Genomic structures of viral agents in relation to the biosynthesis of selenoproteins

By: Ethan Will Taylor, Ram Gopal Nadimpalli, and Chandra Sekar Ramanathan

Taylor, E.W., Nadimpalli, R.G. & Ramanathan, C.S. Genomic structures of viral agents in relation to the biosynthesis of selenoproteins. *Biological Trace Element Research,* **56**, 63–91 (1997). https://doi.org/10.1007/BF02778984

This is a post-peer-review, pre-copyedit version of an article published in *Biological Trace Element Research*. The final authenticated version is available online at: http://dx.doi.org/10.1007/BF02778984.

*****© 1997 Humana Press Inc. Reprinted with permission. No further reproduction is authorized without written permission from Springer. This version of the document is not the version of record and is subject to Springer Nature Terms of Use. *****

Abstract:

The genomes of both bacteria and eukaryotic organisms are known to encode selenoproteins, using the UGA codon for selenocysteine (SeC), and a complex cotranslational mechanism for SeC incorporation into polypeptide chains, involving RNA stem-loop structures. These common features and similar codon usage strongly suggest that this is an ancient evolutionary development. However, the possibility that some viruses might also encode selenoproteins remained unexplored until recently. Based on an analysis of the genomic structure of the human immunodeficiency virus HIV-1, we demonstrated that several regions overlapping known HIV genes have the potential to encode selenoproteins (Taylor et al.*[31], J. Med. Chem.* **37**, 2637–2654 [1994]). This is provocative in the light of over-whelming evidence of a role for oxidative stress in AIDS pathogenesis, and the fact that a number of viral diseases have been linked to selenium (Se) deficiency, either in humans or by in vitro and animal studies. These include HIV-AIDS, hepatitis B linked to liver disease and cancer, Coxsackie virus B3, Keshan disease, and the mouse mammary tumor virus (MMTV), against which Se is a potent chemoprotective agent. There are also established biochemical mechanisms whereby extreme Se deficiency can induce a proclotting or hemorrhagic effect, suggesting that hemorrhagic fever viruses should also be examined for potential virally encoded selenoproteins. In addition to the RNA stem-loop structures required for SeC insertion at UGA codons, genomic structural features that may be required for selenoprotein synthesis can also include ribosomal frameshift sites and RNA pseudoknots if the potential selenoprotein module overlaps with another gene, which may prove to be the rule rather than the exception in viruses. One such pseudoknot that we predicted in HIV-1 has now been verified experimentally; a similar structure can be demonstrated in precisely the same location in the reverse transcriptase coding region of hepatitis B virus. Significant new findings reported here include the existence of highly distinctive glutathione peroxidase (GSH-Px)-related sequences in Coxsackie B viruses, new theoretical data related to a previously proposed potential selenoprotein gene overlapping the HIV protease coding region, and further evidence in support of a novel frameshift site in the HIV *nef* gene associated with a well-conserved UGA codon in the-1 reading frame.

Keywords: AIDS | genomic analysis | glutathione peroxidase | HIV | retrovirus | RNA viruses | selenium | selenoproteins

Article:

INTRODUCTION

An extensive body of evidence, documented in the current symposium proceedings, demonstrates that selenium (Se) deficiency has been linked to the incidence, disease progression, or virulence associated with a number of viral infections, either in humans or by in vitro and animal studies. Furthermore, Se has been shown to have significant chemoprotective effects against a number of human and animal viruses, including hepatitis B infection linked to liver disease and cancer *(1,2),* Coxsackie virus, Keshan disease *(3,4),* viral hemorrhagic fever *(5),* and the mouse mammary tumor virus (MMTV) *(6).* In the case of the acquired immune deficiency syndrome (AIDS) associated with infection by the human immunodeficiency virus (HIV), abnormalities in Se and other antioxidants have been consistently reported over the last decade *(7-25).* Furthermore, there is now extensive evidence of a role for oxidative stress in the pathogenesis of the viral infection *(20,23).* In the light of the chemoprotective effects of Se in the other viral diseases noted above, and the fact that Se has now been proven to be a potent inhibitor of HIV activation in vitro *(26),* definitive large-scale controlled studies of the potential benefits of Se supplementation in HIV+ populations are now strongly justified.

It is possible that the explanation for this inverse correlation between host Se/antioxidant levels and viral pathogenesis lies almost entirely within the realm of known biochemistry, i.e., that it can be largely explained in terms of our current understanding of the importance of antioxidants for proper immune cell functioning, the role of oxidant tone in regulating gene expression, and the demonstrated ability of oxidative stress to activate viral replication. Antioxidant deficiencies can undoubtedly weaken host defenses, creating a more permissive environment for viral replication, yielding a larger pool of mutant viruses, which in turn can favor the emergence of more successful and potentially more virulent viral strains *(27).* However, the cogency and appeal of such explanations of this Se/antioxidant effect should not inhibit us from considering another possible contributing factor in the virus-host-antioxidant equation: the possibility that some viruses may perturb (or mimic) host antioxidant defenses by directly incorporating Se into viral proteins.

In addition to the nonspecific substitution of selenomethionine for methionine that can potentially occur in all proteins when Se is bioavailable, the genomes of both prokaryotic and eukaryotic organisms are known to encode specific selenocysteine (SeC)-containing "selenoproteins." Both bacteria and eukaryotes use the UGA codon for SeC, and both utilize a complex cotranslational mechanism for SeC incorporation into polypeptide chains, involving RNA stem-loop structures *(28,29).* These common features and, most importantly, the similar codon usage, strongly suggest that the molecular biology of SeC incorporation into proteins ("selenobiology') must predate the divergence between bacteria and higher organisms, and is thus an ancient evolutionary development. Bock and coworkers have elaborated on this possibility *(28),* and proposed that on the early earth of several billion years ago (when atmospheric oxygen levels were still very low), selenobiology may have been more prevalent than it is today, and hence that the UGA codon may have originally served primarily as a sense codon for SeC. Under this scenario, UGA may have evolved into its current predominant function as a signal for the termination of protein chains (a "stop codon") primarily by default, owing to the progressive decrease in the bioavailability of SeC in more recent evolutionary times, associated with the emergence of more oxidative conditions in the biosphere.

Although this concept of primordial selenobiology is only a theory, it has significant implications when combined with another interesting hypothesis about RNA viruses and retroviruses, viz., the idea that some of the features or molecular machinery of such viruses may be relics or "molecular fossils" of the ancient RNA or (RNP) ribonucleoprotein world *(30).* In fact, it is now widely accepted that RNA-based life forms predominated early in evolution, prior to the switch to DNA as the primary repository of genetic information, a change that clearly must have required reverse transcriptase, the defining enzyme of retroviruses.

As discussed previously *(31),* if there is any merit to these two theories, then by combining them, one should expect at least that RNA viruses in particular might contain vestigial features consistent with an ancient selenoprotein coding potential, or in some cases, that selenobiology might still be utilized by certain viruses as a modern adaptation of an ancient biochemical strategy. Distinguishing between such a vestigial feature and an active gene might be quite difficult in some cases, for reasons given at the end of the next section.

Probably because of the very low abundance of Se, and the small number of selenoproteins so far identified (leading to the possibly incorrect impression that they are extremely rare), this possibility apparently remained unconsidered and unexplored until our proposal in 1994 of what in retrospect might be called the "viral selenoprotein theory," which arose initially from an analysis of the genomic structure of HIV-1 *(31).* We showed that several regions of the virus have the potential to encode selenoproteins, and that certain UGA codons in overlapping reading frames are well conserved in the many HIV-1 isolates that have been sequenced *(31,32).* Obviously, if HIV could make such proteins, there would have to be a reassessment of the mechanisms by which oxidative stress contributes to AIDS pathogenesis.

In this article, we will assess the current status and scope of the viral selenoprotein theory in light of the rather extensive body of data on the interrelationships between various viral diseases and dietary Se/antioxidant status, much of which is described in detail in the other articles from this symposium. However, before presenting data on specific viruses, such as Coxsackie B (CBV) and HIV-1, we will make some general comments in regard to the protein coding potential of viruses and attempt a brief explanation of some of the structural features that are important for the genomic analyses that will be presented.

THE POTENTIAL COMPLEXITY OF VIRAL GENOMES

Although a nonspecialist might think it remarkable that a new gene could be discovered or proposed in a virus that has been characterized at the sequence level for over a decade, that possibility is by no means improbable if one understands the basis for the potential complexity of viral (or indeed all) genomes. That basis is immediately evident if one considers:

1. The potential benefits to the virus of increasing its protein coding potential;
2. The triplet nature of the genetic code, which creates the potential for three overlapping reading frames--actually six if the complementary strand is also present in the cell as RNA; and
3. The various processes by which the simple linear translation of an "open" genomic region can be diverted to permit the translation of an otherwise inaccessible region (as described below).

Viruses in particular are under considerable evolutionary pressure to maximize the information content of their genomes, because they are very limited in regard to the size of the DNA or RNA that can be packaged in their virions. Thus, they have evolved mechanisms that enable them to maximize the amount of protein coding information within a length of oligonucleotide. For example, by placing overlapping genes in different reading frames of the same nucleotide sequence, they can easily increase their coding density. This and other complex possibilities for increasing protein coding potential are summarized in Fig. 1.

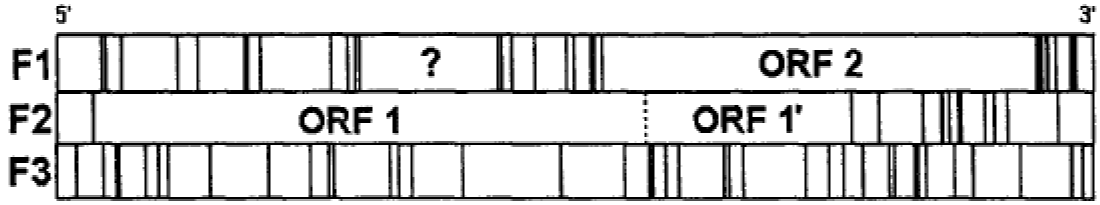


Fig. 1. Schematic diagram of PPCRs in an imaginary RNA transcript, showing the three possible reading frames for translation, F1-F3. Vertical lines are the locations of stop codons in each frame. Two major ORFs are shown as ORF1 and ORF2. Owing to its partial overlap with ORF1, even if it did not have a start codon, ORF2 could be expressed by various alternative mechanisms, including: (1) ribosomal frameshifting from ORF1, (2) RNA editing, or (3) RNA splicing, either of which could bring the ORF2 region in-frame to ORF1 in a subset of modified mRNAs. In all three cases, the protein product would be a fusion protein with its N-terminal derived from part of ORF1 and its C-terminal derived from part of ORF2. ORFI' could be expressed by suppression of the stop codon at the end of ORF1 (dashed vertical line), which would produce an extended form of the ORF1 protein. This analysis shows that there are various additional PPCRs in such a sequence, e.g., that indicated by the question mark, which, despite the lack of a start codon, could have potential to be expressed by alternative mechanisms if appropriate RNA editing, splicing, or frameshifting signals are present. Because specific structural features are involved in such events (e.g., heptameric or other frameshift sequences, RNA pseudoknots, splice acceptor and donor sequences, and so on), a detailed genomic analysis can sometimes reveal which PPCRs are the most likely to be functional, i.e., actually expressed as protein.

