

# RNA Secondary Structure and Sequence Conservation in C1 Region of Human Immunodeficiency Virus Type 1 *env* Gene

OFER PELEG,<sup>1</sup> SØREN BRUNAK,<sup>2</sup> EDWARD N. TRIFONOV,<sup>1</sup> EVIATAR NEVO,<sup>1</sup>  
and ALEXANDER BOLSHOY<sup>1</sup>

## ABSTRACT

We have analyzed amino acid, nucleotide sequence, and RNA secondary structure variability in the *env* gene of human immunodeficiency virus type (HIV-1). In applying algorithms for computing optimal RNA-folding patterns to a nonredundant data set of 178 *env* nucleotide sequences, we found a conserved RNA stem-loop structure in the first conserved (C1) region of the *env* gene. This detailed examination also revealed the known secondary structure conservation of the Rev-responsive element (RRE). This finding is also supported by a higher third position conservation of the translatable reading frame along these subregions. The typical folding of the C1 region consists of two isolated stem-loop structures. These highly conserved structures are likely to have a biological function. This assumption is supported by the conservation of the third position along the coding region of these structures. The third position retains a conservation level above what would be statistically expected.

## INTRODUCTION

### *Diversity in HIV-1*

**D**IFFERENT ISOLATES of human immunodeficiency virus type 1 (HIV-1) exhibit striking genetic diversity.<sup>1–3</sup> This is a result of the high evolutionary rate of HIV-1 caused by the error-prone reverse transcriptase. The diversity is especially high in the *env* and *gag* genes.<sup>3,4</sup> In the region encoding the envelope glycoprotein gp120 the genetic diversity has been shown to directly influence the HIV-1 phenotype.<sup>5</sup> Several conserved and hypervariable regions along the *env* gene have been identified<sup>6,7</sup> and analyzed.<sup>8–10</sup> In particular, intense analysis has been performed for the third hypervariable region (V3).<sup>11</sup> The HIV-1 genome variability gives rise to drug resistance, enables the escape of the virus from immune responses, and apparently has so far prevented the development of an effective vaccine. There is, therefore, a clear need for improved understanding of the underlying selection governing the variability and conservation of the genome. The presence of conserved RNA secondary structures could be one such underlying factor.

A conserved region at the gene level (e.g., a consequence of functional RNA secondary structures) and the protein level (as

a consequence of functional and structural protein constraints) could very well serve both as constant and common target for new therapeutic as well as immunological approaches that would be less susceptible to genetic escape.

There are several well-known RNA secondary structures along the HIV-1 genome that play various functional roles during the virus life cycles. The best known are the *trans*-activation responsive (TAR) elements, which interact with the Tat protein,<sup>12–14</sup> and the Rev responsive element (RRE), which interacts with the Rev *trans*-activator protein.<sup>15–18</sup> In elucidation of these structures both the statistical treatment of the variability and RNA secondary structure prediction had played a crucial role.<sup>15,17,19,20</sup> Some yet undiscovered RNA secondary structures may also be involved in regulation of HIV-1 gene expression, in dimerization of single-stranded HIV genomes, in regulation of the retrovirus mutation rate, in transportation of genomic RNA to the cytoplasm, and in splicing events.

### *HIV-1 RNA structure*

The well-known difficulties in prediction of the minimum free energy RNA secondary structure for the entire HIV-1 genome are caused by its large size.<sup>21</sup> Our efforts focus on pre-

<sup>1</sup>Genome Diversity Center, Institute of Evolution, Haifa University, Mt. Carmel, Haifa 31905, Israel.

<sup>2</sup>Center for Biological Sequence Analysis, Department of Chemistry, Technical University of Denmark, DK-2800 Lyngby, Denmark.

diction of conserved functional RNA secondary structures located in the *env* region. To elucidate features and positions of such structures, we assessed the genetic diversity of all available sequences encoding the complete envelope protein and applied various computational methods to predict common RNA folds within the *env* gene. Studies of variability along the *env* gene by multiple alignment help to outline the conserved and variable regions. Definitions of the conserved and variable regions of the envelope protein are usually related to variability of amino acid *Env* sequences. The common approach concerning these regions is basically related to protein immunogenic determinants and protein structure and function.

However, assuming that the conservation of a protein sequence is the only source of evolutionary pressure would be an oversimplification of the picture. The RRE is the best example of how selection for RRE RNA secondary is consistent with amino acid sequence conservation of the *env* gene at the gp120–gp41 junction cleaved by a nonvariable cellular protease. This region is thereby conserved for at least two reasons: to preserve the gp120–gp41 consensus cleavage motif and to preserve functional RRE RNA secondary structure. We show that the RRE region exhibits the highest conservation among all regions of the gene. Previously, significantly increased similarity was found between secondary structures of the RRE in different lentiviruses such as caprine arthritis–encephalitis virus and visna virus.<sup>22</sup> Thus, at least in the case of the RRE, the presence of conserved RNA secondary structure is a major contribution to the conservation of the genomic sequence.

#### *Thermodynamic RNA structure predictions*

Use of thermodynamic calculations as a method for RNA structure predictions is, to say the least, controversial. In an early study, Trifonov and Bolshoi<sup>23</sup> tried to overcome this problem by using multiple alignments for predicting canonical base matching in 5S rRNA. Using this genomic method, they predicted the 5S rRNA secondary structure. To demonstrate the advantages of combining thermodynamic structure prediction by energy minimization with information obtained by phylogenetic alignment of sequences, Luck *et al.*<sup>24</sup> applied their method to predict the structure of the RRE, the tRNA-like element of cytomegalovirus (CMV), and prion protein mRNA from different organisms. The presence of biological function of the RNA secondary structure regions is indicated only if signals of nucleotide conservation overlap with signals of RNA fold conservation and still overlap with conservation of the third position along the translatable open reading frame.

In this work, we present evidence that in the conserved C1 region of the HIV *env* sequence, there is a previously undetected putative RNA secondary structure.

## MATERIALS AND METHODS

### *Construction of a data set*

A sequence data set compilation is crucial to any study of variability,<sup>25,26</sup> especially in the case of human immunodeficiency virus.<sup>26–28</sup> The most obvious problem is to determine the basic criteria for removal of fragments (clones) to derive a representative set of sequences for further analysis. HIV data

were historically collected quite randomly. In many cases features such as the onset of infection, disease status, antiviral treatment, geography, etc., were not annotated, which prevented accurate data selection according to these criteria. Consequently, we used the “bulk” method: to include as many sequences as possible before further database cleaning.

### *The env data set*

We used the data set of 382 complete HIV-1 genome sequences retrieved from the Los Alamos HIV sequence database. A pool of 382 *env* sequences was obtained by extraction of *env* genes encoding the gp160 protein. This pool was subjected to cleaning and redundancy control in the following steps.

1. Cleaning by size and valid features of coding sequences: Only sequences larger than 2000 bases, starting with ATG and ending with a valid stop codon, with coding region length modulo 3, were kept (234 sequences).

2. gp120–gp41 cleavage validity: Fourteen sequences lacking the consensus REKR fragment in the amino acid sequence for the cleavage site between the gp120 and gp41 proteins were discarded (220 sequences left).

3. Redundancy control: The goal of redundancy reduction is to obtain a data set in which the similarity between any two sequences is below a certain threshold. For this purpose we used CLEANUP, a fast computer program for removing redundancies from nucleotide sequence database.<sup>29</sup> A subset of 180 sequences with pairwise similarity of less than 90% nucleotide identity was selected.

