Thermodynamic stability and statistical significance of potential stem–loop structures situated at the frameshift sites of retroviruses

Shu-Yun Le[1,3], Jih-H.Chen[2] and Jacob V.Maizel[1]

[1]Laboratory of Mathematical Biology, Division of Cancer Biology and Diagnosis, National Cancer Institute, National Institutes of Health, Building 469, Room 151, Frederick, MD 21701, [2]Advanced Scientific Computing Laboratory, Program Resources, Inc., NCI/Frederick Cancer Research Facility, Frederick, MD 21701, USA and [3]Shanghai Institute of Biochemistry, Chinese Academy of Sciences, Shanghai 200031, China

Received March 31, 1989; Revised and Accepted July 3, 1989

ABSTRACT

RNA stem-loop structures situated just 3′ to the frameshift sites of the retroviral *gag-pol* or *gag-pro* and *pro-pol* regions may make important contributions to frame-shifting in retroviruses. In this study, the thermodynamic stability and statistical significance of such secondary structural features relative to others in the sequence have been assessed using a newly developed method that combines calculations of the lowest free energy of formation of RNA secondary structures and the Monte Carlo simulations. Our results show that stem-loop structures situated just 3′ to the frameshift sites are both highly stable and statistically significant relative to others in the *gag-pol* or *gag-pro* and *pro-pol* junction domains (both 300 nucleotides upstream and downstream from the possible frameshift sites are included) of Rous sarcoma virus (RSV), human immunodeficiency virus (HIV-1), bovine leukemia virus (BLV), human T-cell leukemia virus type II (HTLV-II), and mouse mammary tumor virus (MMTV). No other more stable, or significant folding regions are predicted in these domains.

INTRODUCTION

Some retroviruses express their *pol* genes as *gag-pol* fusions which are later cleaved by a virus-encoded protease to yield the mature pol proteins. In the fusion protein, *gag* and *pol* are joined by overlapping different reading frames in RSV and HIV-1, and are interrupted by a third gene (encoding *pro*, the viral protease) which overlaps them both in MMTV, BLV and HTLV-II. This translation mechanism clearly benefits retroviruses in that one kind of mRNA molecule can direct large amounts of structural (*gag*) protein synthesis, relatively small amounts of catalytic (*pro* and *pol*) protein synthesis, while attached *gag* components can direct incorporation and *pol* products into viral cores.

The basis of such translation control has also been proposed (1–5) as moving the ribosome in a −1 reading frame in response either to localized recognition of a primary structure or to some secondary, or higher order, structure of the RNA template. Previous and current studies have revealed that the conserved sequences AAAAAAC and UUUA are the probable frame-shifting sites. The former sequence is present at the *gag-pro* junction, and the UUUA sequence is present at the *pro-pol* or *gag-pol* junction domains. Moreover, Jacks and Varmus(1,4,5), Moore et al.(2) and Rice et al.(3) have also proposed that stem-loop structures situated just 3′ to the frameshift sites may make important contributions to frame-shifting in retroviruses. The thermodynamic stability and statistical significance of such secondary structural features relative to others in the sequence were not assessed, and details about the stability of these secondary structural features were not fully known.

In this study we have analyzed potential RNA secondary structures in the *gag-pol* or *gag-pro* and *pro-pol* junction regions using a recently developed Monte Carlo simulation