There are at least three different ways in which noncontiguous potential coding regions can be brought together to permit the translation of a continuous protein chain: RNA splicing, RNA editing, and frameshifting. Viruses that are transcribed in the nucleus, like retroviruses, can exploit RNA splicing, a characteristic feature of eukaryotic organisms. Alternative RNA splicing can be used to coordinate the expression of different genes at certain stages of infection, as seen in complex retroviruses; it can also potentially be exploited to permit the modular construction of different proteins containing a common module, thus avoiding the needless duplication of precious genetic material. Ribosomal frameshifting can potentially be used to achieve the same sort of modular construction, by placing alternate 3' "exons" in two overlapping reading frames of the same oligonucleotide; when expressed, each is attached to a common module encoded in the 5'-region of the zero reading frame. A third way of attaining the same goal is RNA editing, where an alternate transcript is made with one or more ribonucleotides physically inserted or deleted (usually via slippage of the polymerase on a repetitive sequence), leading to a subset of transcripts in which otherwise out-of-frame regions are brought in-frame. This mechanism enables the Ebola virus to make two different forms of its glycoprotein gene product *(33)*.

Another possibility is that of potential protein coding regions that are in the same frame, but separated by a single stop codon. Some viruses suppress certain stop codons to enable a downstream module to be expressed, usually in low amounts relative to the upstream module, owing to the inefficiency of the readthrough mechanism *(34)*. This process, commonly named after the type of stop codon being suppressed, as in "amber suppression," is also exploited by bacteria and eukaryotes. SeC insertion at UGA codons is conceptually a variant of this "termination suppression" mechanism, since UGA is normally a stop codon.

Because of these multiple mechanisms by which various overlapping or downstream gene regions can be brought in-frame to an upstream region containing a start codon, we find the current terminology of "open" reading frame (ORF) vs "blocked" reading frame to be inadequate, since to call a region an ORF is to imply that it has a start codon at its 5'-end. We will use the more general term "potential protein coding region" (PPCR), which could apply to any gene region where there is at least a theoretical possibility that a protein segment could be translated by any of the mechanisms discussed above.

Overlapping genes and processes like RNA splicing, editing, frameshifting, and suppression also present new opportunities for various forms of genetic regulation, beyond that attainable by transcriptional regulation alone. These processes potentially enable greater control over the expression of a greater variety of gene products than that attainable with a simple linear arrangement of nonoverlapping genes.

Since in many cases the structural features required for such processes are by no means immediately evident, even on a detailed examination of a viral sequence, one must keep an open mind about the possibility of additional coding potential for new genes or gene variants in viruses. Furthermore, because a single virus can infect different cell types and even different host species at different stages of its life cycle, it is also possible that some "specialized" viral genes may be inactive in certain cell types or cell lines, or even within a particular host species, yet be expressed to play an active role in another host organism or cell type. Such considerations imply that in some cases, it may be quite difficult to prove or disprove experimentally or immunologically either the existence or the possible role of a particular hypothetical gene product identified in a viral genome.

Nonetheless, one thing is very likely to be true: many viruses will prove to be more genetically complex than first expected. That has certainly proven to be the case with HIV and other "complex" retroviruses.

FRAMESHIFTING REQUIRES "SHIFTY" SEQUENCES AND RNA SECONDARY STRUCTURES

In retroviruses like HIV, the *pol* gene lacks a start codon, so the *pol* gene products can only be synthesized as *gag-pol* fusion proteins, most commonly by ribosomal frameshifting in the -1 direction *(35,36).* The inefficiency of this process ensures the synthesis of only small amounts of the viral enzymes encoded in *pol* relative to the structural proteins encoded in *gag,* which are required in large quantities. Retroviral *gag-pol* -1 frameshift sites have been extensively studied, leading to the "simultaneous slippage" model developed by Varmus and coworkers *(35).* The frameshift involves the translocation of the RNA transcript by one base while two tRNA molecules are bound to the ribosome, and occurs on a specific nucleotide sequence in the mRNA that has a repetitive, "slippery" quality. For the most efficient frameshifting to occur (which is still only about 5% of the time), 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 are known *(37).* It is also possible 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 regulatory protein might be required. Mutagenesis studies have also shown that considerable deviation from this ideal pattern can still be associated with detectable levels of frameshifting *(38).*

An RNA pseudoknot (PK) located 3' to the frameshift site (or in rare cases merely a stem-loop structure) is usually also required for efficient frameshifting to occur *(39-41).* The primary role of the PK in frameshifting may be merely to cause a pause in translation, but a more active role is also possible.

It must also be noted that frameshifting with only one tRNA bound to the ribosomal P-site is known to occur in some cases, which would only require a 4-base shift signal in the message. This P-site slippage is often facilitated by a "hungry" codon in the second position (usually an Arg or Ser codon, but others are possible), which creates a pause owing to low cellular levels of the required isoacceptor tRNA, analogous to the pausing effect of RNA secondary structures, such as PKs *(42).*

Frameshifting in the +1 direction is also well documented *(39);* a relevant example of a +1 frameshift sequence is CCC followed by a stop codon, particularly CCCUGA *(43),* which will be discussed later in relation to its occurrence in the HIV-1 *nef* coding region.

ALTERNATIVE FUNCTIONS OF THE UGA CODON AND SEC INSERTION MECHANISMS

Even aside from its ability to encode selenocysteine, UGA is probably the most "leaky" of the three stop codons, any of which can be suppressed under appropriate circumstances, by means of mutant "suppressor" tRNAs and/or by programmed readthrough involving PKs, as seen in some retroviruses *(34).* When suppressed by such "conventional" mechanisms, UGA usually encodes an amino acid with a closely related codon, such as Trp, Cys, or Arg, in competition with termination, yielding a mixture of both the terminated product and an extended protein product.

A unique role of the UGA codon is its potential to encode SeC. In eukaryotes, selenocysteine incorporation at UGA depends on the presence of a protein factor and a structural signal in the mRNA 3'-untranslated region, called an SeC insertion sequence or SECIS element *(29,44,45).* A possibly related stem-loop structure is required in bacteria, but must immediately follow the UGA codon *(28).*

Finally, as mentioned above, if preceded by CCC or other shifty codons, an in-frame UGA can also be part of a frameshift signal, which may compete with termination and/or readthrough by any of the mechanisms listed above.

CURRENT STATUS OF THE VIRAL SELENOPROTEIN THEORY

At present, there has been no definitive demonstration that a potential selenoprotein gene is actually expressed by a virus. However, several specific predictions of the viral selenoprotein theory have now been verified. These include the following:

1. One of the key RNA pseudoknots that we predicted in HIV-1, overlapping the codons at the active site of the reverse transcriptase gene and associated with a well-conserved UGA codon in the -1 reading frame, has now been validated experimentally by chemical and enzymatic cleavage studies, reported in a recent paper and thesis from Barbara Carter's group at the University of Toledo *(46).*
2. Our prediction that Se should have an antiviral effect vs HIV has been independently verified in vitro by several investigators; note that this was not a novel prediction, since Schrauzer had first made it more than 10 years previously *(24).* Recently, Favier and coworkers *(26)* have studied the effects of Se supplementation on HIV-1 replication induced by oxidative stress in cell culture; significantly, they were totally unaware of our work at that time and chose to investigate this question based on completely independent lines of evidence. Noting that existing data "implicate an HIV-1 mediated antioxidant imbalance as an important factor in the progressive depletion of CD4 § T cells in AIDS," they demonstrated that in ACH-2 cells, at concentrations of 25-50 µg/L as sodium selenite, Se supplementation inhibits viral cytotoxic effects and the reactivation of HIV-1 by hydrogen peroxide, decreases activation of NF-r an important cellular transactivator of HIV-1, and protects against activation of HIV-1 by tumor necrosis factor. Similarly, preliminary results from the lab of Raymond Schinazi of Emory University, based on the screening of several organic and inorganic Se compounds in a standard assay for anti-HIV activity, indicate that certain simple Se compounds are active against HIV-1 at micromolar concentrations, comparable to plasma Se levels in healthy humans *(47).*
3. Most significantly, our proposal that some viruses may encode selenoproteins, which was received with skepticism a year or two ago, has now been substantiated by new findings from a group at NIH. Bernard Moss of NIAID recently reported the newly sequenced genome of the pox virus *Molluscum contagiosum (48).* This virus contained an open reading frame that is highly homologous to the known mammalian selenoprotein glutathione peroxidase (GSH-Px), with 80% sequence identity at the amino acid level, and an identically placed in-frame UGA codon. Given that GSH-Px is a fairly large protein, it is astronomically improbable that this match could occur by chance; this must

be a real selenoprotein gene. This also shows that viruses can gain some evolutionary advantage by encoding such an antioxidant protein, as we suggested previously *(31).* In the light of this finding, serious consideration must be given to our theoretical genomic evidence that selenoproteins are potentially encoded by HIV and other viruses like Coxsackie virus *(31,32).*

4. We now have firm experimental evidence from our own lab *(49)* demonstrating frameshift activity at a potential novel site in HIV, associated with a previously predicted potential selenoprotein gene overlapping the *nef* coding region, with a highly conserved -1 frameshift signal and UGA codon in the -1 reading frame *(32). The* results suggest that a significant level of frameshifting (around 5%) is taking place in this region. However, given the possible alternative roles of UGA codons discussed above, further work must be done to determine whether SeC is actually encoded by the conserved UGA codon in this novel variant of the HIV-1 *nef* protein. Theoretical data related to this potential gene will be presented below.

POTENTIAL SELENOPROTEIN GENES IN HIV-1 AND COXSACKIE VIRUS: NEW THEORETICAL FINDINGS

Since we have already published studies demonstrating the potential for several new genes in HIV-1 and Coxsackie virus *(31,32),* some of which may be selenoprotein genes, we will focus here only on new data related to several of the most important of these. We will also present new data showing highly distinctive GSH-Px-related sequences in CVB.