4. Alignment consistency: Two *env* sequences with a few nonmutational insertions, probably resulting from recombination, caused multiple gaps in the alignment. Removal of these two sequences resulted in improved alignment, leaving a final pool of 178 sequences. A list of accession numbers of these 178 sequences is shown in Table 1. Automatic multiple alignment is a compromise reflecting gap creation and gap extension penalties, and functions poorly in variable regions. Our data set contains high-variability regions (V1, V2, and V3); therefore manual refinement of alignments for further analysis was inevitable.

### *Information content of multiple alignments*

To find the relevant signal in the multiple alignments we used the Kullback–Leibler measure of information content.<sup>30,31</sup> This measure quantifies the contrast between an actual and an expected distribution of amino acids and nucleotides, respectively. This is used to calculate the total amount of information per position in the alignment. In general, the information content for position *i* in the alignment may be written as

$$I_i = \sum_k q_{ik} \log_2 \frac{q_{ik}}{p_k} \quad (1)$$

where index *k* either sums over all possible amino acids or all possible nucleotides, when in both cases *k* may mean a gap as well. Thus *k* varies from 1 to 5 for nucleotides (A, C, G, T, –) and from 1 to 21 for amino acids. The quantity *q<sub>ik</sub>* is the observed fraction of amino acid/base/gap *k* at position *i*, and *p<sub>k</sub>* is the expected value. Neglecting gaps and using a uniform background distribution reduces this measure to the Shannon infor-

TABLE 1. ACCESSION NUMBERS OF HIV-1 GENOME SEQUENCES

AB032740	AF042105	AF075719	M68894	U39259
AB032741	AF042106	AF076474	M93258	U39362
AF003887	AF049494	AF076475	M95292	U43096
AF004394	AF049495	AF076998	U04908	U43141
AF004885	AF063223	AF077336	U09664	U46016
AF005495	AF063224	AF082394	U12053	U50208
AF015919	AF064699	AF082395	U12055	U51188
AF025749	AF067154	AF082486	U23487	U51190
AF025750	AF067155	AF107771	U32396	U54771
AF025751	AF067156	AF128126	U34603	U63632
AF025753	AF067157	AJ006022	U36859	U71182
AF025754	AF067158	AJ006287	U36860	U82991
AF025755	AF067159	AJ245481	U36865	U82992
AF025756	AF069139	D10112	U36866	U82993
AF025757	AF069670	K02007	U36867	U84819
AF025758	AF069671	K02013	U36869	U84854
AF025759	AF069672	K03454	U36870	U86768
AF025760	AF069673	L02317	U36873	U86769
AF025761	AF069932	L07082	U36875	U86772
AF025762	AF069935	L08655	U36877	U86773
AF025763	AF069937	L20571	U36879	U86774
AF025764	AF069939	L22953	U36880	U86781
AF035532	AF069941	L39106	U36881	U88822
AF041125	AF069943	M17449	U36882	U88823
AF041126	AF069945	M19921	U37270	U88824
AF041127	AF069947	M22639	U39233	U90933
AF041128	AF070705	M26727	U39237	U90934
AF041130	AF070709	M27323	U39238	X04415
AF041132	AF070710	M38427	U39239	X96522
AF041133	AF070711	M38429	U39240	X96526
AF041134	AF070713	M38430	U39242	Y13716
AF041135	AF071473	M38431	U39245	Y13718
AF042101	AF071474	M62320	U39250	Y13719
AF042102	AF075701	M65024	U39253	Z11530
AF042103	AF075702	M66533	U39254	
AF042104	AF075703	M68893	U39256	

mation measure used by Schneider and Stephens<sup>32</sup> to compute sequence logos. For gaps we used the background probability  $p_{-} = 1$  as discussed in the references above. In the case of the Shannon information, the maximum information in bits per position is  $\log_2 2 \approx 4.3$  for amino acids and  $\log_2 4 = 2$  for nucleotides. The quantifier  $p_k$  used here for nucleotides is  $p_N = 0.25$ .

#### Analysis of sequence alignments

The multiple alignments were in all cases made by CLUSTAL W.<sup>33,34</sup> The program was used to search for overall variability. One should be aware that a variable region (a region in the alignment being less conserved in sequence) will by nature contain gaps. Hence, when computing the information content of the alignment, we search for low-entropy regions with the least possible number of gaps.

#### Analysis of multiple alignments of RNAfold predictions

Once profiles of variability have been obtained, further analysis is performed at the RNA level by searching for RNA secondary structure candidates. Such candidates are provided by

the programs Mfold, version 3.0,<sup>35-37</sup> and RNAfold, Vienna RNA package version 1.4.<sup>38,39</sup> The output of RNAfold is a string of dots and brackets, where “.” (i.e., a dot) represents a base not involved in complementary contacts, and “(” and “)” (i.e., brackets) represent 5'- and 3'-complementary bases of the stem (5'ds and 3'ds; ds, double stranded). For instance, the string (((.....))) would refer to a stem-loop structure with four base pairs in the stem and five nucleotides in the loop. Gaps were inserted into strings of dots and brackets according to their positions in the aligned nucleotide sequences. Thus, the alignment of the RNA structures was made in order to reveal RNA structure motifs common for many of the *env* sequences. Conservation of an RNA secondary structure element at position  $i$  is calculated again, using a relative information measure:

$$I_i = \sum_{k = \text{“-”, “.”, “(”, “)”}} q_{ik} \log_2 \frac{q_{ik}}{p_k} \quad (2)$$

where the index  $k$  runs over all RNA secondary structure elements and gaps. The quantities  $q_{i,(}$ ,  $q_{i,.)}$ ,  $q_{i,-}$ , and  $q_{i,-}$  are the observed fractions of 5'ds, 3'ds, ss, and gaps corresponding to position  $i$ . The expected probability for the base to belong to

single-stranded sections at every position  $i$ ,  $p_i$ , has been found empirically to equal 0.5, and consequently ds probabilities  $p_c$  and  $p_g$  are equal to 0.25, and the gap background  $p_{-}$  is equal to 1.

#### Back translation and randomized back translation

To demonstrate the significance of the analysis of the nucleotide sequence and the analysis of RNA fold multiple alignments, two back translation procedures were performed. The first procedure was carried out by correct adding of gaps to the original database (DNA) sequence according to the gaps appearing in the corresponding amino acid sequence. The other procedure involved randomized codon editing within the boundaries of each amino acid. A Perl program picked a random codon from each amino acid codon repertoire according to the original amino acid sequence and added it to the back-translated DNA sequence. At each position in which a gap appeared in the amino acid sequence, three gaps were added to the generated randomized DNA sequence.

