nabet-Belleville, F.; Canton, P. Are zinc and selenium markers of worsening in HIV infected subjects? *Presse. Med.* **1991**, *20*, 1737.

(46) Cirelli, A.; Ciardi, M.; de-Simone, C.; Sorice, F.; Giordano, R.; Ciaralli, L.; Costantini, S., Serum selenium concentration and disease progress in patients with HIV infection. *Clin. Biochem.* **1991**, *24*, 211–214.

(47) Schrauzer, G. N.; Sacher, J. Selenium in the maintenance and therapy of HIV-infected patients. *Chem.-Biol. Interact.* **1994**, *91*, 199–205.

(48) Dworkin, B. M.; Rosenthal, W. S.; Wormser, G. P.; Weiss, L.; Nunez, M.; Joline, C.; Herp, A. Abnormalities of blood selenium and glutathione peroxidase activity in patients with acquired immunodeficiency syndrome and AIDS-related complex. *Biol. Trace Elem. Res.* **1988**, *15*, 167–177.

(49) Zazzo, J. F.; Lafont, A.; Darwiche, H.; Sayegh, F.; Camus, F.; Chappuis, P.; Chalas, J.; Benattar, C. Is non obstructive myocardiopathy (NOMC) in AIDS selenium-deficiency-related? In *Selenium in Medicine and Biology*; Neve, J., Fevier, A., Eds.; Walter de Gruyter: Berlin, 1989; pp 281–282.

(50) Flohe, L., The selenoprotein glutathione peroxidase. In *Glutathione: Chemical, Biochemical and Medical Aspects. Part A*; Dolphin, R. P. D., Avramovic, O., Eds.; Wiley-Interscience: New York, 1989; pp 643–732.

(51) LoPresti, J. S.; Fried, J. C.; Spencer, C. A.; Nicoloff, J. T. Unique alterations of thyroid hormone indices in the acquired immunodeficiency syndrome (AIDS). *Ann. Intern. Med.* **1989**, *110*, 970–975.

(52) Bourdoux, P. P.; De-Wit, S. A.; Servais, G. M.; Clumeck, N.; Bonnyns, M. A. Biochemical thyroid profile in patients infected with the human immunodeficiency virus. *Thyroid* **1991**, *1*, 147–149.

(53) Nduwayo, L.; Nsabiyumva, F.; Osorio-Salazar, C.; Lecomte, P.; Guilmot, J. L.; Renard, J. P. Endocrinological aspects of acquired immunodeficiency syndrome (AIDS). *Med. Trop. Mars* **1992**, *52*, 139–143.

(54) Grunfeld, C.; Pang, M.; Doerrler, W.; Jensen, P.; Shimizu, L.; Feingold, K. R.; Cavalieri, R. R. Indices of thyroid function and weight loss in human immunodeficiency virus infection and the acquired immunodeficiency syndrome. *Metabolism* **1993**, *42*, 1270–1276.

(55) Hommes, M. J.; Romijn, J. A.; Endert, E.; Adriaanse, R.; Brabant, G.; Eeftinck-Schattenkerk, J. K.; Wiersinga, W. M.; Sauerwein, H. P. Hypothyroid-like regulation of the pituitary-thyroid axis in stable human immunodeficiency virus infection. *Metabolism* **1993**, *42*, 556–561.

(56) Byamungu, N.; Mol, K.; Kuhn, E. R. Somatostatin increases plasma T3 concentrations in Tilapia nilotica in the presence of increased plasma T4 levels. *Gen. Comp. Endocrinol.* **1991**, *82*, 401–406.

(57) Geelhoed-Duijvestijn, P. H.; Roelfsema, F.; Schroder-van-der-Elst, J. P.; van-Doorn, J.; van-der-Heide, D., Effect of administration of growth hormone on plasma and intracellular levels of thyroxine and tri-iodothyronine in thyroidectomized thyroxine-treated rats. *J. Endocrinol.* **1992**, *133*, 45–49.