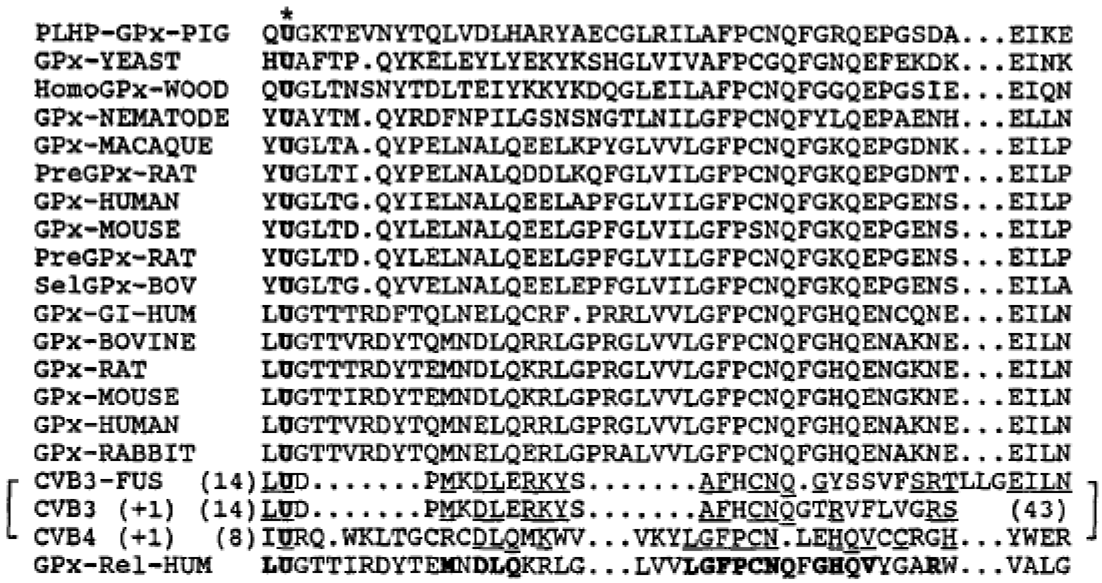
A Potential GSH-Px Gene in Coxsackie Viruses

We previously reported several potential selenoprotein genes in CVB *(32),* which have been implicated as a probable cofactor in Keshan disease, the classical Se-deficiency disease *(3).* Our findings thus suggest the possibility that viral selenoproteins may be an additional factor contributing to the observations of Beck et al., who have shown that the cardiovirulence of CVB3 is highly dependent on the Se status of the host, and that the virus actually mutates into a more virulent form in Se-deficient mice *(4,27).*

Potential selenoprotein genes previously identified in CVB3 included a potential ORF encoded between bases 3442 and 3749, with a start codon and eight in-frame UGA codons, as well as eight Cys codons *(32).* The hypothetical protein has a strong sequence similarity to a whole family of proteins that contain epidermal growth factor (EGF) modules, which are commonly found in various extracellular matrix and cell-adhesion proteins *(50).* This match is also notable in that five out of the eight-UGA codons align precisely with Cys codons in the human EGF protein. Significantly, at least three other viruses are known to encode EGF homologs *(51).*

A second potential selenoprotein in CVB3 may be encoded in a PPCR -1 to the main reading frame of CBV3, overlapping the entire *vpg* gene and much of the p3-c protease coding regions *(32).* This lacks a start codon, but could be expressed by a -1 frameshift owing to the presence of an "ideal" heptameric slippery sequence (GGGUUUU, beginning at position 5287) followed by a potential GC-rich PK located near the beginning of the *vpg* coding region of CBV3. The overlapping reading frame contains five in-frame UGA codons as well as five Cys codons.

Whether these are functional genes in some Coxsackie virus strains, or merely artifacts or evolutionary relics remains to be determined. However, the former possibility seems more viable in the light of our discovery in CVB of a candidate GSH-Px gene--a known selenoprotein (Fig. 2)--which, significantly is the same selenoprotein gene recently identified by Moss and coworkers in *M. contagiosum (48).* This potential GSH-Px gene in CVB was located by creating a data base of all known viral sequences, translating in all six reading frames, and using a mammalian GSH-Px-sequence as a data base probe for sequence similarities *(see* legend to Fig. 2). The CVB4 sequence shown in the figure was a top hit in the search. Significantly, the region of strong sequence similarity found by the search corresponds precisely to the active site of the GSH-Px protein (Fig. 3). In CVB, this sequence overlaps a region of the main polyprotein reading frame that encodes vp3, one of the picornavirus structural (capsid) proteins.



```
                   *
  PLHP-GPx-PIG     QUGKTEVNYTQLVDLHARYAECGLRILAFPCNQFGRQEPGSDA...EIKE
  GPx-YEAST        HUAFTP.QYKELEYLYEKYKSHGLVIVAFPCGQFGNQEFEKDK...EINK
  HomoGPx-WOOD     QUGLTNSNYTDLTEIYKKYKDQGLEILAFPCNQFGGQEPGSIE...EIQN
  GPx-NEMATODE     YUAYTM.QYRDFNPILGSNSNGTLNILGFPCNQFYLQEPAENH...ELLN
  GPx-MACAQUE      YUGLTA.QYPELNALQEELKPYGLVVLGFPCNQFGKQEPGDNK...EILP
  PreGPx-RAT       YUGLTI.QYPELNALQDDLKQFGLVILGFPCNQFGKQEPGDNT...EILP
  GPx-HUMAN        YUGLTG.QYIELNALQEELAPFGLVILGFPCNQFGKQEPGENS...EILP
  GPx-MOUSE        YUGLTD.QYLELNALQEELGPFGLVILGFPSNQFGKQEPGENS...EILP
  PreGPx-RAT       YUGLTD.QYLELNALQEELGPFGLVILGFPCNQFGKQEPGENS...EILP
  SelGPx-BOV       YUGLTG.QYVELNALQEELEPFGLVILGFPCNQFGKQEPGENS...EILA
  GPx-GI-HUM       LUGTTTRDFTQLNELQCRF.PRRLVVLGFPCNQFGHQENCQNE...EILN
  GPx-BOVINE       LUGTTVRDYTQMNDLQRRLGPRGLVVLGFPCNQFGHQENAKNE...EILN
  GPx-RAT          LUGTTTRDYTEMNDLQKRLGPRGLVVLGFPCNQFGHQENGKNE...EILN
  GPx-MOUSE        LUGTTIRDYTEMNDLQKRLGPRGLVVLGFPCNQFGHQENGKNE...EILN
  GPx-HUMAN        LUGTTVRDYTQMNELQRRLGPRGLVVLGFPCNQFGHQENAKNE...EILN
  GPx-RABBIT       LUGTTVRDYTQMNELQERLGPRALVVLGFPCNQFGHQENAKNE...EILN
[ CVB3-FUS    (14)LUD.......PMKDLERKYS.......AFHCNQ.GYSSVFSRTLLGEILN ]
[ CVB3 (+1)   (14)LUD.......PMKDLERKYS.......AFHCNQGTRVFLVGRS   (43) ]
[ CVB4 (+1)    (8)IURQ.WKLTGCRCDLQMKWV...VKYLGFPCN.LEHQVCCRGH...YWER ]
  GPx-Rel-HUM      LUGTTIRDYTEMNDLQKRLG...LVVLGFPCNQFGHQVYGARW...VALG
```

Fig. 2. Multiple sequence alignment of the catalytic core region of representative eukaryotic GSH-Px compared to GSH-Px-like sequences from Coxsackie virus strains B3 and B4. The asterisk indicates the position of the conserved catalytic selenocysteine (translated by the U symbol in the single-letter amino acid code). The viral sequences are shown in square braces; for easier comparison to the CVB4 sequence, one human GSH-Px-related sequence is placed below them, with matches to any of the three viral sequences highlighted in bold. Matches of amino acids in viral sequences to any in the aligned GSH-Px-sequences are indicated by underlining in the viral sequences. Numbers in brackets at the beginning of the viral sequences indicate the number of upstream residues to the preceding start codon (in CVB3) or stop codon (in CVB4). A potential frameshift around the glycine (G) codon following the CNQ sequence *(see* Fig. 4) would lead to a putative GSH-Px-vp3 fusion protein, shown here as CVB3-FUS.

In the CVB4 strain with the highest local similarity to GSH-Px in the overlapping (+1) reading frame (Genbank #S76772), there is no start codon or apparent frameshift site, suggesting that the gene could not be active in that viral strain (unless some type of RNA editing or other unusual event is involved). However, in another CVB4 isolate (Genbank #D00149), the overlapping reading frame extends further upstream, and there is a start codon, as well as additional in-frame

UGA codons. In CVB3 strains, there is also a start codon, and the entire catalytic site of GSH-Px is present within the span of 43 amino acids, suggesting that a highly truncated yet functional GSH-Px (Fig. 3) may be encoded in this potential ORF. It is also notable that, at least in CVB3 strains, there is only a single UGA codon that aligns with the single UGA of GSH-Px that encodes the catalytic SeC.

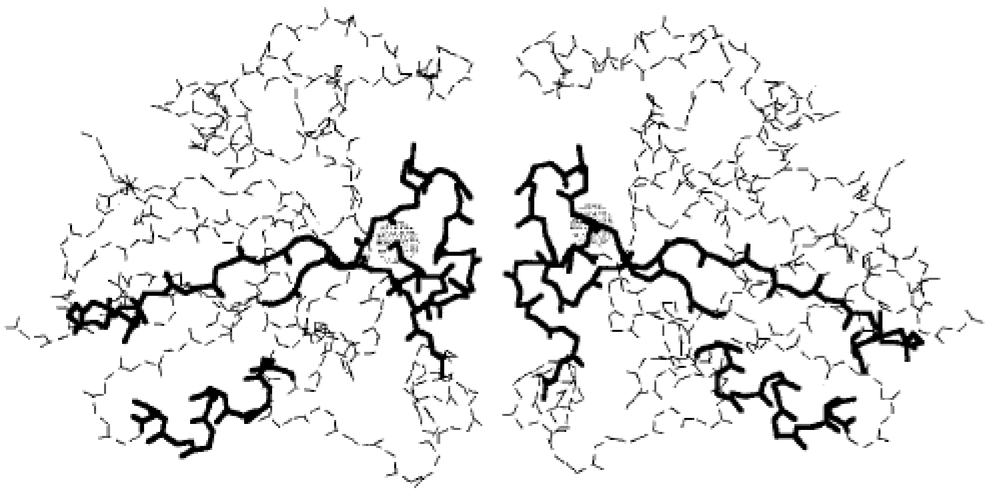


Fig. 3. X-ray crystal structure of GSH-Px-dimer with regions corresponding to the CVB3 GSH-Px-like sequence shown in Fig. 2 highlighted in bold lines. Despite several internal deletions and truncation at the N- and C-terminals relative to the GSH-Px sequences, the 43-amino acid protein potentially encoded in the +1 frame of CVB3 includes the critical active site region of GSH-Px and sufficient supporting structure to function possibly as a highly truncated GSH-Px-like molecule. (Dot spheres show location of Se atoms.)