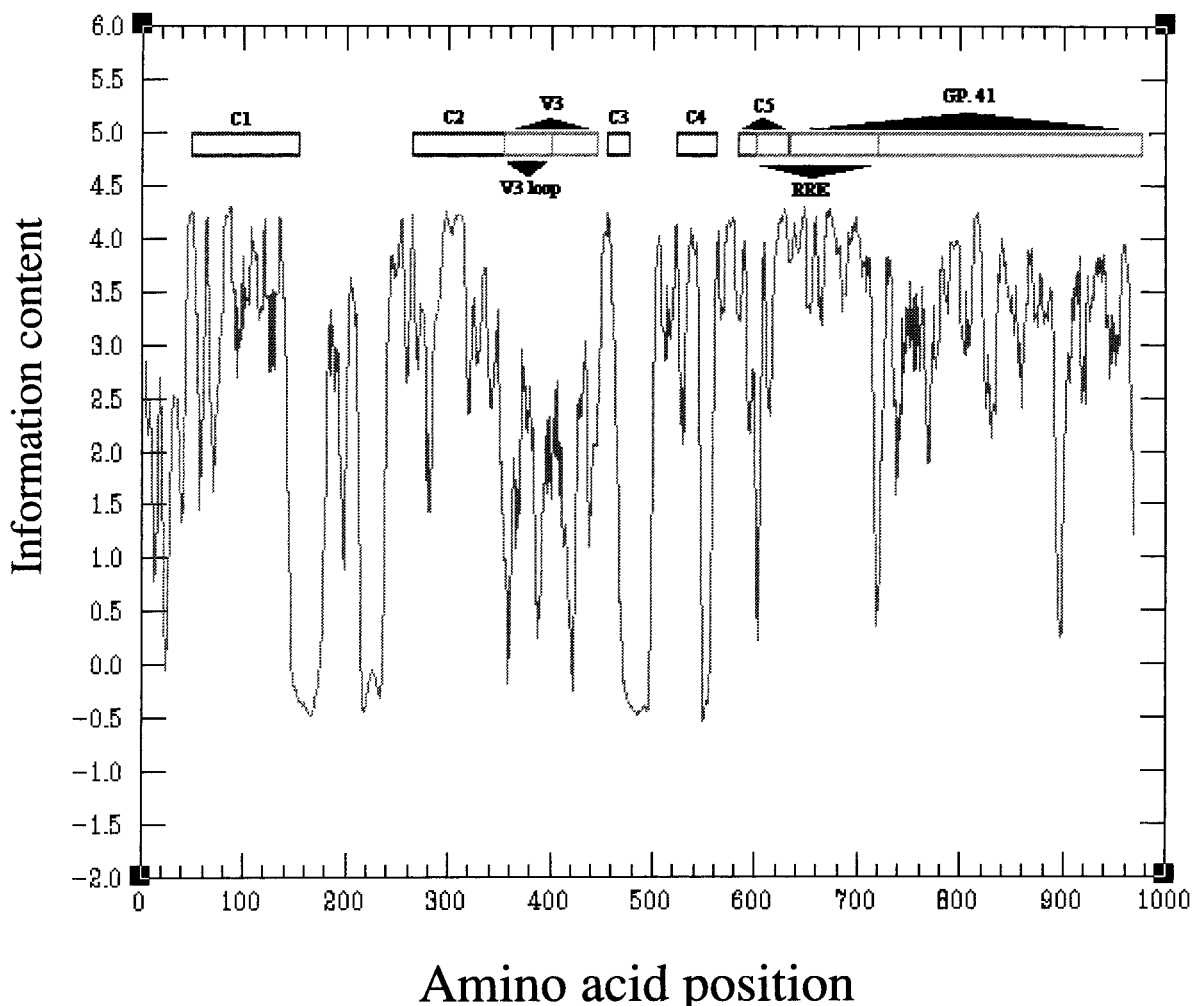
#### Visualization of RNA folds

To illustrate the most conserved features of the putative RNA secondary structures in the C1 subregion we presented two RNA fold predictions: a common C1 RNA secondary structure and one specific representative of C1 fragments. The putative common C1 RNA fold was predicted by GeneBee,<sup>40</sup> a server for RNA secondary structure prediction based on sequence alignment. The aligned C1 sequences were extracted from the aligned *env* sequences described previously. To illustrate the folding of representative sequence, we present C1 RNA structure folded by the Mfold program, using a C1 extract from the sequence HXB2.

## RESULTS

#### Conservation of the *env* gene

For the analysis of sequence conservation we used the information content measure. The *env* gene is characterized by



**FIG. 1.** Conservation of amino acids along different regions in the *env* gene of HIV-1. The conservation is measured by Kullback–Leibler information content computed from protein multiple alignments according to Eq. (1). The curve showing the information content was smoothed by a running average (window size = 6). A map of the subregions is shown above the plot.

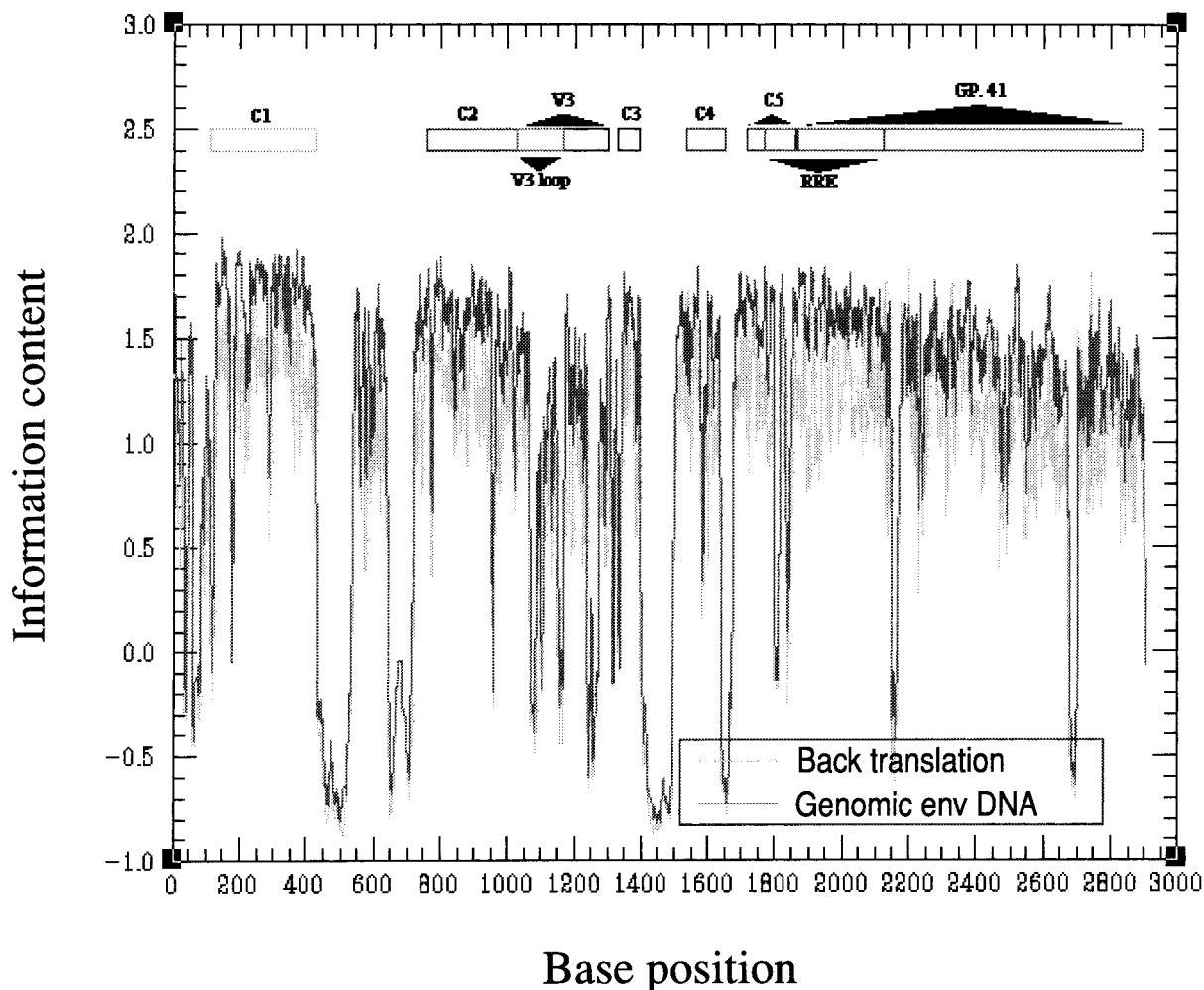
alternating conservative and variable regions.<sup>6</sup> A map of information content distribution along the *env* gene also shows similar variation. Remarkably, our conserved and variable regions coincide well with the published positions of C1–C5 and V1–V5 *env* regions.<sup>6,9</sup> The maps of information content distribution are shown in Figs. 1–5. Here, the known locations of conserved regions (C1–C5) and the third variable region (V3) of the *env* gene are indicated.

