Volume 16 Number 11 1988 Nucleic Acids Research

Stability of RNA stem-loop structure and distribution of non-random structure in the human immunodeficiency virus (HIV-I)

Shu-Yun Le[1], Jih-H Chen[2], Michael J.Braun[3], Matthew A.Gonda[3] and Jacob V.Maizel

Laboratory of Mathematical Biology, Division of Cancer Biology and Diagnosis, National Cancer Institute, National Institutes of Health, Frederick, MD 21701, USA [1]Shanghai Institute of Biochemistry, Chinese Academy of Sciences, Shanghai 200031, China, [2]Advanced Scientific Computer Laboratory, Program Resources, Inc., NCI/FCRF, Frederick, MD 21701 and [3]Laboratory of Cell and Molecular Structure, NCI/FCRF, Frederick, MD 21701, USA

Received October 13, 1987; Revised and Accepted December 11, 1987

ABSTRACT

The stability of potential RNA stem-loop structures in human immunodeficiency virus isolates, HTLV-III and ARV, has been calculated, and the relevance to the local significant secondary structures in the sequence has been tested statistically using a Monte Carlo simulation method. Potentially significant structures exist in the 5'non-coding region, the boundary regions between the protein coding frames, and the 3' non-coding region. The locally optimal secondary structure occurring in the 5' terminal region has been assessed using different overlapping segment sizes and the Monte Carlo method. The results show that the most favorable structure for the 5' mRNA leader sequence of HIV has two stem-loops folded at nucleotides 5-104 in the R region (stem-loop I, 5-54 and stem-loop II, 58-104). A large fluctuation of segment score of the local optimal secondary structure also occurs in the boundary between the exterior glycosylated protein or outer membrane protein and transmembrane protein coding region. This finding is surprising since no RNA signals or RNA processing are expected to occur at this site. In addition, regions of the genome predicted to have significantly more open structure at the RNA level correlate closely with hypervariable sites found in these viral genomes. The possible importance of local secondary structure to the biological function of the human immunodeficiency virus genome is discussed.

INTRODUCTION

The extreme morbidity, prolonged latency period and epidemic proportions of the acquired immune deficiency syndrome (AIDS) have led to intense efforts to elucidate the disease's pathogenesis and to develop effective preventive and treatment measures. Recent studies have revealed that the human immunodeficiency virus (HIV-I), the causative agent of AIDS, has a complex genetic structure (1-8,26). An important clue in understanding the differential regulation of HIV gene expression was provided by the identification of a trans-activator gene(tat) in the HIV genome (9-12). Recently, Cullen(13) presented data suggesting that tat functions in two distinct ways to increase gene expression driven by the HIV long terminal repeat (LTR): first, by enhancing the steady state level

of HIV-specific mRNAs, and second, by relieving the translational block imposed by the presence of HIV R region sequences at the mRNA 5' terminus. Moreover, the conformation of stable stem-loops in the 5' leader of LTR-directed mRNA hasalso been analyzed experimentally (14) using ribonuclease and cobra venom nuclease. The stable stem-loop structure found for the 5' leader of the mRNA suggests a possible role for RNA secondary structure in trans-activation.

This report tests the positional correlation of potentially stable local RNA secondary structures with the genetic map of two human immunodeficiency virus isolates (HTLV-III and ARV). The results show that distinct stable features exist in the 5' and 3' termini of mRNAs and in the env gene at the boundary between the external glycoprotein (EGP) or outer membrane protein (OMP) and transmembrane protein (TMP) coding region. The statistical significance of locally optimal stem-loop structures have been assessed using Monte Carlo simulation. Two highly significant stem-loop structures predicted in the R region of 5' mRNA leader sequence are perfectly consistent with experimental results (14). In addition, several significant open regions at the RNA level are found to correlate with known hypervariable sites in the HIV env gene. These results are discussed with reference to the role of secondary structure and stability of single-stranded RNA in the life cycle of the AIDS virus.

METHOD

RNA sequences of HTLV-III and ARV were derived from available genomic DNA sequences in the GenBank data base. The program RANFOLD was run to compute stability of local secondary structure along single-stranded RNA and assess its statistical significance by Monte Carlo simulation. RANFOLD is programmed in Fortran 77 and run on the Cray Operating System of a CRAY X-MP/24 computer. In this approach, successive overlapping segments along the RNA chain are folded, and the free energy of the optimal secondary structure of each overlapping segment is computed (28). In the calculation, a thermodynamic parameter set proposed by Freier et al(15) for stacking energies and the loop destabilizing free energies was used. Since the parameters for unpaired terminal nucleotides and terminal mismatches are not included in the set, the above parameters can be conveniently implemented in the algorithm.

In order to assess the statistical significance of each "optimal" structure folded in the overlapping segments of the biological sequence,

the free energy of the "optimal" structure is compared to the mean free energy of the "optimal" structures for a large number of random permutations of the sequences (e.g. 200 randomized segment sequences having the same nucleotide compositions as the actual biological segment sequence are folded for the 100 base-long overlapping segment). The randomized sequences are produced by using a random number generator of the CRAY system. Thus, the segment score is calculated as the difference between the free energy of the locally optimal secondary structure of the actual biological sequence and the average optimal free energy from these randomized sequences divided by the SD of the random sample set. The calculation is carried out repeatedly by sliding one base at a time along a RNA sequence. The distribution of segment scores against the positions of the overlapping segments can then be obtained.