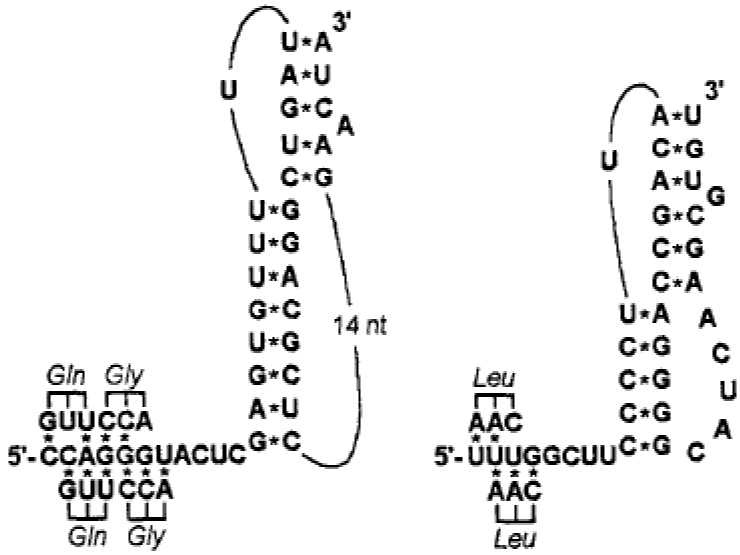


Fig. 4. Potential -1 frameshift sites associated with GSH-Px-related sequences in CVB3 strains (left) and in some CVB4 strains (right). Schematic tRNAs are shown before (below) and after the shift (above) relative to the hypothetical slippery sequences (asterisks indicate cognate base pairings). Slippage here would involve a shift from the overlapping +1 frame into the vp3 coding region of CVB, giving a GSH-Px vp3 fusion protein.

In both CVB3 and CVB4, it is also intriguing that downstream of this truncated GSH-Px coding region, sequence similarities to the C-terminal region of GSH-Px can still be observed in the main reading frame, i.e., within the vp3 protein sequence itself. In both the CVB3 and CVB4 strains that have start codons in the GSH-Px-related ORF, there are potential -1 frameshift signals (Fig. 4) that could permit this GSH-Px module to be fused to the C-terminal half of vp3, in which certain residues might contribute to the GSH-Px activity, because they match conserved GSH-Px residues in the alignment (e.g., the EILN sequence at the far right of Fig. 2, just past the frameshift site). Such a hypothetical fusion protein sequence is shown in Fig. 2 as CVB3-FUS, which extends past the end of the CVB sequences in the +1 frame that terminate following the conserved GSH-Px active site sequences.

If this frameshift site is active, out of the many copies of the vp3 protein synthesized, a small percentage (determined by frameshift efficiency and SeC availability) would contain a fused GSH-Px module. Because vp3 is a picornavirus structural protein present in about 60 copies/virion, this implies that GSH-Px activity may be associated with the virion itself. This hypothesis can be tested using standard assays by harvesting virus from infected cells supplied with Se and assaying purified viral filtrates for GSH-Px activity.

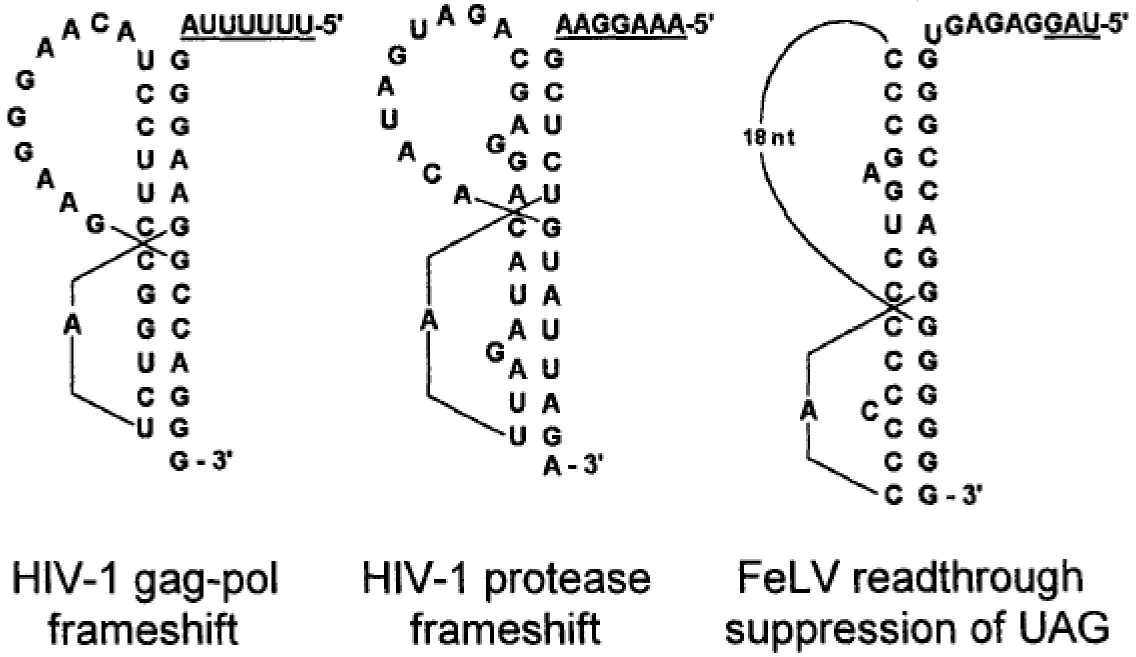
Independent of that possibility, the hypothetical truncated GSH-Px encoded in the +1 frame could potentially accumulate in infected cells and serve the virus in some regulatory or defensive role, e.g., to antagonize apoptotic defense mechanisms of the immune system that use oxidative processes to kill infected cells *(52).*

This potential GSH-Px gene in CVB3 may also be of some significance in regard to the specific mutations observed by Beck et al. in the conversion of an avirulent strain into a more cardiovirulent strain *(27),* because one of the mutations consistently associated with the phenotypic change was in the vp3 gene. Thus, it is possible that this mutation might in some way modify the GSH-Px-related activity in the hypothetical GSH-Px-vp3 fusion protein described above, possibly thereby contributing to the greater virulence observed even in Se-adequate animals.

A Potential Selenoprotein Gene Overlapping the HIV-1 Protease Coding Region

One of the most intriguing potential new genes we predicted in HIV overlaps the protease gene and would have to be expressed by a -1 frameshift from protease; thus, we have tentatively called it the *protease fs* or *pro fs* gene *(31). The* hypothetical protein it encodes has significant similarities to a number of known DNA-binding proteins, which is consistent with its high content of basic amino acids (computed isoelectric point > 11). Thus, if real, it is likely to be either a nuclear protein or possibly an RNA binding protein. We showed that it conformed to a multiple sequence alignment of a known type of viral DNA-binding protein, the family of papillomavirus E2 proteins *(31).* A collaborator at Ohio State University (Robert J. Hondal in the lab of Ming-Daw Tsai) has created a bacterial expression vector for the cysteine homolog of this hypothetical HIV selenoprotein (because it is impossible to clone mammalian selenoproteins in bacteria directly). By mutagenesis, Cys codons were substituted for the two UGA codons in this protein; however, one of these aligns with a conserved Cys codon in the E2 proteins, so the substitution is not necessarily inconsistent with activity. Results to date indicate that this protein appears to have biological activity, because activation of the promoter of the expression construct leads to growth inhibition in the transfected bacteria (this did not occur with control constructs from which the viral coding sequence was omitted). This result was somewhat expected, because the E2-1ike palindrome that we pointed out in the HIV-1 long-terminal repeat as a potential target for this protein, although extremely rare in human DNA other than in a few cytokine-related genes *(31),* is found in a number of bacterial genes associated with intermediary metabolism, leading to a concern about the feasibility of cloning it in bacteria (i.e., if it binds to that palindrome, it could interfere with expression of some bacterial genes).

Our interest in this potential gene has recently been stimulated by the discovery of a new family of RNA PKs, initially in certain bacteriophages, sharing an unusual topology involving a single base in the 5'-loop, spanning a 3'-stem of 6 or 7 bp *(53).* Du et al. showed not only that this remarkable structure was feasible (by NMR studies), but also that a large number of known frameshift and suppression sites in retroviruses had potential PKs consistent with this topology, which they called common pseudoknot motif 1 (CPK1).



HIV-1 gag-pol frameshift

HIV-1 protease frameshift

FeLV readthrough suppression of UAG

Fig. 5. The known HIV-1 *gag-pol* -1 frameshift site and pseudoknot in CPK1 topology *(53),* compared to a proposed -1 frameshift site in the HIV-1 protease coding region *(31).* In both cases, the slippery sequence comes right up to the base of the 5'-stem of the PK, and the PKs are similar except for several extruded bases in the helices of the protease PK structure. A potential CPK1 previously identified in feline leukemia virus (FeLV) *(53),* is shown for comparison to illustrate that such single-base extrusions are relatively common.

One such known frameshift site where they found a potential CPK1-type PK was the *gag-pol* frameshift site in HIV-1. This may have implications for the potential frameshift in the hypothetical *pro fs* gene, because, by extending the 3'-stem of the PK structure that we originally proposed for the protease PK *(31),* a CPKl-type PK (with a single A in the 5'-loop) is obtained

(Fig. 5). A comparison of this structure and the HIV-1 *gag-pol* PK in CPK1 topology also shows that in both cases, the heptameric frameshift sequence comes right up against the base of the 5'-stem of the PK (Fig. 5). Thus, this hypothetical *pro fs* frameshift signal is structurally almost identical to the known *gag-pol* frameshift signal. Although the *pro fs* heptameric shift sequence is not "ideal," as we pointed out previously *(32),* it still permits two out of three cognate base pairs on both tRNAs after slippage, with the mismatched base pairs potentially both being GU, with the appropriate tRNA isoacceptors. We are now initiating in vitro studies to attempt to quantitate the frameshift efficiency associated with this novel site in HIV-1.