In Fig. 1, we present the information content results of multiple alignment of 178 Env protein sequences. Note that the gp41 part of the *env* gene is highly conserved except for a small region around position 900 in the Env amino acid enumeration. The most conserved part of the whole gp160 protein is the border region at the gp120–gp41 cleavage site, where the RRE is located.<sup>15–18,41</sup> One likely explanation of the extraordinary conservation is the joint evolutionary pressure both on the RNA secondary structure of the RRE and on the protein se-

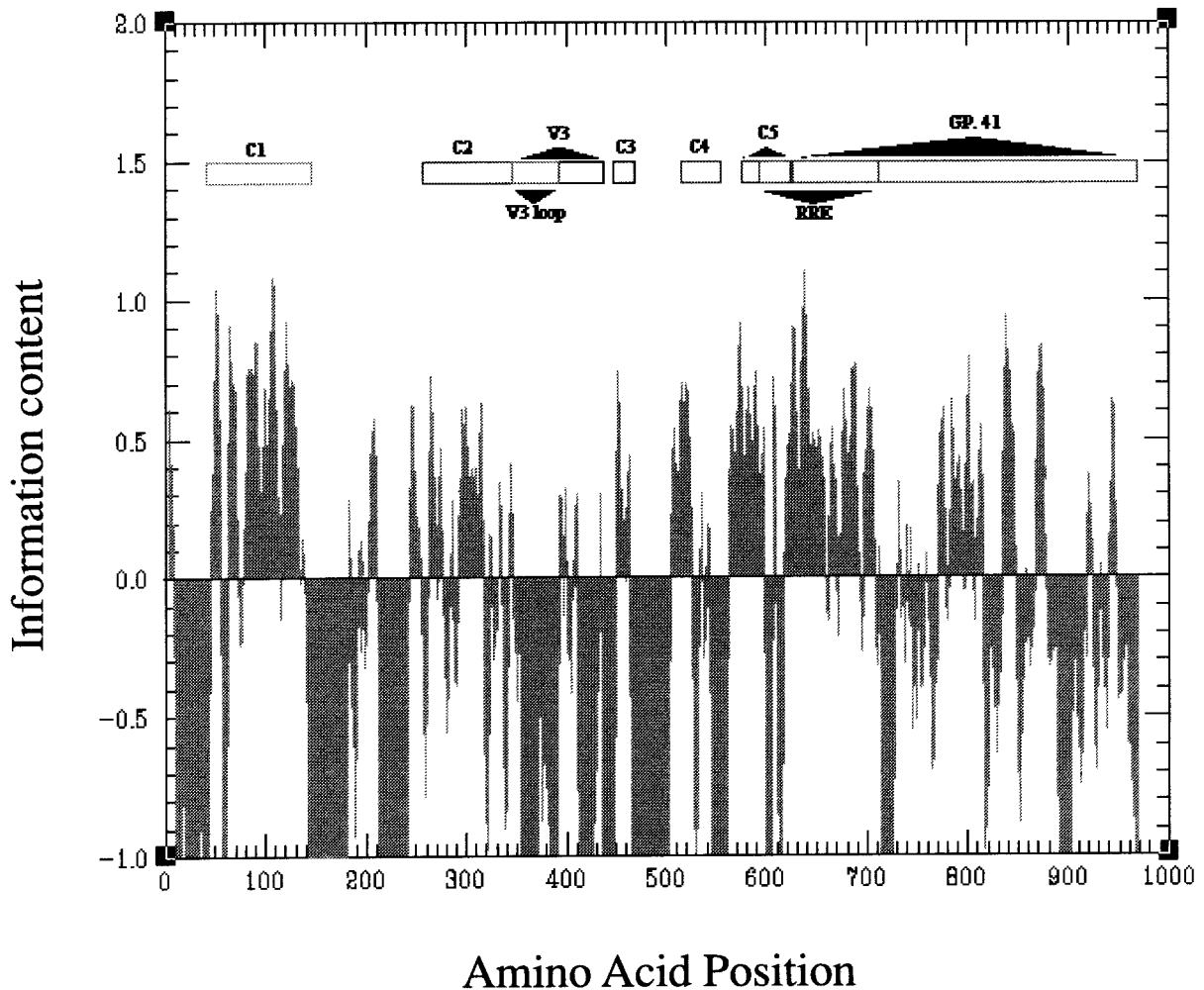
quence, for preservation of the gp120–gp41 consensus cleavage motif.

*RNA secondary structure conservation along the env gene*

To discover DNA codes different from the obvious protein coding we compared conservation patterns at the amino acid and nucleotide levels. We did not perform DNA multiple alignment per se. Rather, we used the back-translation technique. The results of the back translation are presented in Fig. 2. Here both randomized back translation and actual coding sequence are reflected. The randomized back translation was made by replacing every amino acid with one of the corresponding randomly selected codons (equiprobable within the group). Although, as Fig. 2 shows, the randomized back-translated DNA sequences are more variable than the genomic DNA, the con-



**FIG. 2.** DNA conservation of *env* DNA of HIV-1. To calculate the information content of *env* DNA sequences we used a back-translation technique: randomized back translation and correct transformation. The correct back translation was made by replacing every amino acid by the corresponding codon from the real, known gene sequence; the randomized back translation was made by replacing every amino acid by one of the corresponding randomly selected codons (equiprobable within the group). The curve showing the information content was smoothed by a running average with a window size of six. The dark gray line describes the DNA conservation of the *env* sequences retrieved from the original database, whereas the light gray line describes the conservation of the back-translated *env* sequences. A map of the subregions is shown above the plot.



**FIG. 3.** Conservation of the third codon position along the *env* gene. Nucleotides at the third position of each codon of the *env* gene had been extracted from the multiple alignment data sets of the DNA sequences of the HIV-1 *env* gene. The conservation of the third position was computed by using information content as a measure. We present the normalized plot of the information content. For the normalization, we used the average and standard deviation values that we obtained from ungapped alignments to avoid disruptions of the results. The normalized information content was smoothed by a running average (window size = 6). A map of the subregion is shown above the plot.