(58) Roederer, M.; Staal, F. J.; Anderson, M.; Rabin, R.; Raju, P. A.; Herzenberg, L. A.; Herzenberg, L. A., Disregulation of leukocyte glutathione in AIDS. *Ann. N. Y. Acad. Sci.* **1993**, *677*, 113–125.

(59) Holroyd, K. J.; Buhl, R.; Borok, Z.; Roum, J. H.; Bokser, A. D.; Grimes, G. J.; Czerski, D.; Cantin, A. M.; Crystal, R. G. Correction of glutathione deficiency in the lower respiratory tract of HIV seropositive individuals by glutathione aerosol treatment. *Thorax* **1993**, *48*, 985–989.

(60) Kalebic, T.; Kinter, A.; Poli, G.; Anderson, M. E.; Meister, A.; Fauci, A. S. Suppression of HIV expression in chronically infected monocyte cells by glutathione, glutathione ester, and N-acetyl cysteine. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 986–990.

(61) Staal, F. J. T.; Anderson, M. T.; Staal, G. E. J.; Herzenberg, L. A.; Gitler, C.; Herzenberg, L. A. Redox regulation of signal transduction: Tyrosine phosphorylation and calcium influx. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 3619–3622.

(62) *Selenium in biology and medicine*; Combs, G. F., Jr.; Spallholz, J. E., Levander, O. A., Oldfields, J. E., Eds.; Van Nostrand-Reinhold: New York, 1987; Vols. A and B.

(63) Schrauzer, G. N. Selenium. Mechanistic aspects of anticarcinogenic action. *Biol. Trace Elem. Res.* **1992**, *33*, 51–62.

(64) Schrauzer, G. N.; Molenaar, T.; Kuehn, K.; Waller, D. Effect of simulated American, Bulgarian, and Japanese human diets and of selenium supplementation on the incidence of virally induced mammary tumors in female mice. *Biol. Trace Elem. Res.* **1989**, *20*, 169–178.

(65) Ge, K.-Y.; Bai, J.; Deng, X.-J.; Wu, S.-Q.; Wang, S.-Q.; Xue, A.-N.; Su, C.-Q. The protective effect of selenium against viral myocarditis in mice. In *Selenium in biology and medicine*; Combs, G. F., Jr., Spallholz, J. E., Levander, O. A., Oldfields, J. E., Eds.; Van Nostrand-Reinhold: New York, 1987; Vol. B; pp 761–768.

(66) Yu, S. Y.; Li, W. G.; Zhu, Y. J.; Yu, W. P.; Hou, C. Chemoprevention trial of human hepatitis with selenium supplementation in China. *Biol. Trace Elem. Res.* **1989**, *20*, 15–22.

(67) Balansky, R. M.; Argirova, R. M. Sodium selenite inhibition of some oncogenic RNA viruses. *Experientia.* **1981**, *37*, 1194–1195.

(68) Lazymovam, Z. A.; Abdullaev, I. I.; Abdullaev, F. I.; Asadullaev, T. A. Inhibiting action of sodium selenite on influenza virus reproduction. *Voprosy Virusol.* **1986**, *31*, 236–238.

(69) Ching, P. S., Occurrence of selenium-containing tRNAs in mouse leukemia cells. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 3010–3013.

(70) Yan, L.; Yee, J. A.; Boylan, L. M.; Spallholz, J. E. Effect of selenium compounds and thiols on human mammary tumor cells. *Biol. Trace Elem. Res.* **1991**, *30*, 145–162.

(71) Yu, S. Y.; Zhu, Y. J.; Li, W. G.; Huang, Q. S.; Zhi-Huang, C.; Zhang, Q. N.; Hou, C. A preliminary report on the intervention trials of primary liver cancer in high-risk populations with nutritional supplementation of selenium in China. *Biol. Trace Elem. Res.* **1991**, *29*, 289–294.