It must also be noted that the first UGA codon in this region of the -1 frame overlapping the HIV-1 protease gene is conserved in all known isolates and subtypes of HIV-1 currently archived in the Los Alamos data base; this is the UGA that aligns with the conserved Cys codon in the DNA recognition helix of the papillomavirus E2 proteins *(31)*. The second in-frame UGA is conserved throughout HIV-1 subtypes B and D, except for a single type B isolate in which it has mutated to an Arg codon (perhaps significantly, in the alignment of the hypothetical protein sequence with papillomavirus E2 proteins, the most common residue at this point is Arg). The frameshift slippery sequence is conserved only in subtypes B, D, and U, but not in subtypes A and O.

Potential structural features within the HIV protease coding region that are possibly associated with SeC insertion, i.e., capable of acting as SeC insertion elements *(29),* will be described in a later section.

A Novel Frameshift Site and Conserved UGA Codon Overlapping the HIV-1 nef Coding Region

As pointed out previously *(32),* the *nef* coding region of HIV-1 contains a highly conserved "ideal" -1 frameshift sequence (UUUAAAA) followed by an RNA PK, with a well-conserved UGA codon in the -1 reading frame. The high degree of conservation of this UGA codon raises the possibility that this region overlapping the *nef* gene may encode a selenoprotein, although there are several possible alternatives (discussed below).

The frameshift heptamer and PK structure appear to be remarkably well conserved in all of the HIV-1 subtypes for which data are available. In only 5 of 76 *nef* sequences in the data base is there variation from the UUUAAAA sequence, and in all of those there is still considerable frameshift potential. Four of those five have mutated to UUUAAGA, which still has high slip potential because of the known role of Arg codons (AGA in this case) as "hungry codons," thus facilitating slippage at the ribosomal P-site (on UUUA). In the remaining exception, P-site slippage involving a single tRNA (on UUUU) is also a possibility.

In addition, the potential PK structure immediately downstream of the slippery sequence is very well conserved in HIV-1 isolates, as shown in Fig. 6. Note that this PK topology is somewhat different than what we originally proposed *(32);* this structure is more consistent with the nucleotide variations in the various HIV-1 subtypes, as well as within individual subtypes. Again, it is interesting that the structure appears to be another example of the CPKl-type PK, because there is a single nucleotide in the 5'-loop spanning a 6-bp helical stem (an A in over 80% of all HIV-1 isolates).

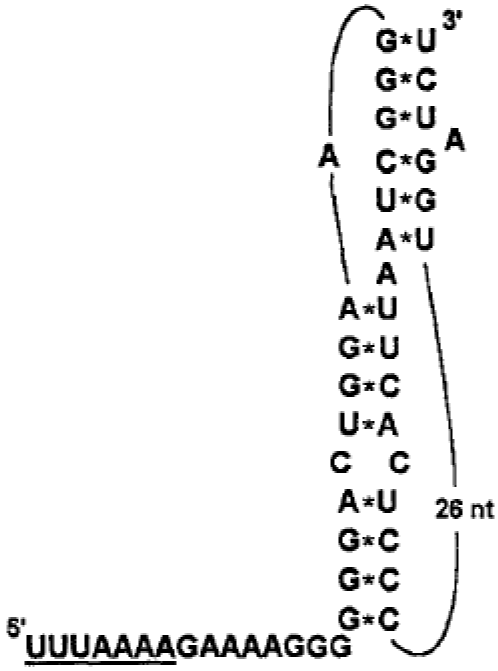


Fig. 6. An ideal heptameric -1 frameshift sequence and PK in the *nef* coding region of HIV-1. Both the PK structure and the heptamer are well conserved in HIV-1 variants *(see text). The* overlapping -1 reading frame contains several well-conserved UGA codons *(see* Fig. 7). Experiments have confirmed that this viral sequence induces a -1 frameshift, at about 5% efficiency in a standard assay *(49).*

As mentioned above, we now have firm experimental evidence that these conserved structural features (slip site and PK) are functional, because we have been able to demonstrate significant levels of frameshifting in an in vitro frameshift assay *(49),* using a construct containing an oligonucleotide corresponding to this 60-base sequence from the HIV-1 *nef* coding region (Fig. 6). Thus, we can say with confidence that at the least this *neffs* gene may encode a truncated isoform or variant of the *nef* protein, produced by the frameshift into this region followed by termination at the UGA codon. Alternatively, any type of opal suppression, including SeC insertion, would permit readthrough of the UGA and thus the attachment of an additional protein sequence encoded downstream in the -1 frame.

A compelling argument that there must be readthrough of the UGA comes from the fact that there is additional conservation of sequences in the -1 frame downstream of that first UGA, including a second UGA that occurs in the context of the sequence CCCUGA (beginning at base 8752 in the HIV-1 sequence, Genbank #K02013). CCCUGA is a known +1 frameshift signal in *Escherichia coli (43).* This suggests the intriguing possibility that the *nef fs* -1 frameshift discussed above could ultimately be followed by an efficient +1 frameshift returning to the original *nef* reading frame, leading to the "swapping" of the portion of the normal *nef* protein encoded in this region for an alternate module encoded in the overlapping -1 frame. Since the number of amino acids in both isoforms would be the same, this hypothetical variant *nef* protein would probably not be distinguishable from the normal *nef* protein on Western blots, and most antibodies would crossreact. Note that for this variant to be formed, there would have to be readthrough rather than termination at the first conserved UGA codon in the -1 frame (Fig. 7).

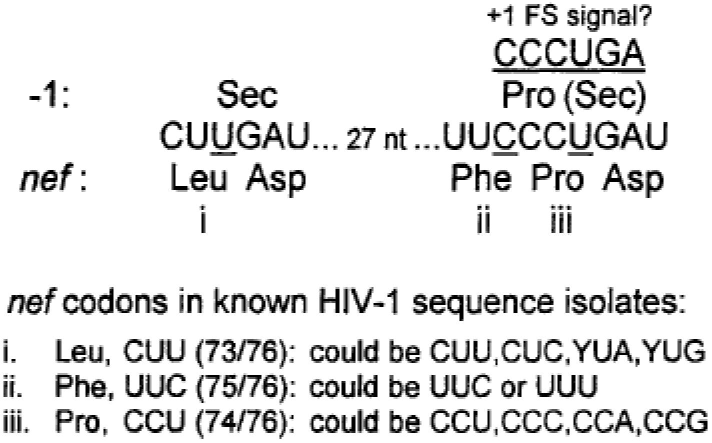

+1 FS signal?
CCCUGA
-1:
Sec
Pro (Sec)
CUUGAU... 27 nt ...UUCCCUGAU
nef:
Leu Asp
Phe Pro Asp
i
ii
iii
nef codons in known HIV-1 sequence isolates:
i. Leu, CUU (73/76): could be CUU,CUC,YUA,YUG
ii. Phe, UUC (75/76): could be UUC or UUU
iii. Pro, CCU (74/76): could be CCU,CCC,CCA,CCG

Fig. 7. Codon usage in the HIV-1 *nef* gene in relation to conservation of a UGA codon and CCCUGA sequence in the overlapping -1 reading frame. This region is downstream of the frameshift site shown in Fig. 6. As indicated, the *nef* codons at *i* (Leu) and *iii* (Pro) could use any base in their third codon positions (owing to the degeneracy of the genetic code), yet U is almost always seen in the 76 different HIV-1 isolates, suggesting conservation of UGA codons in the -1 frame. Similarly, the Phe codon at *ii* is almost always UUC rather than UUU, which suggests conservation of a CCC (Pro) codon in the -1 frame. The sequence CCCUGA is a known +1 frameshift signal, which could cause a return to the *nef* ORF after insertion of an alternate protein module encoded in the -1 frame. Conservation of the CCCUGA downstream of the first UGA suggests that the latter must be suppressed by readthrough, i.e., the first UGA encodes an amino acid rather than a termination event.

The rare HIV-1 subtype O (with only two isolates) is clearly unique in regard to this overlapping reading frame, because upstream from the first UGA in the -1 frame, it has a UAA stop codon not found in the other subtypes, and the UGA has mutated to GGA (Gly).

Excluding the two type O isolates, the first in-frame UGA codon in the overlapping -1 reading frame is present in all but 2 of the remaining 74 *nef* sequences in the Los Alamos HIV-1 sequence data base (one of which has mutated to an AGA [Arg] codon, and one to a UAA stop codon). This is remarkable in light of the fact that if only the *nef* gene in the zero frame is considered, there is no reason for the U of the UGA to be conserved, because the U is the last base of a leucine codon in *nef,* and owing to the degeneracy of the genetic code, any of the 4 bases should be permitted in that position (Fig. 6). Because of properties inherent in the transcriptional infidelity of reverse transcriptase, the HIV genome also tends toward a high A content ("A pressure"), so in a large set of divergent HIV-1 sequences, one would expect to see a high frequency of A bases in the third position of leucine codons, in the absence of overlapping genes.

As detailed in the legend to Fig. 7, the same argument applies to the well-conserved potential +1 frameshift sequence CCCUGA located slightly downstream. This second UGA is conserved in all but three isolates, where it has mutated to an Arg codon, which might still permit it to function as part of a +1 frameshift signal via a hungry codon mechanism.

The extent of conservation of these sequences is essentially identical to the degree of conservation of the *nef* ORF itself within this set of sequences, i.e., in 3 of 76 HIV-l-related sequences in the data base, the zero frame *nef* ORF is truncated because of premature termination at a stop codon in the middle of the frame (this includes the original HTLVIII sequence first reported by Gallo and coworkers *(56),* also called BH10). This may reflect the fact that the selective pressure to maintain the *nef* gene is only fully present in vivo, but not in cultured cells. Thus, we can fairly say that the sequence and structural features related to this hypothetical *neffs* gene (-1 frameshift signal, UGA and CCCUGA) are as well conserved within this set of HIV-1 sequences as the *nef* ORF itself.

In summary, the existence of this highly conserved UGA codon and CCCUGA sequence in the -1 frame, downstream of an "ideal" heptameric shift and fairly large PK, is strongly suggestive of the presence of a novel protein module overlapping the *nef* gene, possibly a selenoprotein module. However, we cannot rule out the possibility that simple termination or perhaps some other type of opal suppression could be taking place at this conserved UGA codon. In fact, all three possibilities could occur depending on intracellular conditions, such as the relative abundance of $tRNA_{Sec}$, opal suppressor tRNAs, and release factor.