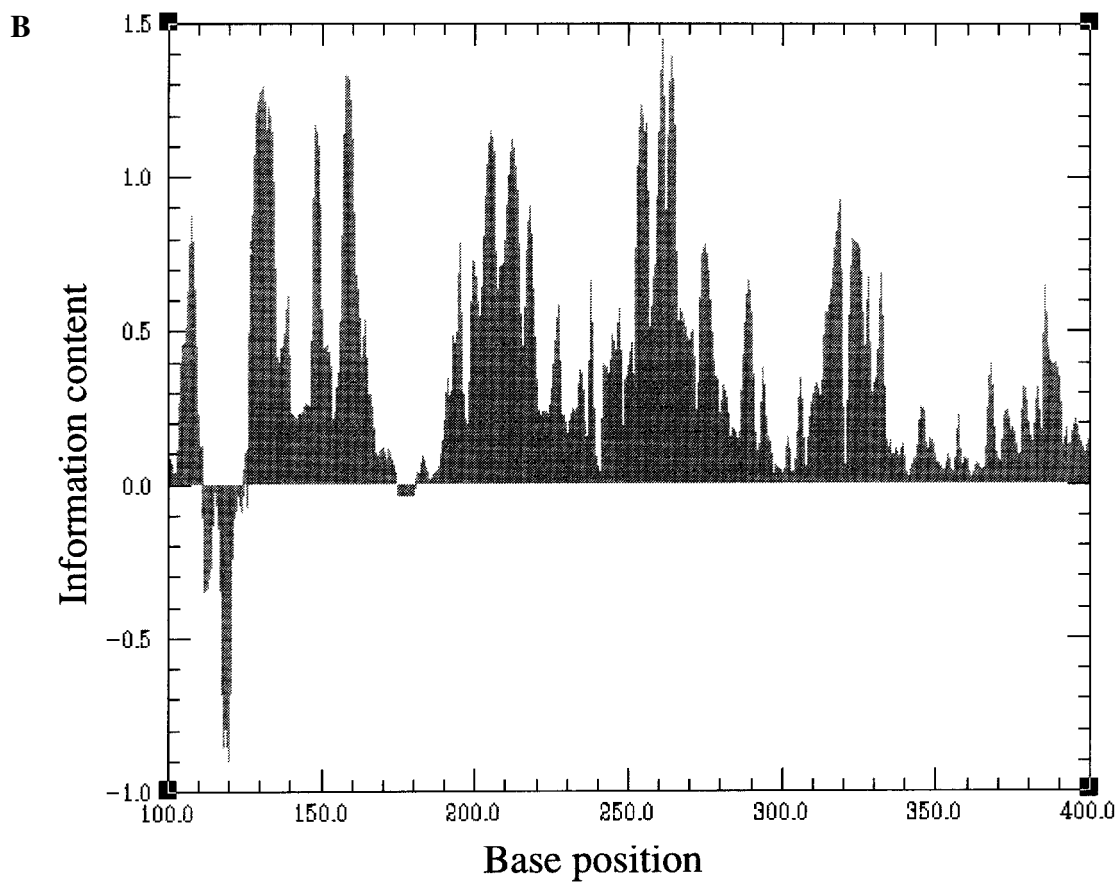
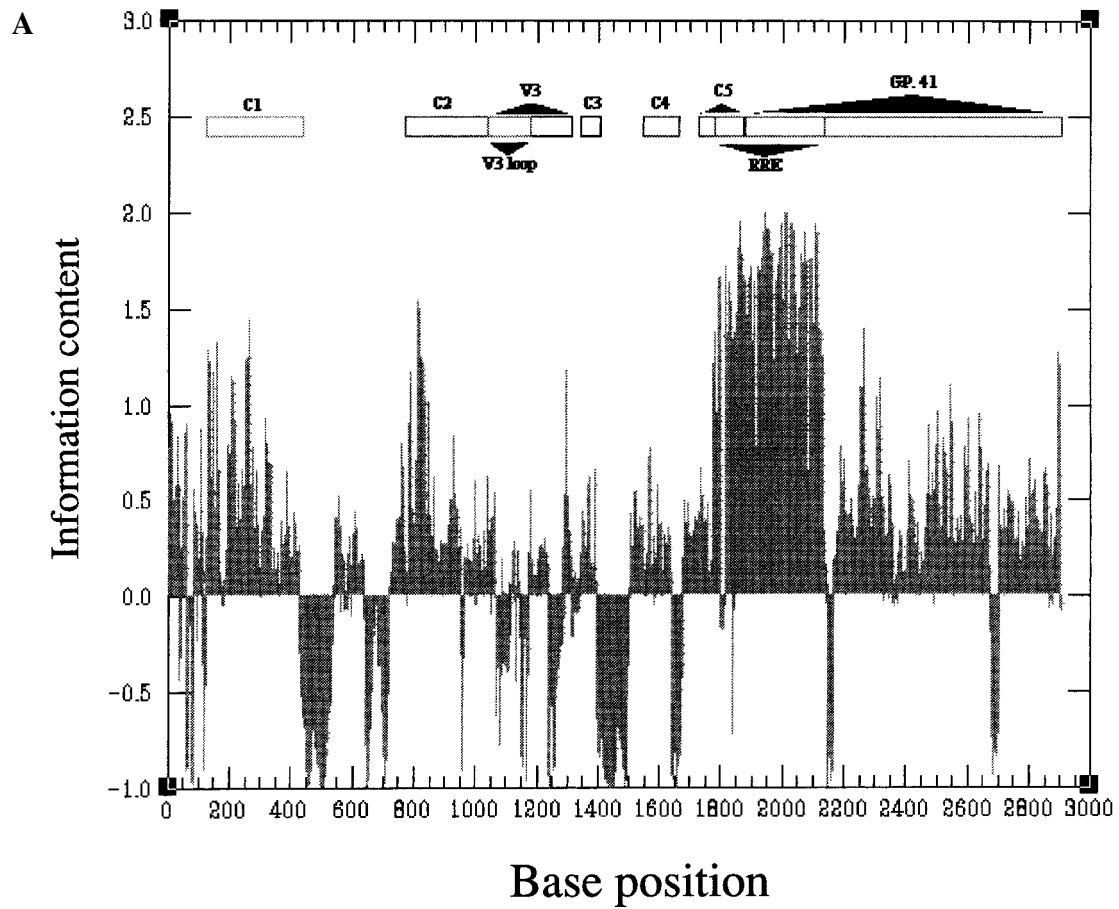
conservation pattern of the back-translated sequences still follows the pattern of the real genomic DNA. This, apparently, reflects the conservation of the first and second positions in each codon.

Degeneracy of the genetic code leads to higher variability of the third codon position in genomic sequences as compared with randomized back translation. Figure 3 shows the information content of the third position only. These data were normalized in order to reveal the regions where the third position is preserved above expectation. Two regions demonstrate third position preservation above the expected level: the RRE and C1. Additional regions, such as C2 and V3, dem-

onstrate similar conservation behavior but the former two regions show the highest values. Such an outcome was expected for the RRE region with its important RNA secondary structure. However, information content maximum in the C1 region is unexpected and indicates the presence of another, hidden message in this sequence, beside the messages for amino acid translation.