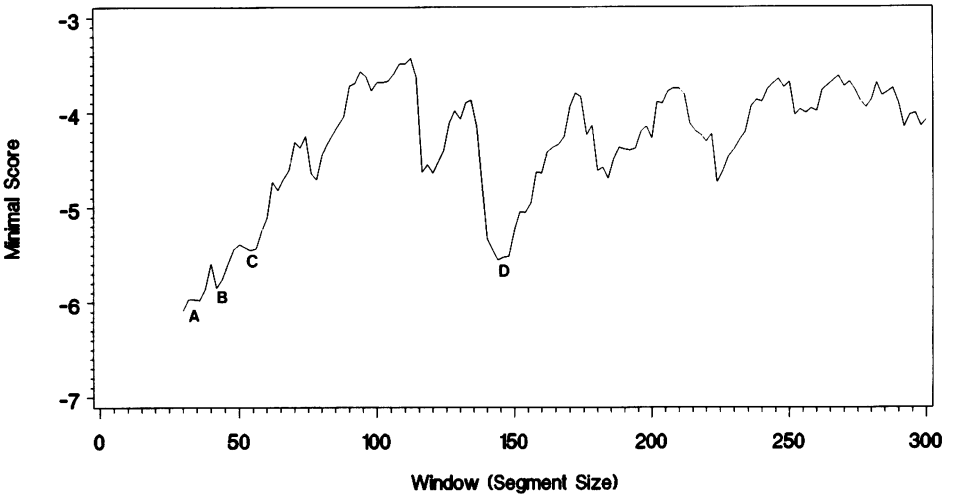


Figure 1. Distribution of significant folding region sizes in the *gag-pol* junction region of HIV-1. The segment scores of the secondary structure with the lowest free energy were computed(7) using Tinoco energy rules (18, 19). For each specific segment (window) size, the simulation was carried out by sliding one base along the RNA sequence. The global minimum score was picked up for each simulation and plotted against the window size. The exhaustive simulation was completed for window sizes ranging from 30 to 300 bases by increasing two bases to the window. The suitable size of significant regions for RNA folding were identified (30, 42, 56, and 144) and labelled by letters A–D. The corresponding folding regions are A(1639–1668), B(1638–1679), C(1625–1680), and D(1561–1704).

method (6,7) which incorporates a dynamic programming algorithm (9,10). In this approach, the free energy corresponding to each segment which measures the thermodynamic stability, and the segment score which measures the statistical significance of the optimal secondary structure folded within the window, were computed for segments at successive positions along the sequence. The optimal sizes of the segments containing maximal statistically significant secondary structures in the sequence are predicted by an exhaustive Monte Carlo simulation, in which the Monte Carlo simulation is carried out repeatedly as the window size changes. The extensive simulation is effective for the detection of the predicted stable folding regions and RNA secondary structures with statistically high significance in both the putative RNA target sequence for trans-activation by the viral *tat* gene product(6) and the *cis*-acting *rev* response element(8) of HIV-1.

The distributions of free energies and segment scores for potential secondary structures in the junction regions of *gag-pol* in RSV(11), HIV-1(12), as well as *gag-pro* and *pro-pol* in BLV(13), HTLV-II(14) and MMTV(2) are presented. The minima of these scores and free energies in their distributions are located at the neighborhood of the frameshift sites of *gag-pol* or *gag-pro* and *pro-pol* of these viruses. Our analyses show that stem-loop structures situated just 3′ to the frameshift sites are both extremely stable and highly significant relative to others in all these sequences.

METHODS

The program RANFOLD is designed for assessing segment scores which measure the statistical significance of the optimal secondary structures (6,7). In the segment score of