(72) Gelpi, C.; Sontheimer, E. J.; Rodriguez-Sanchez, J. L. Autoantibodies against a serine tRNA-protein complex implicated in cotranslational selenocysteine insertion. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 9739–9743.

(73) Karimpour, I.; Cutler, M.; Shih, D.; Smith, J.; Kleene, K. C., Sequence of the gene encoding the mitochondrial capsule selenoprotein of mouse sperm: identification of three in-phase TGA selenocysteine codons. *DNA Cell Biol.* **1992**, *11*, 693–699.

(74) Bedwal, R. S.; Nair, N.; Sharma, M. P.; Mathur, R. S. Selenium—its biological perspectives. *Med. Hypotheses* **1993**, *41*, 150–159.

(75) Shih, A.; Coutavas, E. E.; Rush, M. G. Evolutionary implications of primate endogenous retroviruses. *Virology* **1991**, *182*, 495–502.

(76) Katz, R. A.; Skalka, A. M. Generation of diversity in retroviruses. *Annu. Rev. Genet.* **1990**, *24*, 409–445.

(77) Talal, N.; Flescher, E.; Dang, H. Are endogenous retroviruses involved in human autoimmune disease? *J. Autoimmun.* **1992**, *5*, 61–66.

(78) Taylor, E. W.; Jaakkola, J. A transposition of the reverse transcriptase gene reveals unexpected structural homology to *E. coli* DNA polymerase I. *Genetica* **1991**, *84*, 77–86.

(79) Lazcano, A.; Valverde, V.; Hernandez, G.; Gariglio, P.; Fox, G. E.; Oro, J. On the early emergence of reverse transcription: theoretical basis and experimental evidence. *J. Mol. Evol.* **1992**, *35*, 524–536.

(80) McClure, M. A. Evolutionary history of reverse transcriptase. In *Reverse Transcriptase*; Skalka, A. M., Goff, S. P., Eds.; Cold Spring Harbor Laboratory Press: New York, 1993; pp 425–444.

(81) Storz, G.; Tartaglia, L. A.; Ames, B. N. Transcriptional regulator of oxidative stress-inducible genes: Direct activation by oxidation. *Science* **1990**, *248*, 189–194.

(82) Vernon, S. D.; Hart, C. E.; Reeves, W. C.; Icenogle, J. P. The HIV-1 tat protein enhances E2-dependent human papillomavirus 16 transcription. *Virus Res.* **1993**, *27*, 133–145.

(83) Burcham, J. L.; Tindall, B.; Marmor, M.; Cooper, D. A.; Berry, G.; Penny, R. Incidence and risk factors for human immunodeficiency virus seroconversion in a cohort of Sydney homosexual men. *Med. J. Aust.* **1989**, *150*, 634–639.

(84) Soderberg, L. S.; Barnett, J. B. Inhaled isobutyl nitrite compromises T-dependent, but not T-independent, antibody production. *Int. J. Immunopharmacol.* **1993**, *15*, 821–827.

(85) Brutlag, D. L.; Dautricourt, J.-P.; Maulik, S.; Relph, J. Sensitive similarity searches of biological sequence databases. *Comput. Appl. Biosci.* **1990**, *6*, 237–245.

(86) McLachlan, A. D. Repeating sequences and gene duplication in proteins. *J. Mol. Biol.* **1972**, *64*, 417–437.

(87) Smith, T. F.; Waterman, M. S. Comparison of biosequences. *Adv. Appl. Math.* **1981**, *2*, 482–489.

(88) Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **1970**, *48*, 443–453.

(89) Gribskov, M.; Burgess, R. R. Sigma factors from E. coli, B. subtilis, phage SPO1 and phage T4 are homologous proteins. *Nucleic Acids Res.* **1986**, *14*, 6745–6763.

(90) Devereux, J.; Haeberli, P.; Smithies, O. A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **1984**, *12*, 387–395.

(91) ORFwriter is a Macintosh program by Dr. James S. Gibbs, Harvard Medical School (gibbs@husc.harvard.edu).