A number of secondary structures that play a functional role during the virus life cycle of HIV-1 have been determined. To the best of our knowledge, the RRE is the only known functional RNA secondary structure that is located within the *env*

**FIG. 4.** (A) Conservation of predicted RNA folds along the *env* gene. RNA secondary structures were predicted by the Vienna package. The outputs of these predictions were aligned and the gaps were inserted according to the amino acid multiple alignments. Conservation of RNA secondary structure at every position was computed as the information contribution of stem or loop relative to their expected distribution according to Eq. (2). (B) Conservation of predicted RNA secondary structure in the C1 region of the *env* gene. Shown is an enlargement of the C1 region from (A).



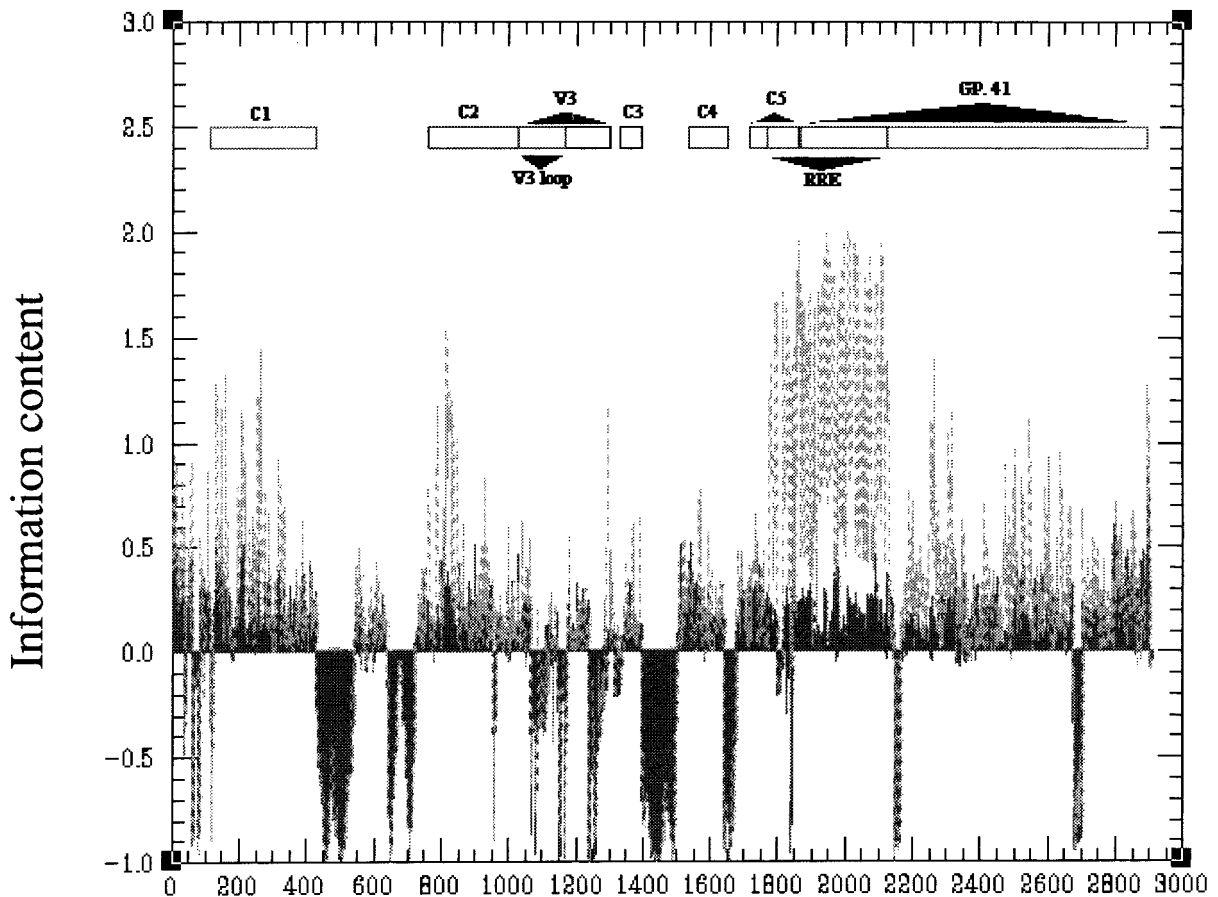
gene. The question is, are there other functional RNA secondary structures within the *env* gene?

It is generally agreed that comparative methods are most reliable for determination of RNA secondary structure common for a set of related RNA sequences.<sup>42</sup> As a first simple test, we took the above-described CLUSTAL W alignment, and calculated energetically optimal RNA secondary structures for each sequence. The distribution of RNA secondary structure conservation along the *env* gene is shown in Fig. 4A. There are few regions where putative RNA secondary structures are highly conserved; most prominent is the RRE region located around positions 1778–2137, with the gp120–gp41 cleavage site at position 1818. The previously undetected C1 secondary structure conservation is clearly seen in Fig. 4B. In Fig. 4A and B one can also observe additional regions with potentially conserved RNA secondary structures, specifically C2 and two spots

downstream of the RRE, around positions 2200 and 2450. These regions are probably 5' and 3' strands of one and the same RNA fold. Indeed, more detailed analysis shows that in the C2 region mainly 5' stems of RNA secondary structure are conserved, whereas downstream of the RRE a region of 3'-stem conservation is situated.

#### *A novel RNA secondary structure within the C1 region*

The visualization of "RNA secondary structure alignments" in Fig. 4A, which reveals the RRE structure in finest detail, points also to other possibly conserved RNA stem-loop structures. Indeed, this simple approach identifies RNA secondary structures that are common to practically all versions of the *env* gene in the C1 region (positions 124–438).



**FIG. 5.** Conservation of hypothetical *env* RNA secondary structure obtained by back translation of *env* amino acid sequences. *env* amino acid sequences from the data set were back translated, whereas the codons were randomly selected within the limits of each amino acid codon pulled. Obtained randomized nucleotide sequences were folded by RNAfold software. The back-translated RNA secondary structures predicted by the Vienna package were realigned and gaps were inserted. Conservation of RNA secondary structure at every position was computed by using the information content as a measure, in exactly the same way as described in Fig. 4A for the real RNA sequences. The dark area demonstrates RNA secondary structure conservation of the back-translated *env* sequences, whereas the fainter area describes the RNA secondary structure conservation of the *env* sequences retrieved from the original database (as in Fig. 4A).

Another piece of evidence to support the existence of a common RNA fold in the C1 region comes through comparison of the RNA secondary structure conservation predicted inside the genomic *env* C1 sequences with the randomized back-translated C1 sequences as demonstrated in Fig. 5. It is clear that the conservation pattern of the randomized back-translated RNA second structure plot is flattened, compared with the information content of the natural HIV-1 RNA secondary structure, and the respective peaks of C1 and RRE are low. The most common features of the conserved RNA structures could be easily reconstructed from the pattern in Fig. 4B, and confirmed by the common RNA fold (Fig. 6A). Figure 6 shows these folds in detail. The consensus RNA fold presented in Fig. 6A can be further compared with a representative individual HXB2 RNA fold in Fig. 6B. The fold consists of two long imperfect stems. The first stem is located at positions 1 to 150 in C1 or at positions 124 to 274 in the whole *env* gene. In Fig. 6A the fold has a complex structure, branching at position 30 (or 154). The second structure is located at positions 150 (174) to 300 (324), and has a long stem with two major bulges. The common structure in Fig. 6A is similar to the predicted structure in genome HXB2 of Fig. 6B. The numbering of the positions in Fig. 6A and B is different because in case of a common structure (Fig. 6A) the numbering includes gaps. One should not expect the HXB2 predicted structure and common structure to be identical. The common structure contains gaps required by the multiple alignment procedure; it should approximately reflect HIV-1 subtype, which is likely to have its own version of the stem-loop structure. It may well be that individual structures reflect alterations caused by changes in physiological conditions. One interesting observation is that the UAAA loop at the top of the first structure in the HXB2 prediction (positions 174–177 or 76–79 in Fig. 6B) is conserved, as is the GGAUUA sequence (positions 308–315 or 210–217 in Fig. 6B) at the top loop of the second structure. The difference between the common structure and the HXB2 structure may be explained by the specifics of the GeneBee software.<sup>40</sup> This algorithm tends to fold the most conserved bases first, thus sometimes forcing the above-described motifs into the stems. We consider the common secondary structure in Fig. 6A as more reliable because its stem elements are based on the multiple alignment of 178 sequences. It is important to emphasize that our conclusion about the biological relevance of the C1 RNA secondary structure is based largely on the conservation of the third position and the multiple alignments of the secondary structure.