The segment score is expressed in SD units. As the optimal free energies are negative, the lower the segment score, the more significant the "optimal" structure of the segment. Thus, regions in the sequence in which distinct and non-random secondary structure can potentially occur are predicted by means of the distribution of the segment score. To further characterize the significance of the stability of the local secondary structure in HTLV-III and ARV genomes, the SAS programs (Procedures MEANS(16)) were used to calculate the average thermal stability of each individual HIV gene or functional region.

RESULT

Stability Map of Local Secondary Structure of HIV-I

Figures 1(a), (1b) and (2a) show the thermal stability maps of local secondary structures of HTLV-III and ARV. Each map was obtained by plotting the free energy of the optimal secondary structure of a local segment in mRNA against the position of the middle base in the segment along the mRNA strand. Fig.1(a) and 2(a) are the maps for 40 and 100-base long overlapping segments of HTLV-III, respectively. Fig.1(b) is a map for 40-base long overlapping segments of ARV. The figures 1(a) and 2(a) clearly indicate that both 5' and 3' termini, have more stable local secondary structures than those of gag, pol and env coding region. Interestingly, a conspicuous fluctuation of the stability distribution of local secondary occurs at the boundary of the OMP/EGP and the TMP coding region. The coding region of TMP has more stable local stem-loop structures than those of the OMP/EGP. Moreover, although about a 5.6%

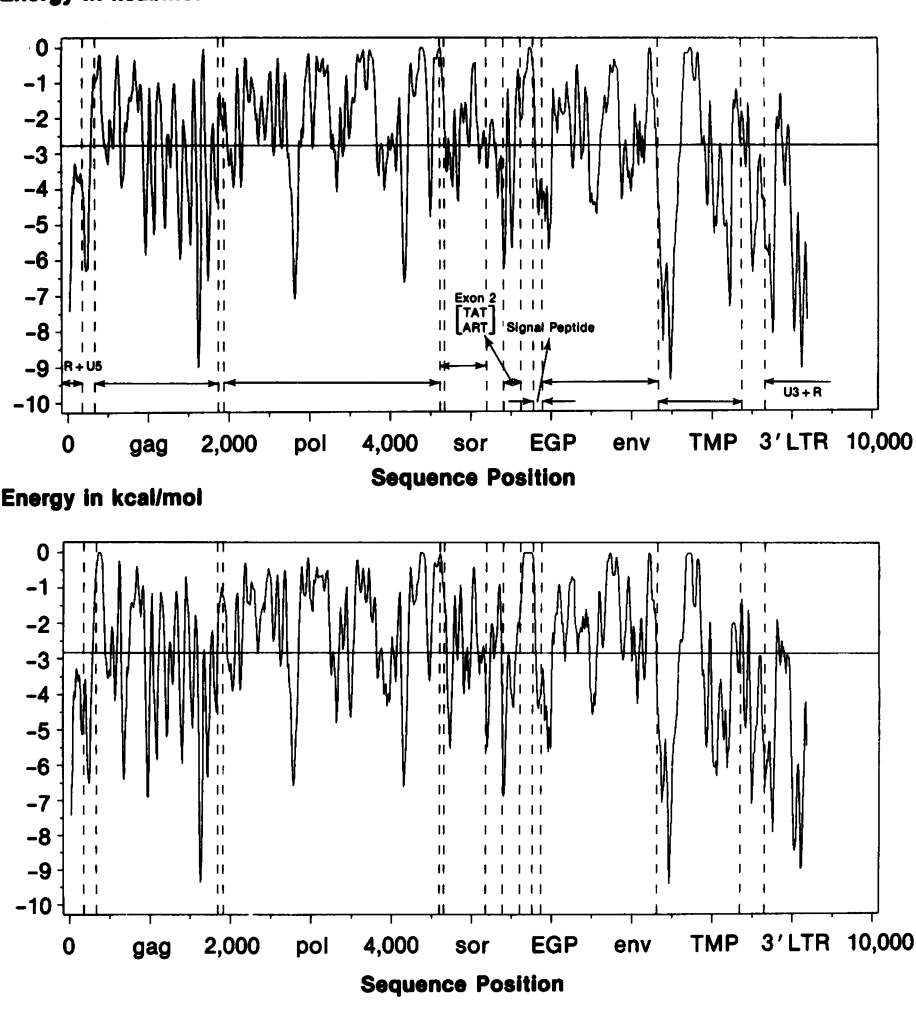

Figure 1. Stability map of local secondary structure of HTLV-III and ARV. The free energy of the optimal secondary structure of 40-base long overlapping segments are plotted against the position of the middle base of the segments along HTLV-III mRNA shown in Fig. 1(a) and ARV in 1(b), respectively. A horizontal line represents the average free energy value of the local optimal secondary structure in HTLV-III or ARV mRNA. The unit of free energy is kcal/mol. The structural viral gene maps were obtained from their respective sequences deposited in GENBANK. In the maps, energies have been averaged in groups of ten positions and then smoothed in overlapping groups of 5 positions along the sequences.