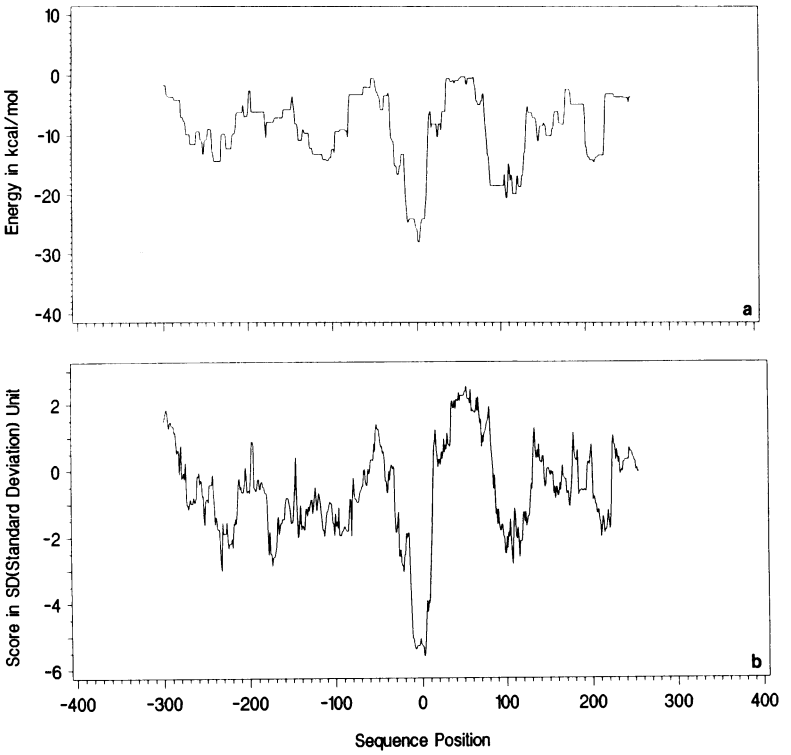


Figure 2. (a) Thermodynamic stability distribution of stem-loop structures in the *gag-pol* domain of HIV-1. The unit of free energy is kcal/mol. (b) Statistical significance distribution of stem-loop structures in *gag-pol* domain of HIV-1. The segment scores are represented in standard deviation(SD). The free energies and segment scores of the stem-loop structures of 46-base long overlapping segments (window) are plotted against the position of the first base of each segment along the *gag-pol* domain of HIV-1. In the maps the frameshift site is taken as position zero.

the RNA secondary structure, S is defined as the difference between the lowest free energy (E) of the real biological sequence and the mean (E_r) of the lowest free energies from a large number of random permutations of the real sequence, divided by the standard deviation (SD) of the random sample:

$$S = (E - E_r) / SD$$

In the equation, E_r and SD can be computed using empirical formulas(15, 24) based on the length and base composition of the segment. The distribution of the lowest free energies generated from these randomized sequences of the same nucleotide compositions as the biological segment has been revealed as an approximately normal distribution(16). Thus, the probability that a score of the real biological segment would be obtained in a comparison of randomized sequences can be approximately determined using the table of the standardized normal distributions. The lower the probability of occurrence of a particular segment score by chance, the more significant the secondary structure of the real biological segment. In practice, the statistical significance of a RNA secondary structure relative to

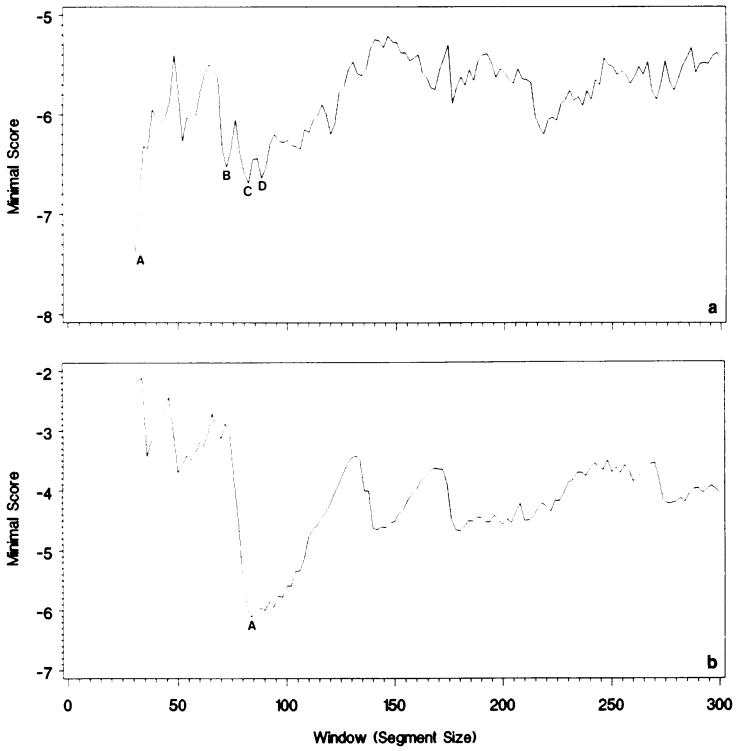


Figure 3. Distribution of significant folding region sizes in the (a) *gag-pro* and (b) *pro-pol* junction regions of MMTV. The suitable size of significant regions for RNA folding were identified (30, 72, 82, and 88) and labelled by letter A–D in (a), as well as by letter A (window size: 84) in (b). For further details see the caption to Figure 1.

others in the sequence are estimated by an exhaustive Monte Carlo simulation. In this study, the exhaustive simulation of each RNA sequence of retroviruses was performed for window sizes ranging from 30 to 300 bases by adding two bases to the window size. For each window size, the calculation of segment score was carried out by sliding one base along the RNA sequence in the *gag-pol* or *gag-pro* or *pro-pol* domains. In the calculation the global minimum segment score and its corresponding segment starting position in the sequence were selected for each window. As a result, regions of significant predicted structures relative to others in the sequence can be detected in plots of these minimum scores versus window sizes and versus sequence positions.