These results are also supported by a computed covariance analysis we have performed for this region (data not shown).

## DISCUSSION

The aim of our study was to search for factors constraining the variability of the genome primarily in relation to conserved RNA secondary structures located in the *env* gene. We were able to predict a new, apparently functional (and as a result, conserved), RNA secondary structure. This was done by a simple combination of thermodynamic and genomic approaches. This method allowed checking each factor separately, aligning it in order to enhance the major and important signals, and comparing it with the other aligned factors.

### *Putative RNA secondary structures within the env gene*

Straightforward alignment of thermodynamically optimal RNA folds of the entire sequence indicates the existence of several common RNA stem-loop structures. One of them is the well-known RRE structure, whose topology can be reconstructed in detail from the RNA secondary structure alignment (Fig. 4A; see Materials and Methods). Two other highly conserved RNA secondary structures are located in the C1 region (Figs. 4B and 6), with a level of conservation almost as high as for the RRE. Despite the high conservation of the C1 region, optimal RNA secondary structures for individual sequences possess potentially significant differences. RNA secondary structures in the other constant region of *env* (C2–C5) are less conserved than in C1, although more sophisticated analysis of potential RNA structures may confirm their statistical significance.

### *Back translation*

Support for the assumption of the existence of functional RNA secondary structures along the C1 region is provided by the difference between the genomic RNA secondary structure and the randomized back-translated RNA structure on the one hand, and the high conservation of the third codon position in this region on the other hand. The degeneracy of the genetic code in such that the third codon position only rarely changes the amino acids. Hyperconservation of the third position in a coding sequence implies the existence of another biological message besides the protein coding. RNA secondary structure is indeed such an additional message. The hyperconservation of the third position implies a functionality of the encoded RNA secondary structure. The striking difference between RNA secondary structures derived from the randomized back-translated *env* sequences and the real genomic *env* sequences is far more prominent than differences in those sequences and points, indeed, to the important contribution of the third position to the formation and conservation of C1 RNA secondary structure.

### *The C1 RNA secondary structure*

The role of the C1 region in the gp120 glycoprotein had been indicated by Wyatt *et al.*<sup>43</sup> as part of the gp41–gp120 interactive region together with the C5 conserved region. Our natural assumption is that a significant portion of amino acids in the C1 region does not participate in the gp120–gp41 protein–protein interaction. Nevertheless, the entire C1 nucleotide sequence is well conserved. We explain it by the existence of the conserved RNA fold. The actual functional role of the conserved RNA secondary structure remains to be elucidated.

Unlike the C1 RNA structure, both the Rev protein-binding element (RBE) and TAR element have well-studied specific protein-binding features. A nonspecific binding feature of these secondary structures to the chromosomal protein HMG-D has been discovered.<sup>44</sup> The high-mobility group protein HMG-D is known to bind preferentially to DNA of irregular structure, with little or no sequence specificity. HMD-D can also bind to double-stranded RNA. It seems likely that this feature of nonspecific binding to HMG-D plays a role in the development of HIV-1 in the host cell. It may well be that the stem-and-loop



RNA structures in C1 are an evolutionary design for nonspecific binding. Indeed, a closer look at the loops of the C1 RNA structure and the RBE loop reveals certain similarities. For example, the AUUUAU consensus at positions 19–23 (or 143–147 in the *env* gene) at the first bulge of the common structure is similar to the stem-loop IID AUUUAU of RRE. (Remarkably, the first bulge of the HXB2 predicted secondary structure keeps the AUUUAU site as well.)

### Overlapping messages

The question concerning whether highly conserved genetic codes tend to overlap, or whether this overlapping of genetic codes creates highly conserved sequences, has no answer yet. The two major RNA secondary structures detected in the *env* gene sequence are located in regions of conserved amino acid sequences: the RRE region and the C1 region. A possible explanation for this phenomenon is that each of the superimposed biological signals reduces the freedom of its overlapped companion to vary. Indeed, the existence of RNA secondary structure in these regions produces additional constraints reducing the potential of the amino acid sequence to change. Another speculation is that overlapping of various codes is advantageous from the point of view of “venture distribution.” That is, if a given region is biologically important and the preservation of its sequence is critical, then the appearance of another important code in this region, superimposed on the first, is of evolutionary preference, in order to minimize the number of vulnerable sites in this sequence. Thus, one would expect to find more hidden codes within the sequences of highly conserved genes.

This is definitely not the first case of overlapping messages in the HIV-1 genome. Overlapping of genes such as *pol/vif*, *env/tat2*, and *env/vpu* is well known. One of the interesting findings is that in most cases the regions containing superimposed messages have characteristics of RNA secondary structures. Overlapping the RNA secondary structure with protein-coding sequence such as the RRE is also well known. Characterization of RNA secondary structures along the HIV-1 genome is still far from completion. Furthermore, the biological function of these structures is yet to be discovered.

In this article we have reported on the finding of a previously undescribed RNA secondary structure in the C1 region by employing informational characterization of the sequences. The same technique indicates more overlapping messages in other regions of the HIV-1 genome, for example, in the *nef* gene.

### ACKNOWLEDGMENTS

We thank Drs. Irit Or, Marilyn Safran, Shifra Ben Dor, and Vered Halifa-Caspi (Bioinformatics and Biological Service, Weizmann Institute of Science) for helpful support. A.B. was partially supported by the Fondation Scata-Rachi, by a grant from the Danish National Research Foundation, and by the FIRST Foundation of the Israel Academy of Science and Humanities. S.B. is supported by a grant from the Danish National Research Foundation.