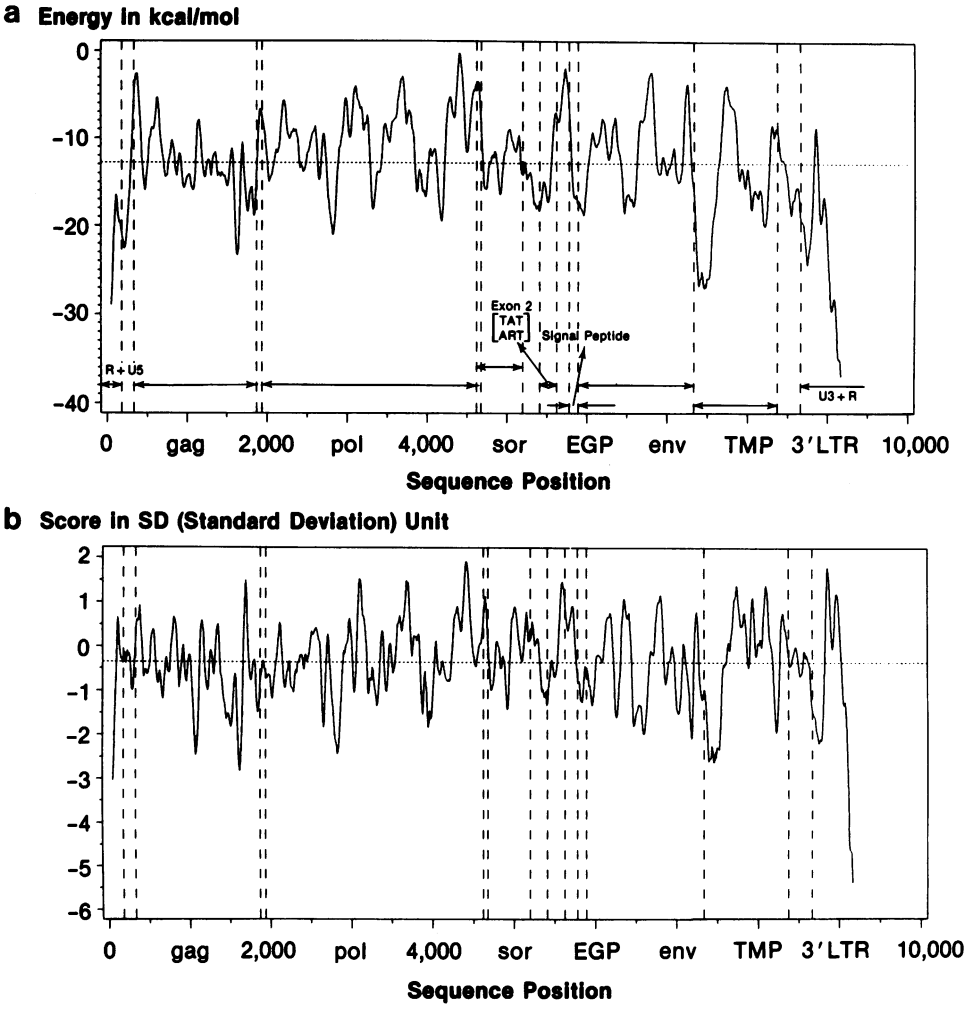

Figure 2. (a) Stability map of local secondary structure of HTLV-III. The overlapping segment length is 100 bases. For further details see the caption to Figure 1. (b) A segment score of the local optimal secondary structure in HTLV-III. The score of a 100-base long overlapping segment is plotted against the position of the middle base in the segment along HTLV-III. The number of random permutations of the sequence segment is 200. The calculations are carried out repeatedly by sliding one base along the sequence. The scores have been averaged in groups of ten positions, and then smoothed in overlapping groups of 5 positions along the sequence. A horizontal line represents the average score value of the overlapping segments in HTLV-III.

difference of nucleotide sequences occurs between HTLV-III and ARV, the stability features mentioned above are conserved.

In order to analyze the positional correlation of thermal stability with the gene structure map of HTLV-III and ARV, the stability values of the overlapping segments were averaged for each genetic unit of HIV. The mean thermal stability for the total genome is taken as a unit of measure in the statistical analyses. Comparisons of the stability of local secondary structures within each individual gene are shown in Figures 3-4. The 5' and 3' terminal regions have the most stable local secondary structures in the genome of HTLV-III or ARV-2. Moreover, the mean thermal stability of the local stem-loop structure of genes is not greatly affected by changing the size of the overlapping segments (window size: 40, 60, and 100 nucleotides). The stability scale of local secondary structure resulting from statistical analyses combined all data of three different sizes of overlapping segment in both HTLV-III and ARV mRNAs depict that the 5' LTR regions R and U5 have the minimal average free energy (the most stable folding region), and the envelope signal peptide has the least stability in the total genome (see Fig.5).

Statistical Significance of the Locally Optimal Secondary Structure

The statistical significance of the locally optimal secondary structure was assessed using Monte Carlo simulation. Fig.2(b) shows the