RESULTS AND DISCUSSIONS

The possible frameshift sites of *gag-pol* or *gag-pro* and *pro-pol* of retroviruses have been proposed by Jacks and Varmus(1,4,5). In our study, the *gag-pol* domains of RSV and HIV-1 (BH10 isolate) and *gag-pro* and *pro-pol* domains of BLV, HTLV-II and MMTV, are decided according to their open reading frames in the genomes of these retroviruses. These junction regions consist of about 600 nucleotides (i.e., they include both 300 bases upstream and downstream from the possible frameshift sites of RSV(17), HIV-1(4), BLV(3,

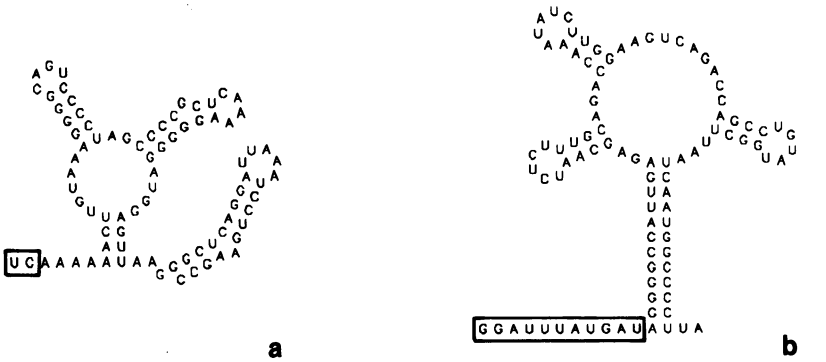


Figure 4. Predicted secondary structures followed close the frameshift sites of MMTV. (a) in *gag-pro* and (b) in *pro-pol* junction regions. The RNA stem-loop structures presented here are both extremely stable ($\Delta G = -49.7$ kcal/mol in (a) and -45.2 kcal/mol in (b) are computed using Tinoco energy rules, and -26.5 and -27.2 kcal/mol using Turner energy rules) and highly significant (segment score in standard deviation Units = -6.64 in (a) and -6.11 in (b)). The secondary structure models derived from Tinoco and Turner energy rules are very similar even though their free energies are quite different. The nucleotides in the box are added for computing the distributions of thermodynamic stability and statistical significance of potential stem-loop structures in these two domains (see text). The graphs shown were generated by using DRAW program(22). For further details see the caption to Figure 1.

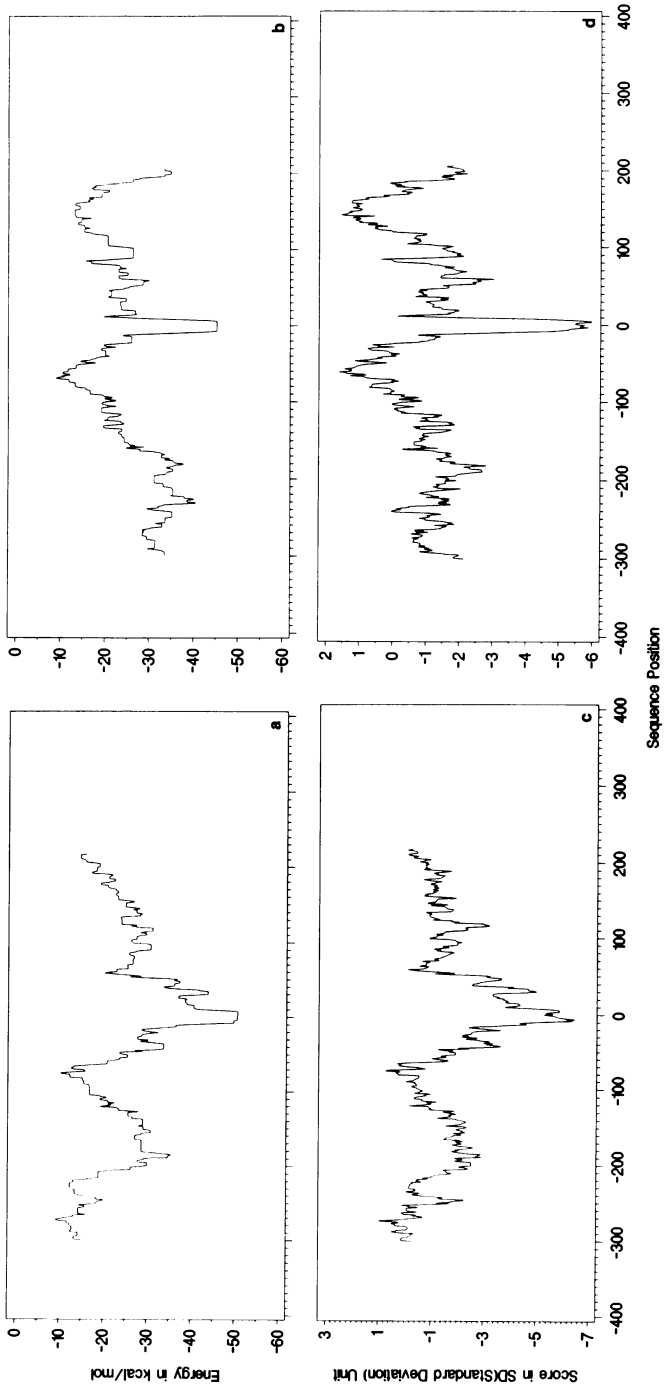
5), HTLV-II(5, 14), and MMTV(1, 2)). The exhaustive Monte Carlo simulations for all these viruses were carried out in these junction domains using the Cray Operating System of a CRAY X-MP/24 supercomputer. The program was vectorized and optimized (7).