OTHER VIRUSES THAT MAY ENCODE SELENOPROTEINS

Mouse Mammary Tumor Virus

In the light of our HIV results, MMTV is a prime candidate for this analysis, because it is probably the only retrovirus that has been proven to be sensitive to the chemoprotective effects of Se. In numerous studies, Schrauzer et al. has demonstrated that the incidence of mammary tumors consequent to infection with MMTV is dramatically reduced by increasing dietary Se *(6).* The relevance of this animal model to human disease is considerable: aside from the obvious implications for AIDS and other retroviral diseases, the implications for human cancer are now quite inescapable in light of the recent discovery of sequences highly homologous to the MMTV envelope gene in 38% of human breast cancer specimens *(54).*

A preliminary examination of MMTV for possible selenoprotein coding potential reveals several regions overlapping the MMTV *gag* gene that have appropriate frameshift signals capable of accessing PPCRs containing multiple in-frame UGA codons (Fig. 8). There are two PPCRs in the -1 frame to *gag,* both having potential -1 frameshift entry sites at their 5'-ends (indicated at A and B in Fig. 8). Site A has the potential slip sequence UAAAAGA, which is an excellent candidate for P-site slippage on UAAA, with AGA acting as a hungry Arg codon. This is followed 3 bases downstream by a small PK having >50% GC base pairs (not shown). The overlapping PPCR associated with site A has eight in-frame UGA codons. The second UGA-rich PPCR overlapping the MMTV *gag* gene contains four tandem UGA codons, downstream of potential frameshift site B (Fig. 8) with an ideal heptameric -1 frameshift sequence (AAAGGGA) and a PK that, like the HIV-1 *gag-pol* and potential HIV protease PKs (Fig. 5), appears to be another example of the Hoffman CPK1 type PK topology (Fig. 9).

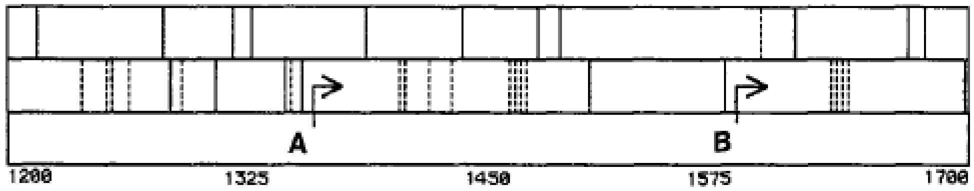


Fig. 8. PPCRs overlapping part of the MMTV *gag* gene. Numbers correspond to nucleotides in the genomic sequence (Genbank #D16249). A and B are possible -1 frameshift sites that could provide entry into the indicated PPCRs. Dashed lines are UGA codons.

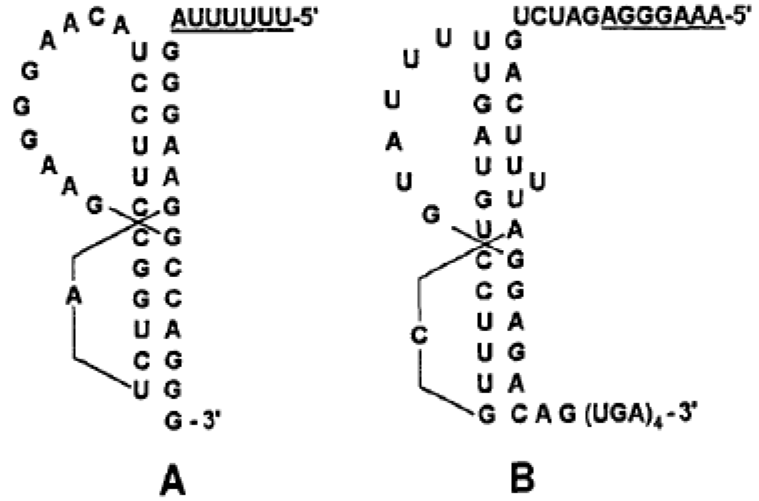


Fig. 9. A: The known HIV-1 *gag-pol* -1 frameshift site and PK in CPK1 topology *(53),* for comparison to **B:** potential frameshift site B in MMTV, with ideal heptamer and CPKl-like PK, located upstream from four tandem in-frame UGA codons *(see* Fig. 8).

Schrauzer et al. *(57)* has noted antagonism between dietary zinc and the chemoprotective effects of Se vs MMTV. Thus, it may be significant that the closely spaced UGA codons in these PPCRs are somewhat reminiscent of the Cys residues in metallothionein, suggesting the possibility of a metal ion chelating role. It should also be noted that the MMTV *gag* gene encodes a zinc finger module just downstream of the region shown in Fig. 8. Thus, if expressed, these hypothetical selenoproteins would be formed in small amounts as alternative modules to the zinc finger region. Whether they might interact with zinc ions or zinc fingers, contributing to the observed antagonism, remains to be determined.

Hepatitis B Virus (HBV)

Of other viruses we have examined, HBV is notable for having been shown to have reduced pathogenicity in Se-supplemented human populations *(1,2).* Although classified as DNA viruses, hepadnaviruses like HBV are closely related to retroviruses in that they utilize an RNA intermediate and encode a reverse transcriptase enzyme. Thus, it is highly significant that a potential -1 frameshift sequence and large RNA PK can be found overlapping the region encoding the YMDD active site codons of the HBV reverse transcriptase (Fig. 10), because this is precisely the same location where a PK that we predicted in HIV-1 has now been supported by chemical and enzymatic stability studies *(46).* The fact that both HIV-1 and HBV have quite divergent slippery sequences and PK topologies suggests that what are being conserved in these two distant relatives are simply the elements necessary to produce a frameshift into the -1 frame, along with its conserved UGA codon overlapping the YMDD sequence of the polymerase. If this overlapping PPCR is functional, it is possible that it may be an ancient feature of reverse transcriptase, possibly predating the divergence of these two types of viruses. Alternatively, the existence of these similar structural elements in HBV and some retroviruses might reflect a homologous recombination event that could have taken place in more recent times.

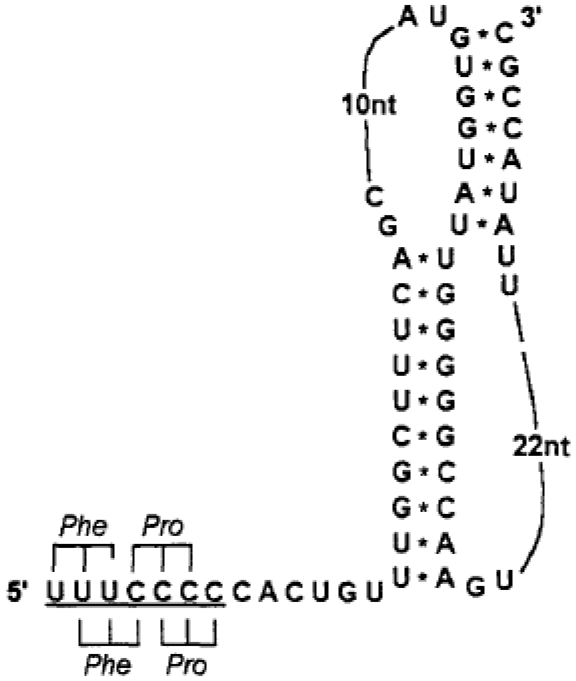


Fig. 10. A large potential PK and heptameric -1 frameshift sequence in HBV (Genbank #X51970), overlapping the active site codons of the viral reverse transcriptase. A PK predicted in precisely the same location in HIV-1 *(31)* has now been supported by enzymatic and chemical stability studies *(46).*

DNA Viruses

At this point in time, the most compelling evidence for a virally encoded selenoprotein is actually in a DNA virus: the pox virus *M. contagiosum,* which, as mentioned previously, was recently shown to encode an ORF with 80% sequence identity to the known mammalian selenoprotein GSH-Px *(48).* Although such a high level of sequence homology probably reflects the recent acquisition of a copy of a cellular gene, the fact that the virus has retained it shows that a virus can gain some advantage from encoding its own antioxidant protein. This would permit a virus to regulate oxidant tone within the infected cell, possibly to repress its own transcription (since that is typically activated by oxidative stress), and also for defensive purposes. The latter could be quite important in light of accumulating evidence *(52)* that cytotoxic T-cells can invoke oxidative stress in target cells (e.g., by Fas-mediated stimulation) in order to kill them by programmed cell death (apoptosis). A virally encoded GSH-Px or other antioxidant protein might be used to counter such an attack, and permit the infected cell to live longer and ultimately produce more viral progeny.

The fact that DNA viruses can have massive genomes in comparison to most RNA viruses clearly presents a greater potential for incorporating specialized genes, such as antioxidant or selenoproteins, but at the same time can make finding them more difficult. We have observed several quite large UGA-rich PPCRs in certain herpesviruses (e.g., cytomegalovirus and Epstein-Barr) with start codons and up to 11 in-frame UGA codons spanning over 400 amino acids. However, we have so far not been able to find any obvious analogs of known eukaryotic SECIS elements closely associated with these UGA-rich PPCRs. We have also noted a remarkable stretch of multiple tandemly repeated TGA codons near the replication origin of some herpesviruses, but can presently offer no insights regarding their possible significance.

RNA Viruses and the Possibility of Nonspecific Selenoprotein Synthesis

Because of the known role of Se in blood-clotting mechanisms and the thrombotic and even hemorrhagic effects of Se deficiency, we have been particularly interested in the possibility that hemorrhagic fever viruses might be able to incorporate SeC in some viral proteins. This is discussed at length in the accompanying article *(55).*

The causative agents in viral hemorrhagic fevers are generally single-stranded RNA viruses, which as a class have recently been shown to have significantly higher than expected numbers of UGA codons in the -1 reading frame overlapping the known ORFs of many of their genes, e.g., up to threefold higher than expected in paramyxovirus N genes *(30).* Given that frameshifting in viruses is most commonly in the -1 direction, the -1 frame is precisely where one would expect to find a virally encoded selenoprotein, for which expression by the inefficient frameshifting process would be preferred because of the low abundance of $tRNA_{Sec}$. It is also possible that some of these overlapping UGA-rich PPCRs might be vestigal remnants of ancient (seleno)proteins now relegated to inactive reading frames.