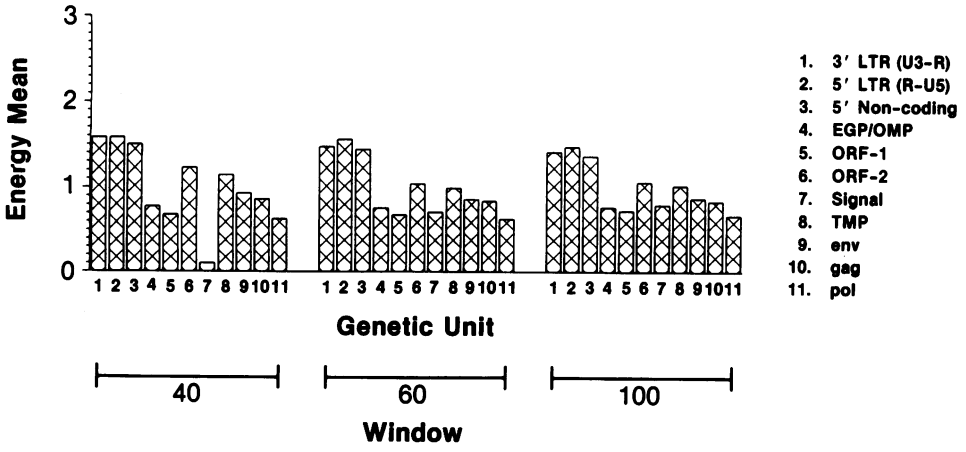

Figure 3. Local stability of individual genes of HTLV-III. Ratios of the energy mean of the local optimal secondary structures in each structural gene of HTLV-III to the average free energy of the local optimal secondary structures in the entire genome were ploted for overlapping segment length of 40, 60, and 100-bases.

variation in segment scores along the HTLV-III genome. Because the optimal free energies are negative, the deeper valleys in Fig.2(b) represent the more statistically significant segment scores. The local secondary structures in the R region of both 5' and 3' termini have distinct statistical significance. The minimal free energies of the calculated optimal secondary structures for the actual sequence segments are at least three SD units less than the average free energies of randomized sequences. Another region with statistically significant secondary structure occurs in the TMP coding region near the boundary of the OMP and TMP of env. The nucleotide sequence of this region is well conserved among several HIV isolates(29). In addition, there are several distinct deep valleys of significance in the coding region of gag. One of them occurs at the boundary between the gag and pol open reading frames. The statistical significance of the locally optimal secondary structure for ARV mRNA was also tested. A plot of the segment scores for the locally optimal secondary structure of ARV revealed the same distinct features as those of HTLV-III(data not shown). Moreover, the thermal stability of the regions mentioned above are also high in HTLV-III (see Fig.2(a)) and ARV.

To further study the potential stem-loop structure of the 5' terminal region(1-332), RANFOLD simulations were run in which the window size of the overlapping segments were varied from 12 to 200 bases in increments of 2 bases and varied from 205 to 320 bases in increments of 5 bases. In the

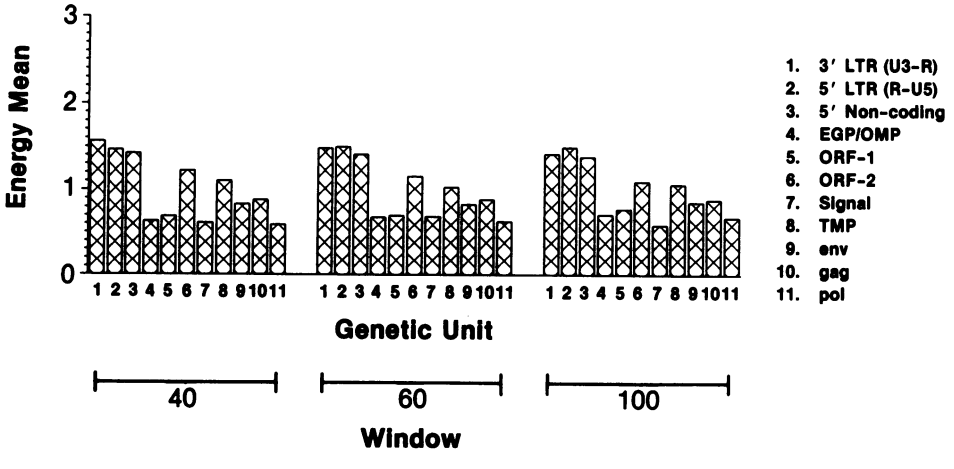

Figure 4. Local stability of individual genes of ARV. For further details see the caption to Figure 3.

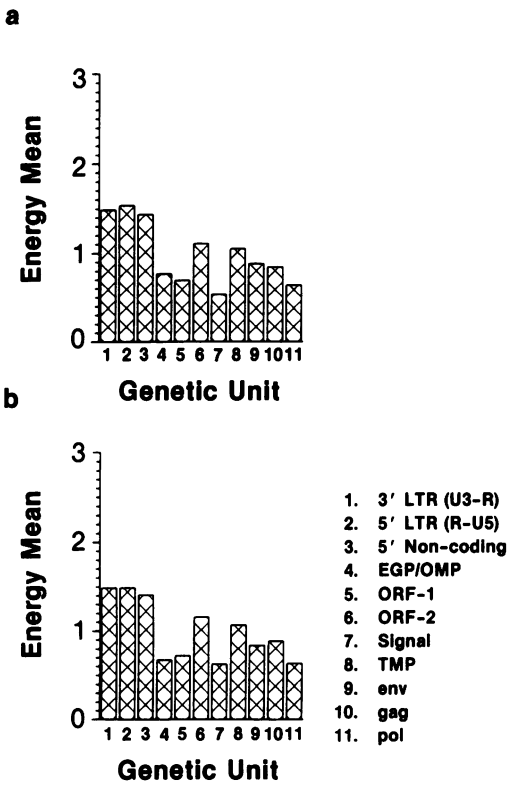

Figure 5. Combined local stability for individual genes of HTLV-III ans ARV. (a) HTLV-III. (b) ARV. The histogram is plotted using average free energy of the optimal secondary structures folded from 40, 60, and 100-base long segments in each genetic unit divided by the energy mean of the optimal secondary structure of those segments in the total genome.