### REFERENCES

- Hahn B, Shaw G, Taylor M, *et al.*: Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. *Science* 1986;232:1548–1553.
- Holmes E, Zhang L, Simmonds P, Ludlam C, and Brown A: Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proc Natl Acad Sci USA* 1992;89:4835–4839.
- Wong-Staal F, Shaw G, Hahn B, Salahuddin S, Popovic M, Markham P, Redfield R, and Gallo R: Genomic diversity of human T-lymphotropic virus type III (HTLV-III). *Science* 1985; 229:759–762.
- Starcich B, Hahn B, Shaw G, McNeely P, Modrow S, Wolf H, Parks W, Joseph S, Gallo R, and Wong-Staal F: Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. *Cell* 1986;45: 637–648.
- Cheng-Mayer C, Seto D, Tateno M, and Levy J: Biologic features of HIV-1 that correlate with virulence in the host. *Science* 1988;240:80–82.
- Modrow S, Hahn B, Shaw G, Gallo R, Wong-Staal F, and Wolf H: Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: Prediction of antigenic epitopes in conserved and variable regions. *J Virol* 1987;61: 570–578.
- Willey R, Rutledge R, Dias S, Folks T, Theodore T, Buckler C, and Martin M: Identification of conserved and divergent domains within the envelope gene of the acquired immunodeficiency syndrome retrovirus. *Proc Natl Acad Sci USA* 1986;83:5038–5042.
- Andreassen H, Bohr H, Bohr J, Brunak S, Bugge T, Cotterill R, Jacobsen C, Kusk P, Lautrup B, Petersen S, *et al.*: Analysis of the secondary structure of the human immunodeficiency virus (HIV) proteins p17, gp120, and gp41 by computer modeling based on neural network methods. *J Acquir Immune Defic Syndr* 1990;3: 615–622.
- Hansen J, Lund O, Nielsen J, Brunak S, and Hansen J-S: Prediction of the secondary structure of HIV-1 gp120. *Proteins* 1996;25:1–11.
- LaRosa G, Davide J, Weinhold K, Waterbury J, Profy A, Lewis J, Langlois A, Dreesman G, Boswell R, Shaddock P, *et al.*: Conserved sequence and structure elements in the HIV-1 principal neutralizing determinant. *Science* 1990;249:932–935.
- Javaherian K, Langlois A, McDanal C, Ross K, Eckler L, Jellis C, Profy A, Rusche J, Bolognesi D, and Putney S: Principal neutralizing domain of the human immunodeficiency virus type 1 envelope protein. *Proc Natl Acad Sci USA* 1989;86:6768–6772.
- Feng S and Holland E: HIV-1 Tat trans-activation requires the loop sequence within TAR. *Nature* 1988;334:165–167.
- Klaver B and Berkhout B: Evolution of a disrupted TAR RNA hairpin structure in the HIV-1 virus. *EMBO J* 1994;13:2650–2659.
- Selby M, Bain E, Luciw P, and Peterlin B: Structure, sequence, and position of the stem-loop in TAR determine transcriptional elongation by Tat through the HIV-1 long terminal repeat. *Genes Dev* 1989;3:547–558.
- Dayton E, Powell D, and Dayton A: Functional analysis of CAR, the target sequence for the Rev protein of HIV-1. *Science* 1989;246:1625–1629.
- Kjems J, Brown M, Chang D, and Sharp S: Structural analysis of the interaction between the human immunodeficiency virus Rev protein and the Rev response element. *Proc Natl Acad Sci USA* 1991;88:683–687.
- Malim M, Hauber J, Le S-Y, Maizel J, and Cullen B: The HIV-1 Rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* 1989; 338:254–257.

18. Malim M, Tiley L, McCarn D, Rusche J, Hauber J, and Cullen B: HIV-1 structural gene expression requires binding of the Rev trans-activator to its RNA target sequence. *Cell* 1990;60:675–683.
19. Le S-Y, Chen J, and Maizel J: Detection of unusual RNA folding regions in HIV and SIV sequences. *Comput Appl Biosci* 1991;7:51–55.
20. Le S-Y, Malim M, Cullen B, and Maizel J: A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. *Nucleic Acids Res* 1990;18:1613–1623.
21. Huynen M, Perelson A, Vieira W, and Stadler P: Base pairing probabilities in a complete HIV-1 RNA. *J Comput Biol* 1996;3:253–274.
22. Huynen M, Konings D, and Hogeweg P: Multiple coding and the evolutionary properties of RNA secondary structure. *J Theor Biol* 1993;165:251–267.
23. Trifonov EN and Bolshoi G: Open and closed 5S ribosomal RNA, the only two universal structures encoded in the nucleotide sequences. *J Mol Biol* 1983;169:1–13.
24. Luck R, Steger G, and Riesher D: Thermodynamic prediction of conserved secondary structure: Application to the RRE element of HIV, the tRNA-like element of CMV and the mRNA of prion protein. *J Mol Biol* 1996;258:813–836.
25. Brunak S, Engelbrecht J, and Knudsen S: Cleaning up gene databases. *Nature* 1990;343:123.
26. Brunak S, Engelbrecht J, and Knudsen S: Neural network detects errors in the assignment of mRNA splice sites. *Nucleic Acids Res* 1990;18:4797–4801.
27. Korber B, Learn G, Mullins J, Hahn B, and Wolinsky S: Protecting HIV databases. *Nature* 1995;378:242–244.
28. Learn, G., Korber B, Foley B, Hahn B, Wolinsky S, and Mullins J: Maintaining the integrity of human immunodeficiency virus sequence databases. *J Virol* 1996;70:5720–5730.
29. Grillo G, *et al.*: CLEANUP: A fast computer program for removing redundancies from nucleotide sequence databases. *Comput Appl Biosci* 1996;1–8.
30. Cover T and Thomas JA: *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
31. Kullback S and Leibler RA: On information and sufficiency. *Ann Math Stat* 1951;22:79–86.
32. Schneider T and Stephens R: Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res* 1990;18:6097–6100.
33. Higgins D, Thompson J, and Gibson T: Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* 1996;266:383–402.
34. Thompson J, Higgins D, and Gibson T: CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680.
35. Zuker M: On finding all suboptimal foldings of an RNA molecule. *Science* 1989;244:48–52.
36. Zuker M: Prediction of RNA secondary structure by energy minimization. *Methods Mol Biol* 1994;25:267–294.
37. Zuker M and Stiegler P: Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res* 1981;9:133–148.
38. Hofacker I, Fontana W, Stadler P, Bonherffer L, Tacker M, and Schuster P: Fast folding and comparison of RNA secondary structures. *Monatsh Chem* 1994;125:167–188.
39. Jaeger J, Turner D, and Zuker M: Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol* 1990;183:281–306.
40. Brodsky LI, Ivanov VV, Kalaydzidis YL, Lenotovich AM, Nikolaev VK, Ferancuk SI, and Drachev VA: GeneBee-NET: Internet-based server for analyzing biopolymer structure. *Biochemistry (Moscow)* 1995;60:923–928.
41. Charpentier B, Stutz F, and Rosbash M: A dynamic in vivo view of the HIV-1 Rev-RRE interaction. *J Mol Biol* 1997;266:950–962.
42. Chan L, Zuker M, and Jacobson A: A computer method for finding common base paired helices in aligned sequences: Application to the analysis of random sequences. *Nucleic Acids Res* 1991;19:353–358.
43. Wyatt R, Desjardin E, Ulshevsky U, Nixon C, Binley J, Olshevsky V, and Sodrosky J: Analysis of the interaction of the human immunodeficiency virus type 1 gp120 envelope glycoprotein with the gp41 transmembrane glycoprotein. *J Virol* 1997;71:9722–9731.
44. Arimondo PB, Gelus N, Hamy F, Payet D, Travers A, and Bailly C: The chromosomal protein HMG-D binds to the TAR and RBE RNA of HIV-1. *FEBS Lett* 2000;485:47–52.

Address reprint requests to:

Alexander Bolshoy

Genome Diversity Center

Institute of Evolution

University of Haifa

Mount Carmel, Haifa 31905, Israel

E-mail: bolshoy@research.haifa.ac.il