Alternatively, even if there is no programmed synthesis of selenoproteins taking place, a low frequency of translational slippage into these frames would be expected to occur on a purely random basis, leading to encounters with in-frame UGA codons, proportional to the density of UGAs in the overlapping frames. Rima suggests that this might be the purpose of the high stop codon density: to bring about rapid termination of erroneously frameshifted protein chains *(30).* However, because stop codon recognition by termination factors is known to be slow relative to tRNA recognition (the basis of the "shifty stop" frameshift mechanism *[39])* and because $tRNA_{Sec}$ has the cognate anticodon for UGA, it is possible that during the translational pause before the release factor is able to bind, SeC could be inserted at some small fraction of the UGA codons encountered during such random translational events. Even if such events were quite rare, they might make an impact on intracellular host SeC levels if the virus was replicating at extremely high rates, particularly if it had a high enough density of UGAs in the overlapping frames. In essence, this would be nonspecific selenoprotein synthesis, i.e., taking place without the involvement of RNA stem-loop SeC insertion elements. Although such aberrant SeC-containing proteins would be difficult to detect individually owing to their low concentrations and structural diversity, their cumulative effect on cellular SeC pools could be significant. The possible implications for hemorrhagic pathology are discussed elsewhere *(55).*

SPECIFIC SELENOPROTEIN SYNTHESIS: THE PROBLEM OF POTENTIAL VIRAL SECIS ELEMENTS

To produce consistent and usable amounts of protein, any specific programmed selenoprotein synthesis by a virus would require a cotranslational mechanism identical or analogous to that involving eukaryotic SECIS elements or the equivalent RNA stem-loop structures in bacteria. Such elements are necessary in order for SeC insertion at UGA to compete successfully with release-factor-mediated protein chain termination.

The greatest challenge to the viral selenoprotein theory at present is to demonstrate such SeC-insertion activity at any specific viral UGA codon, which to date has not been accomplished. However, very little experimental effort has been expended in this direction. To our knowledge, only three candidate viral elements have so far been tested for ability to direct SeC insertion in a human 5'-deiodinase gene construct, all with negative results (Nadimpalli et al., unpublished data). These include a region of the HIV-1 *pol* gene and a region downstream of the potential CVB3 selenoprotein with EGF homology, both suggested previously as potential viral SECIS elements *(32).* The third was one of three potential SECIS elements that we identified in Ebola virus, shown as putative SECIS A in Fig. 3 of ref. *(55).* Of these, the one in CVB3 still merits further study because it was tested using a synthetic oligo that was designed with terminal "clamps" owing to its short length, which may have adversely affected activity by stabilizing the incorrect (computed rather than actual) base stem structure. It will be retested in its natural RNA structural context by using a larger region of CVB3 obtained by subcloning from the viral genome.

A number of points merit consideration in regard to the possibility of specific SeC insertion in viral proteins. The fact that GSH-Px homologs or distinctive GSH-Px-related sequences have now been reported in a pox virus *(48)* and in Coxsackie viruses (this work) strongly suggests that such insertion is likely to be taking place, at the least providing incentive for further study of this possibility. These are also the first examples of hypothetical viral selenoproteins whose activity could be predicted with confidence; since there are established assays for GSH-Px activity, it should be easy to prove in vitro that these are functional selenoproteins, even before the exact location of the viral SECIS element might be identified.

We must also consider the possibility that viruses might use a unique mechanism for SeC insertion or a variant of a known mechanism, particularly considering the significant differences that exist between bacterial and eukaryotic SeC insertion methods. For example, it may be significant that in some bacterial selenoproteins, the RNA structure downstream of the coding UGA can alternatively exist as a PK rather than as a simple stem-loop structure (e.g., Fig. 11A, in glycine reductases; Fig. 11B, in formate dehydrogenase). This raises the possibility that RNA PKs might be able to play a role in SeC insertion, which would also be consistent with the well-established role of RNA PKs in retroviral suppression of other stop codons *(34).* Many of the potential coding UGAs that we have pointed out, particularly in HIV-1, are on or near PKs. Even the potential coding UGA in the hypothetical CVB3 GSH-Px homolog is adjacent to a possible

RNA PK, analogous to the situation in some bacterial selenoprotein genes (Fig. 11). This suggests that SeC insertion in some viral genes might involve a local signal akin to that in bacteria, rather than the more versatile eukaryotic SECIS mechanism, which permits any in-frame UGA to be decoded as SeC.

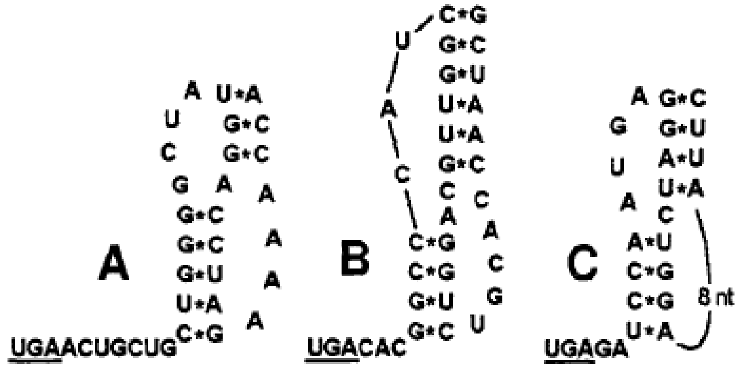


Fig. 11. Potential RNA pseudoknots adjacent to SeC-encoding UGA codons in bacterial selenoproteins. A: In *Clostridium sticklandii* glycine reductase (Genbank #M60399); B: in *E. coli* formate dehydrogenase (Genbank #M13563); C: a possibly similar situation in the hypothetical CVB3 GSH-Px homolog. The underlined UGA is the one shown aligned with the catalytic SeC of GSH-Px in Fig. 2.

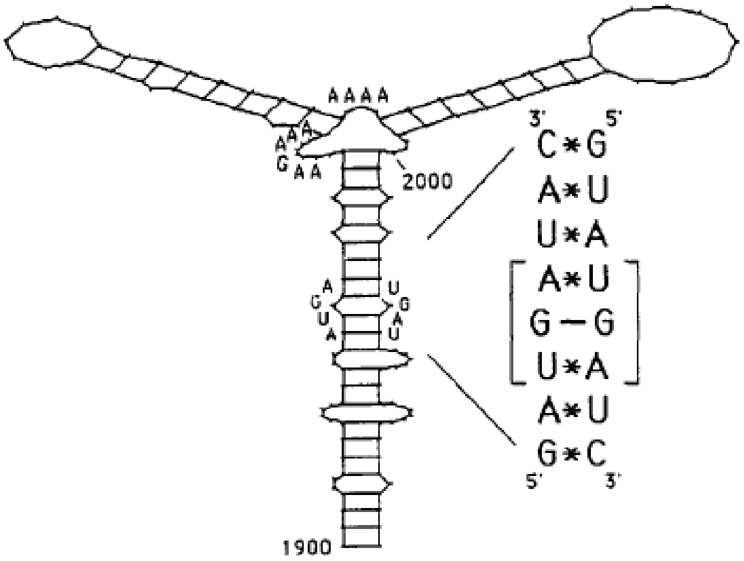


Fig. 12. RNA structure predicted for a portion of the HIV-1 protease coding region, where a selenoprotein may be encoded in the overlapping -1 reading frame *(31).* The two potential coding in-frame UGAs (shown in braces) are located in complementary regions that form a stem with an embedded 5'-AUGA, forced to pair with UGAU-3' as shown in detail in the inset. This structure has many features characteristic of a eukaryotic SECIS element, yet rather than being in an untranslated region, the AUGA and UGAU involve the potential coding UGAs, a situation akin to a bacterial SeC insertion element. This structure is the global energy minimum predicted by the Zuker FOLD program (stability = -17.2 kcal/mol), as implemented in the GCG software package (Program Manual for the Wisconsin Package, Ver. 8, September 1994, Genetics Computer Group, 575 Science Drive, Madison, WI 53711).

Along these lines, we have identified possible structural features in HIV-1 that may suggest how SeC might be inserted at the two conserved UGA codons in the hypothetical protein overlapping the HIV protease coding region. It seems more than coincidental that these two in-frame UGAs are both in the context of local sequences that are complementary to each other, and thus can pair to make an RNA stem structure in which the coding UGAs are forced to base pair with each other. As shown in Fig. 12, this gives a stem with an AUGA on the 5'-arm pairing with a UGAU on the 3'-arm, with a forced noncanonical G-G base pair in the center. These are precisely the consensus bases found in the stem region of most eukaryotic SECIS elements *(29),* and the pairing to form the G-G base pair is consistent with current models of SECIS structure and function (Berry and Low, personal communication). Note that this AUGAUGAU paired stem region is about 8-10 bp below a bulge region with runs of unpaired A bases, which is also a conserved feature of SECIS elements. Thus, this region of HIV has many features of a eukaryotic SECIS element. However, at the same time, it is more like a bacterial element, because it is located in the coding region, and the UGAs in the stem are the potential coding UGAs, whereas eukaryotic SECIS elements are always located in 3'-untranslated regions.

We previously pointed out an RNA stem-loop structure in the repeat region of the HW-1 long-terminal repeat that is very similar to the SECIS of the 5'-deiodinase gene, except that the stem is shorter and lacks the AUGA-UGAU sequences in the lower stem *(31).* Since that AUGAUGAU region probably interacts with the anticodon loop of $tRNA_{Sec}$, it is possible that the upper stem-loop with the AAA sequence may be all that is required to bind the eukaryotic SelB homolog. If so, SelB homolog binding to the HIV-1 stem-loop might be able to inhibit partially the termination factor even in the absence of an ability to bind t $tRNA_{Sec}$, and the presence of copies of this structure at both ends of the HIV-1 mRNA might decrease termination efficiency at all in-frame UGA codons in the HIV-1 genome. That could favor *any* type of termination suppression at UGA, even under conditions of Se depletion, which might be advantageous because it would permit various amino acids to be inserted at UGA depending on cellular conditions.

Whatever its function, it would appear that this structure is of some importance to the virus, because the stem and the UAAAG sequence on it (which is identical to that of the 5-deiodinase SECIS) are absolutely conserved in all known HIV-1 isolates. Such a mechanism, particularly if combined with a local SECIS-like signal favoring SeC insertion (possibly a pseudoknot or a structure in the coding region, such as that shown in Fig. 12) would ultimately be even more versatile than the eukaryotic SECIS system, because it could simultaneously permit various amino acid translations of different UGA codons in the genome, depending on their context.

Clearly, such speculations aside, there is as yet no concrete evidence of a specific SeC insertion mechanism in any viral genes. However, that possibility must be evaluated in the context of the rest of the data, both theoretical and otherwise, that we have reviewed in the previous sections.

CONCLUSIONS

Although there is still no definitive proof that some viruses can encode selenoproteins, we have reviewed a fairly extensive body of evidence that tends to favor that possibility. It is probably only a matter of time before functional GSH-Px-like activity will be demonstrated in several viruses where GSH-Px-related sequences have now been identified *(48)*. In addition, certain previously predicted genomic features in HIV-1 that are associated with highly conserved UGA codons in overlapping reading frames have now been experimentally confirmed, including a novel RNA PK *(46)* and a functional -1 frameshift site in HIV-1 *(49)*.