simulations, the sizes of the random sample sets were taken as 80 for window size 12-20, 120 for 22-48, 150 for 50-58, 180 for 60-78, 200 for 80-120, and 100 for 122-320. In the calculations, about 1.28 million randomized sequence segments were processed. The minimal scores obtained for each window size were abstracted and plotted against the window size of the overlapping segments (see Fig.6). The most statistically significant stem-loop structures appear with a window size of about 50 and 100 nucleotides. The optimal stem-loop structures occur in the regions 5-54, and 58-104 (i.e. region 5-104). The scores are -6.01 and -6.90 SD units respectively. These predictions are perfectly consistent with the empirical results from analyses using ribonuclease and cobra venom nuclease

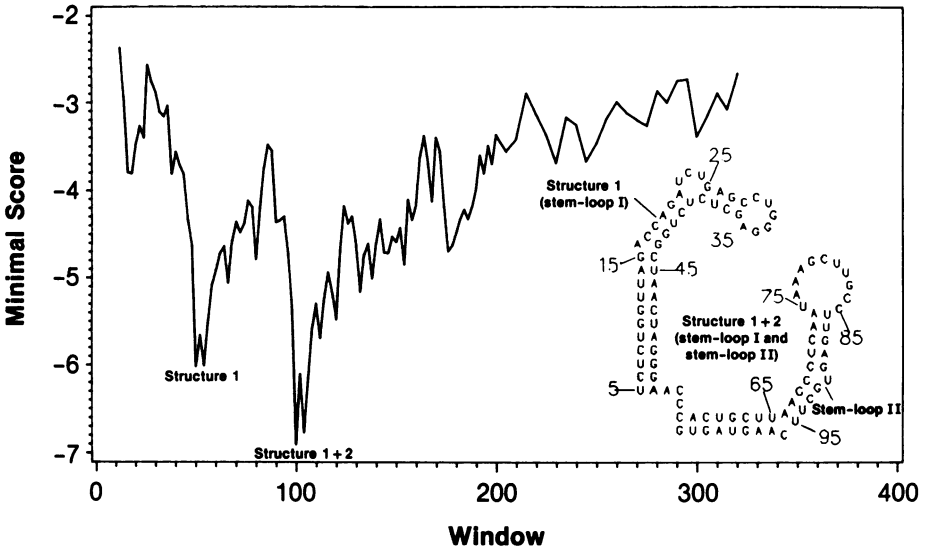

Figure 6. Distribution of significant secondary structures in the leader sequence of HTLV-III mRNA(1-332) under different segment sizes. For each specific segment size(window size), the calculation is carried out by sliding two bases (window size: 12-100), three bases(window size: 102-160), four bases(window size: 162-180), and five bases(window size: 182-320) along the leader sequence. The calculation is then carried out repeatedly by increasing two bases to the segment of from 12 to 200-base long sizes and five bases to the segment of from 205 to 320 along the leader sequence. The number of random permutations of the sequence segments are 80 for 12-20 bases of segment size, 120 for 22-48, 150 for 50-58, 180 for 60-78, 200 for 80-120, and 100 for 122-320. The structures 1 and 1+2 correspond to two distinct deep valleys. The nucleotide number labeled in the structure 1+2 (or 1) is taken from its position in HTLV-III mRNA. The structure 1+2 consists of two stem-loop structures, stem-loop I (i.e. structure 1) and stem-loop II in R region of 5' terminal.

digestions(14). Our analyses show that the computer prediction from thermal stability considerations and statistical analyses can be used to identify local secondary structures which can be verified experimentally by nuclease cleavage studies. Of great practical interest is the fact that the statistical significance of both structure 1 and structure 1+2 diminished very quickly as the segment size increased beyond that where the full structure was first able to form. This indicates that potentially relevant structures can easily be masked by the inclusion of flanking sequences in the calculation. Proper window size appears to be critical for the sensitivity of the statistical analysis and several window sizes should normally be tested.

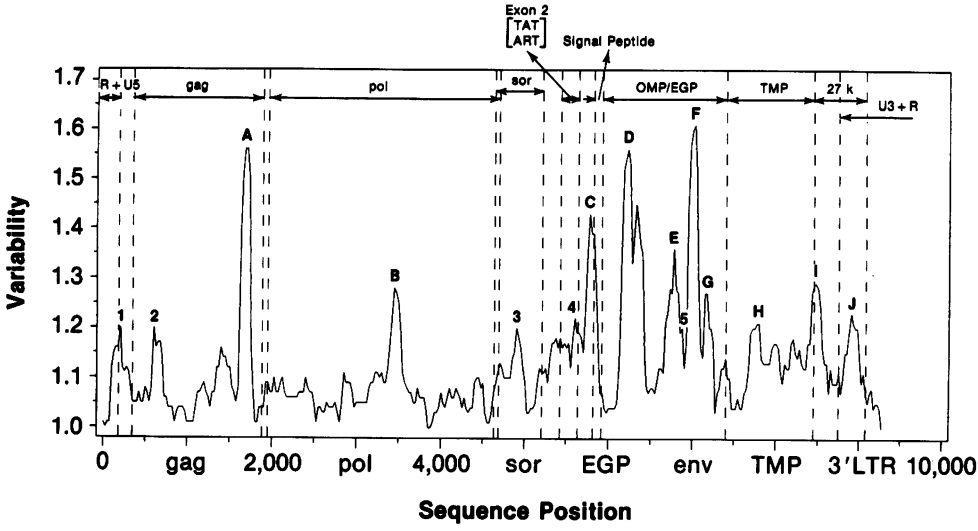

Figure 7. Variability map of nucleotides at each position of AIDS virus sequences from HTLV-III (BH10), ARV, LAV, and HTLV-III (H9) strains. The variabilities in successive groups of 25 positions are averaged and smoothed in overlapping groups of five positions along the sequences. Peaks A-J and 1-5 correspond to the regions and sites labeled in these letters in Table 1.

Open Regions with Statistical Significance at the RNA Level

Several interesting open regions(locally unstable secondary structure or single stranded region) seen in the RNA which were statistically more significant have also been identified. The interesting features of these open regions are that the minimal free energies of calculated optimal secondary structures of the actual, biological sequence segment are equal to zero (no stable secondary structures can be formed in these regions), however the average free energies of randomized segments are less than zero (random permutations of the biological sequence can form locally stable secondary structure). Thus, these open or single-stranded regions appear to have significantly less secondary structure than might be expected.

In order to explore the possible relationship between hypervariable sites in the AIDS virus and these significant open regions, we aligned four HIV-1 sequences from different isolates: two from different strains of HTLV-III(5,35), LAV(33,34), and ARV(32) using the NUCALN program(30). The variability was then computed as the number of different bases at a given position divided by the frequency of the most common base at the position. Fig. 7 shows that most hypervariable sites are located in the envelope

Table 1 Nucleotide Positions of Hypervariable Regions in the Sequences of HIV-I

Peaks	No. in the Maximum Alignment**	No. in HTLV-III	No. in ARV
1	201		
2	601		
A*	1626-1775	1614-1785	1624-1759
B	3450-3551	3439-3539	3388-3513
3	4901		
4	5601		
C	5751-5900	5733-5873	5712-5848
D	6176-6425	6142-6394	6117-6381
E	6676-6850	6619-6790	6606-6774
5	6876		
F	6951-7100	6890-7045	6875-7014
G	7151-7250	7082-7203	7051-7175
H	7726-7825	7651-7756	7601-7731
I	8426-8575	8351-8447	8326-8434
J	8901-9000	8814-8913	8801-8900

* Peaks A-J and 1-5 are labeled in Fig.7. ** The gaps are inserted to get maximum alignment for four isolates, LAV, ARV, HTLV-III(BH10), and HTLV III(BH9).

coding region except for a sharp peak in the gag coding region. This result agrees with previous reports showing that env contains hypervariable sites where the nucleotide sequence has diverged rapidly between HIV isolates (17,27,29,31). Each of the four major regions(17) of hypervariablity of amino acid sequences (positions 137-211(I), 316-370(II), 412-440(III), and 651-675(IV)) in the HIV env corresponds to one of the sharp peaks in the Fig.7. Among them, the hypervariable region I corresponds to the peak D, region II to the peak E, region III to the peaks F and G, and region IV to peak H. The nucleotide positions of these hypervariable regions in the sequences of HIV-I are shown in Table 1. Table 2 shows the correlation of HIV hypervariable sites with significant open regions at the RNA level. From Table 2 it can be seen that the major open regions are located in the coding region of the envelope protein of both HTLV-III and ARV. The lack of secondary structure in these regions is conserved. Each of the hypervariable peaks identified in Fig.7 lies in or overlaps the regions lacking the potential to form secondary structure(Table 2) Among them, hypervariable peaks E and H have statistically more significant open structure. Moreover, single-stranded segments at least 60-bases long occur in the regions E and H. It is worth noting that there is another significant open region in the signal peptide of the env gene of both HTLV-III (5708-5822) and ARV (5686-5813) (peak C in the Fig.7). In the single-stranded region no stable local secondary structure can be folded within about 100-base. We propose that the lack of

Table 2 Correlation of HIV Hypervariable Sites with Lack of RNA Secondary Structure

Hypervariable Sites Peak and Position*1 From Fig.7		Open Regions*2 in HTLV-III W=40	W=60	in ARV w=40	w=60**
A	1614-1785	1669-1749	1677-1749	1679-1733	-
B	3439-3539	3448-3491	-	3436-3478	-
C	5733-5873*3	5708-5827	5708-5822*5	5619-5815	5630-5813*6
D	6142-6394*4	6308-6384	6318-6383	6295-6317	-
E	6619-6790*4	6727-6792	6727-6792	6693-6795	6706-6795
F	6890-7045*4	-	-	6939-7012	6939-7012
G	7082-7203*4	7122-7171 7159-7207	-	7032-7071	-
H	7651-7756*4	7655-7730	7656-7729	7686-7748	7662-7771*6
I	8351-8447	8324-8380	-	-	-
J	8814-8913	8873-8929	-	-	-

*1 Nucleotide position numbered according to HTLV-III(BH10) sequence(31). *2 Regions listed are all those where free energy of the optimal secondary structure of the actual sequence segment(40 or 60 nucleotides) was zero, i.e. no secondary structure is predicted, but the average free energy of the randomized segments(with the same base composition and same size) was less than zero(i.e. some structures can form). ** Window size was set at 40 and 60 nucleotides. *3 The region corresponds to the env signal peptide. *4 The region corresponds to previously recognized HIV hypervariable sites (17,29). *5 the secondary structure with 100 nucleotides can't be predicted either in the open region 5708-5822 of HTLV-III. *6 the secondary structure with 100 nucleotides can't be predicted either in the open regions 5686-5813 and 7666-7771 of ARV.

secondary structure at the RNA level may contribute in some way to the hypervariability observed at these sites in HIV.