Despite these encouraging developments, an immense amount of work remains to be done before we will be able to assess definitively the extent to which viruses may directly or indirectly utilize or impact on selenobiological processes inside infected cells, if indeed they do at all. However, there is no doubt that Se has an effect on viruses, as these symposium proceedings attest. The question is; What are the mechanisms underlying that effect? Hopefully, the data presented here will inspire researchers from related fields to consider ways in which this fascinating question might be best resolved.

ACKNOWLEDGMENTS

The authors would like to thank Gerhard Schrauzer for the invitation to participate in this symposium. We would also like to acknowledge Robert J. Hondal of Ohio State University for pointing out the fact that the potential RNA pseudoknot in the HIV-1 protease coding region might be an example of the CPK1 family, as well as for other helpful discussions.

REFERENCES

1. S.Y. Yu, W. G. Li, Y. J. Zhu, W. P. Yu, and C. Hou, *Biol. Trace Element Res.* 20, 15-22 (1989).

2. S. Y. Yu, Y. J. Zhu, W. G. Li, Q. S. Huang, C. Z. Huang, Q. N. Zhang and C. Hou, *Biol. Trace Element Res.* 29, 289-294 (1991).

3. J. Bai, S. Wu, K. Ge, X. Deng and C. Su, *Acta Acad. Med. Sin.* 2, 29-31 (1980).

4. M. A. Beck, P. C. Kolbeck, L. H. Rohr, Q. Shi, V. C. Morris and O. A. Levander, *J. Med. Virol.* 43, 166-170 (1994).

5. J. C. Hou, Z. Y. Jiang and Z. F. He, *Chung Hua I Hsueh Tsa Chih* 73, 645-646 (1993).

6. G. N. Schrauzer, T. Molenaar, K. Kuehn and D. Waller, *Biol. Trace Element Res.* 20, 169-178 (1989).

7. B. M. Dworkin, G. P. Wormser, W. S. Rosenthal, S. K. Heier, M. Braunstein, L. Weiss, R. Jankowski, D. Levy and S. Weiselberg, *Am. J. Gastroenterol.* 80, 774 (1985).

8. B. M. Dworkin, W. S. Rosenthal, G. P. Wormser and L. Weiss, *J. Parenter. Enteral Nutr.* 10, 405-407 (1986).

9. J. F. Zazzo, J. Chalas, A. Lafont, E Camus and P. Chappuis, *J. Parenter. Enteral Nutr.* 12, 537-538 (1988).

10. B. M. Dworkin, W. S. Rosenthal, G. P. Wormser, L. Weiss, M. Nunez, C. Joline and A. Herp, *Biol. Trace Element Res.* 15, 167-177 (1988).

11. B. M. Dworkin, P. P. Antonecchia, R Smith, L. Weiss, M. Davidian, D. Rubin and W. S. Rosenthal, *J. Parenter. Enteral Nutr.* 13, 644-647 (1989).

12. L. Olmsted, G. N. Schrauzer, M. Flores-Arce and J. Dowd, *Biol. Trace Element Res.* 20, 59-65 (1989).

13. K. W. Beck, P. Schramel, A. Hedl, H. Jaeger and W. Kaboth, *Biol. Trace Element Res.* 25, 89-96 (1990).

14. L. Kavanaugh-McHugh, A. Ruff, E. Perlman, N. Hutton, J. Modlin and S. Rowe, *J. Parenter. Enteral Nutr.* 15, 347-351 (1991).

15. A. Cirelli, M. Ciardi, C. de-Simone, F. Sorice, R. Giordano, L. Ciaralli and S. Costantini, *Clin. Biochem.* 24, 211-214 (1991).

16. C. Allavena, B. Dousset, T. May, C. Amiel, F. Nabet-Belleville and P. Canton, *Presse. Med.* 20, 1737 (1991).

17. E. Mantero-Atienza, R. S. Beach, M. C. Gavancho, R. Morgan, G. Shor-Posner and M. K. Fordyce-Baum, *J. Parenter. Enteral Nutr.* 15, 693-694 (1991).

18. J. P. Revillard, C. M. Vincent, A. E. Favier, M. J. Richard, M. Zittoun and M. D. Kazatchkine, *J. Acquired Immune Defic. Syndr.* 5, 637-638 (1992).

19. J. Constans, J. L. Pellegrin, E. Peuchant, M. E Thomas, M. F. Dumon, C. Sergeant, and M. Simonoff, *Rev. Med. Interne.* 14, 1003 (1993).

20. A. Favier, C. Sappey, P. Leclerc, P. Faure and M. Micoud, *Chem.-Biol. Interact.* 91, 165-180 (1994).

21. B. M. Dworkin, *Chem.-Biol. Interact.* 91, 181-186 (1994).

22. R. Bologna, F. Indacochea, G. Shor-Posner, E. Mantero-Atienza, M. Grazziutti, M.-C. Sotomayor, M. Fletcher, C. Cabrejos, G. B. Scott and M. K. Baum, *J. Nutr. Immunol.* 3, 41-49 (1994).

23. C. Sergeant, M. Simonoff, C. Hamon, E. Peuchant, M. E Dumon, M. Clerc, M. J. Thomas, J. Constans, C. Conri, J. L. Pellegrin, and B. Leng, in *Oxidative Stress, Cell Activation and Viral Infection* C. Pasquier, ed., Birkhauser Verlag, Basel, pp. 341-351 (1994).

24. G. N. Schrauzer and J. Sacher, *Chem.-Biol. Interact.* 91, 199-206 (1994).

25. C. Allavena, B. Dousset, T. May, F. Dubois, P. Canton and E Belleville, *Biol. Trace Element Res.* 47, 133-138 (1995).

26. C. Sappey, S. Legrand-Poels, M. Best-Belpomme, A. Favier, B. Rentier and J. Piette, *AIDS Res. Human Retroviruses* 10, 1451-1461 (1994).

27. M. A. Beck, Q. Shi, V. C. Morris and O. A. Levander, *Nature Med.* 1, 433-436 (1995).

28. Bock, K. Forchhammer, J. Heider, W. Leinfelder, G. Sawers, B. Veprek and F. Zinoni, *Mol. Microbiol.* 5, 515-20 (1991).

29. M. J. Berry and P. R. Larsen, *Biochem. Soc. Trans.* 21, 827-832 (1993).

30. B. K. Rima, *Biochem. Soc. Trans.* 1, 1-13 (1996).

31. W. Taylor, C. S. Ramanathan, R. K. Jalluri and R. G. Nadimpalli, *J. Med. Chem.* 37, 2637-2654 (1994).

32. W. Taylor, C. S. Ramanathan, and R. G. Nadimpalli, in M. Witten, ed., *Computational Medicine, Public Health and Biotechnology: Building a Man in the Machine* Part 1, World Scientific, London, pp. 285-309 (1996).

33. A. Sanchez, S. G. Trappier, B. W. J. Mahy, C. J. Peters and S. T. Nichol, *Proc. Natl. Acad. Sci. USA* 93, 3602-3607 (1996).

34. D. L. Hatfield, J. G. Levin, A. Rein and S. Oroszlan, *Adv. Virus Res.* 41, 193-239 (1992).

35. T. Jacks, M. D. Power, F. R. Masiarz, P. A. Luciw, P. J. Barr and H. E. Varmus, *Nature* 331, 280-283 (1988).

36. T. G. Parslow, in *Human Retroviruses,* B. R. Cullen, ed., Oxford University Press, New York, pp. 101-136, (1993).

37. T. Jacks, H. D. Madhani, F. R. Masiarz and H. E. Varmus, *Cell* **55,** 447-458 (1988).

38. I. Brierley and J. A. Jenner, *J. Mol. Biol.* 227, 463-479 (1992).

39. R. B. Weiss, *Current Opin. Cell Biol. 3,* 1051-1055 (1991).

40. M. Chammoro, N. Parkin and H. E. Varmus, *Proc. Natl. Acad. Sci. USA* 89, 713-717 (1992).

41. E. ten Dam, K. Pleij and D. Draper, *Biochemistry* 31, 11,665-11,676 (1992).

42. J. Gallant and D. Lindsley, *Biochem. Soc. Trans.* 21, 817-821 (1993).

43. M. H. de Smit, J. van Duin, P. H. van Knippenberg and G. H. van Eijk, *Gene* 143, 43-47 (1994).

44. Q. Shen, F. F. Chu and P. E. Newburger, *J. Biol. Chem.* 268, 11463-9 (1993).

45. M. J. Berry, L. Banu, J. W. Harney and P. R. Larsen, *EMBO J.* 12, 3315-3322 (1993).

46. J.-M. A. Battigello, M. Cui, S. Roshong and B. Carter, *Bioorganic Med. Chem.* 3, 839-849 (1995).

47. E. W. Taylor, C. S. Ramanathan, R. G. Nadimpalli, and R. E Schinazi, *Antiviral Res.* 26, A271, #86 (1995).

48. T. G. Senkevich, J. J. Bugert, J. R. Sisler, E. V. Koonin, G. Darai, and B. Moss, *Science* 273, 813-816 (1996).

49. R. G. Nadimpalli, J. A. Hamilton, A. Thakur, R. G. Dean, E. W. Taylor, and B. M. Blumberg, *Virus Res.,* submitted for publication (1997).

50. J. Engel, *FEBS Lett.* 251, 1-7 (1989).

51. A. Opgenorth, N. Nation, K. Graham and G. McFadden, *Virology* 192, 701-709 (1993).

52. T. M. Buttke and P. A. Sandstrom, *Free Radical Res.* 22, 389-397 (1994).

53. Z. Du, D. P. Giedroc and D. W. Hoffman, *Biochemistry* 35, 4187-4198 (1996).

54. Y. Wang, J. F. Holland, I. J. Bleiweiss, S. Melana, X. Liu, I. Pelisson, A. Cantarella, K. Stellrecht, S. Mani, and B. G.-T. Pogo, *Cancer Res.* 55, 5173-5179 (1995).

55. C. S. Ramanathan and E. W. Taylor, Computational genomic analysis of hemorrhagic fever viruses, *Biol. Trace Element Res.* 56, 93-106 (1997).

56. Wong-Staal, R. C. Gallo, N. T. Chang, J. Ghrayeb, T. S. Papas, J. A. Lautenberger, M. L. Pearson, S. R. Petteway Jr., L. Ivanoff, K. Baumeister, E. A. Whitehorn, J. A. Rafalski, E. R. Doran, S. J. Josephs, B. Starcich, K. J. Livak, R. Patarca, W. A. Haseltine, and L. Rather, Complete nucleotide sequence of the aids virus, htlv-iii, *Nature* 313, 277-284 (1985).

57. N. Schrauzer, D. A. White, and C. J. Schneider, *Bioinorg. Chem.* 6, 265-270 (1976).