DISCUSSION

Correlation between Boundaries of Genes and Stability Map of Local Secondary Structure of HIV

In a rigorous examination of the viral RNAs of HTLV-III and ARV, the boundary of stability of local secondary structure of these RNAs is found to have a significant correlation with the boundary of their protein coding regions. The statistical analyses of the stability map have also shown that the average free energies between two adjacent protein coding regions have considerable fluctuation (see Figs.5a-5b). It is worth mentioning that there is a large stability fluctuation in the splicing site region between the intervening sequences and Exon 2 of the tat and art mRNA (see Figs. 1(a), 1(b), and 2(a)). Interestingly, a quite large stability fluctuation occurs in the boundary between EGP/OMP and TMP of env gene. This finding is surprising since no RNA signal or RNA processing are expected to occur at this site. In addition, the statistical feature of the stability distribution of local secondary structure is independent of

the overlapping segment size in the simulation calculation. In the case of viral DNAs (ϕX174, G4, fd, SV40 and BKV) the stability distribution pattern also show a statistically significant correlation with the boundary of their protein coding regions(18-20). Some papers have also suggested that the distribution of stable features in DNA may be correlated with the origin of replication and initiation and termination sites of transcription(21-23). Thus, the profile of the stability distribution of the local secondary structure of single-stranded retroviral RNA may be suggested to play a role in the processing of mRNA precursors to determine site-specificity and efficiency.

Secondary Structure of the HIV 5' mRNA Leader

It was shown that the translational efficiency and steady-state level of HIV LTR-directed mRNA are both significantly enhanced in COS cells by the HIV tat gene(13). Cullen(13) has presented evidence suggesting that the sequences encoded within the HIV R region, specifically within the first 37 nucleotides of the R region, inhibit mRNA translation when located in cis at the 5' terminus of the mRNA. Moreover, deletion analysis of the LTR indicates that a large inverted repeat containing a smaller direct repeat may be an important feature of the trans-activation site(14). In our computer simulations, these sequences efficiently fold into a stable stem-loop in the corresponding mRNA. A possible role for the RNA secondary structure in trans-activation has also been suggested(13,14,24). RNA secondary structure was analyzed using ribonuclease and cobra venom nuclease, and a structure consisting of two adjacent stem-loops has been proposed(14). Our results clearly indicate that a statistically significant stem-loop structure can be predicted in the region 5-104 of 5' terminus of the HIV mRNA(Fig.6), and completely support Muesing's experimental results(14). A two-fold symmetry of seven base-pairs exist within the helical stem of the stem-loop I(UCUGGUU and UCUGGCU). It is reasonable to suppose that this dyad element provides the possible recognition site for a dimeric RNA binding protein. Further studies of the intermolecular interaction between the trans-activator site and tat gene product will be necessary to confirm this hypothesis.

Sequence Divergence and Open Region of RNA Secondary Structure

Genetic variation during the course of infection of an individual is a remarkable feature of the acquired immune deficiency syndrome (AIDS) disease. Such variation has been observed for ARV, LAV, African (Z3) and North American (NY5) AIDS retroviruses(17). It has been observed that the

preponderance of mutations among different HIV isolates occur in the env gene(17,29,31). The correlation between hypervariable sites and regions with little intramolecular base pairing in both the visna virus genome, a pathogenic retrovirus related to HIV(36-38) and HIV genomes has been noted(25). In this paper, we show that there are several evident hypervariable regions when the HTLV-III, ARV, and LAV isolates are compared(see Fig.7). Ten hypervariable clusters (A-J) and five sites(1-5) with a variability of at least 1.2 are shown in Fig.7. The frequency of their occurring in the total genome is 0.163. Furthermore, the frequencies of significant open regions occurring in the HTLV-III are 0.124 for window size 40, 0.040 for 60 and 0.005 for 100, respectively. The probabilities of coincidence or overlap between the hypervariable regions and the significant open regions in HTLV-III are 0.02 for window size 40, 0.007 for 60 and 0.0008 for 100 according to random distribution. Similarly, the probabilities for ARV are 0.016 for 40, 0.0065 for 60 and 0.0016 for 100. However, our results show that the hypervariable regions appear to coincide quite well with those open and loose regions predicted to have significantly less secondary structure than might have been expected at random. Obviously, the coincidence is not a random event. The correlation between hypervariable sites and single-stranded regions at the RNA level may provide a clue to mechanisms inducing mutability. It may imply that in these regions there are no base-pairing constraints which limit the rate of sequence divergence. Alternatively, the nucleotides of these regions may be more exposed to chemical or physical mutagens than those with stable secondary structure. The fact that most hypervariable sites observed among different AIDS retrovirus isolates occur in the env gene, which codes for the major viral surface proteins, suggests the involvement of the host immune system in selecting for env mutants which can escape immune surveillance. However, the close correspondence between hypervariable sites and regions lacking secondary structure and the fact that most of these open regions occur in the env gene, together suggest that the virus may be preadapted in such a way that the env gene is intrinsically hypermutable. It is known that some of the mutations found at the RNA level are not silent substitutes. The majority of the predicted antigenic determinants in the EGP were found in regions of high sequence variability in the viral isolates(27). This would be advantageous to the virus, because it would concentrate mutations where they would do the most good in allowing the virus to avoid elimination by the host.

Acknowledgements

Research sponsored, at least in part, by the National Cancer Institute, DHHS, under contract NO1-CO-23910 with Program Resources, Incorporated.

REFERENCES

1. Barre-Sinoussi, F. et al. (1983) Science 220, 868-870.
2. Popovic, M., Sarngadharn, M.G., Read, E. and Gallo, R.C. (1984) Science 224, 497-500.
3. Ratner, L. et al. (1985) Nature 313, 636-637.
4. Sarngadharan M. G. Popovic, M., Bruch, L., Shupbach, J., and Gallo, R. C. (1984) Science 224, 506-508.
5. Ratner, L. et al. (1985) Nature 313, 277-283.
6. Sanchez-Pescador, R. et al. (1985) Science 227, 484-492.
7. Muesing, M.A. et al. (1985) Nature 313, 450-458.
8. Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S. and Alizon, M. (1985) Cell 40, 9-17.
9. Sodroski, J. et al. (1985) Science 227, 171-173.
10. Sodroski, J., Patarca, R., Rosen, C., Wong-Staal, F., and Haseltine, W. (1985) Science 229, 74-77.
11. Rosen, C., Sodroski, J., and Haseltine, W. (1985) Cell 41 , 813-823.
12. Rosen, C., Sodroski, J., Goh, W., Dayton, A., Lippe, J. and Haseltine, W. (1986) Nature 319, 555-559.
13. Cullen, B. (1986) Cell 46, 973-982.
14. Muesing, M., Smith, D., and Capon, D. (1987) Cell 48 , 691-701.
15. Freier, S. et al. (1986) Proc. Natl. Acad. Sci. USA 83 , 9373-9377.
16. SAS Institute, Inc. (1985) SAS User's Guide: Basics. Version 5 ed. Cary, N. C.: SAS Institute Inc.
17. Willey, R. et al. (1986) Proc. Natl. Acad. Sci. USA. 83 , 5038-5042.
18. Wada, A. and Suyama. A. (1983) J. Phys. Soc. Jpn. 52 , 4417-4422.
19. Wada, A. and Suyama. A. (1984) J. Biomolec. Struct. Dyn. 2 , 573-591.
20. Wada, A., Ueno, S., Tachibana, H. and Husimi, Y. (1979) J. Biochem. 85, 827-832.
21. Wada, A. and Suyama. A. (1986) Prog. Biophys. Molec. Bio. 47,113-157
22. Wada, A. and Suyama. A. (1984) Molecular Basis of Cancer (ed. R. Rein) pp. 37-46, Alan R. Liss Inc. New York.
23. Wells, R. et al. (1980) Prog. Nucl. Acid. Res. Molec. Biol. 24, 167-267.
24. Okamoto, T. and Wong-Staal, F. (1986) Cell 47, 29-35.
25. Braun, M.J., Clements, J.E. and Gonda, M.A. (1987) J. Virol., in press.
26. Guyader, M., Emerman, M., Sonigo, P., Clavel, F., Montagnier, L., and Alizon, M. (1987) Nature 326, 662-669.
27. Modrow, S., Hahn, B.H., Shaw, G.M., Gallo, R.C., Wong-staal F. and Wolf, H. (1987) J. Virol. 61, 570-578.
28. Zuker, M. and Stiegler, P. (1981) Nucl. Acids Res. 9, 133-148.
29. Starcich, B.R., Hahn, B.H., Shaw, G.M., McNeely, P.D., Modrow, S., Wolf, H., Parks, E.S., and Parks, W.P. (1986) Cell 45, 637-648.
30. Wilbur, W. J. and Lipman, D.J. (1983) Proc. Natl. Acad. Sci. USA 80, 726-730.
31. Myers G., Rabson, A.B., Josephs, S.F., Smith, T.F., and Wong-staal, F. (1987) A compilation and analysis of nucleic acid and amino acid sequences. In Human Retroviruses and AIDS. pp. II-33. Los Alamos, New Mexico, USA.
32. Sanchez-Pescador, R. et al. (1985) Science 227, 484-492.

33. Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S. and Alizon, M. (1985), Cell 40, 9-17.
34. Alizon, M., Wain-Hobson, S., Montagnier, L. and Sonigo, P. (1986) Cell 46, 63-74.
35 Muesing, M.A., Smith, D.H., Cabradilla, C.D., Benton, C.V., Lasky, L.A. and Capon, D.J. (1985) Nature 313, 450-458.
36 Gonda, M.A., Braun, M.J., Clements, J.E., Pyper, J.M., Gallo, R.C. Wong-Staal, F. and Gilden, R.V. (1986) Proc. Natl. Acad. Sci., U.S.A. 83, 4007-4011.
37. Gonda, M.A., Wong-Staal, F., Gallo, R.C., Clements, J.E., Narayan, O. and Gilden, R.V. (1985) Science 227, 173-177.
38. Sonigo, P., Alizon, M., Staskus, K., Klatzmann, D., Cole, S., Danos, O., Retzel, E. and Wain-Hobson S. (1985) Cell 42, 369-382.