Predicting Highly Significant Folding Regions in the gag-pol Junction Domain of HIV-1

Suitable folding regions for the potential highly significant secondary structures in *gag-pol* junction domain of HIV-1 were assessed using the exhaustive Monte Carlo simulation. The relation between the global minimum score and window size abstracted from each simulation of HIV-1 *gag-pol* junction domain is plotted in Fig.1. The deep troughs where the scores are less than -5.0(SD) are labelled A$-$D. The corresponding highly statistically significant folding regions of Valleys A$-$D in Fig. 1 are 1639$-$1668 (window size: 30), 1638$-$1679 (window size: 42), 1625$-$1680 (window size: 56) and 1561$-$1704 (window size: 144) respectively, where the sequence position of frameshift site in HIV-1 *gag-pol* expression is position 1634 (i.e., the sequence position of base A in the frameshift signal UUUA, the starting position of HIV-1 mRNA is numbered 1). A potential stem-loop structure situated just 3′ to the frameshift site in HIV-1 *gag-pol* expression proposed in (4) can be predicted by folding any of regions A$-$C. The stem-loop structure is highly statistically significant (the segment score is at least less than -5.0 SD units from the mean).

Thermodynamic Stability and Statistical Significance of the Stem-loop Structures in the gag-pol Domain of HIV-1

The thermodynamic stability and statistical significance of the distinct stem-loop structure proposed in (4) relative to others in the *gag-pol* region of HIV-1 were also computed by the Monte Carlo simulation. In the simulation, the window size was taken 46 bases according to the size of the distinct stem-loop structure presented in (4) (the segment sequence encompasses 3 nucleotides upstream from the frameshift signal). The proposed structure in (4) also displays a lowest level of segment score (-5.59 SD unit from the mean) and


a
b
c
d
Energy in kcal/mol
Score in SDXStandard Deviation) Unit
Sequence Position

free energy ($\Delta G = -27.9$ kcal/mol) in the junction domains (see Fig.2). The data presented in Fig. 2 indicate that the stem-loop structure situated just 3′ to the frameshift site of HIV-1 *gag-pol* expression is the most stable and statistically significant in relation to others in the *gag-pol* junction domain.

Predicting Highly Significant Folding Regions in the gag-pol and pro-pol Junction Domains of MMTV

According to MMTV sequence data(2), the positions of the translational frameshift sites of *gag-pro* and *pro-pol* in MMTV are position 3402 (the sequence position of base C in the frameshift signal AAAAAAC, the starting position of mRNA is numbered 1), and 4210 (the position of base A in UUUA). The potentially high statistically significant stem-loop structures that follow close to the frameshift site in *gag-pro* and *pro-pol* translation of MMTV(2) were predicted using the same simulation (see above). The data presented in Fig. 3a revealed that the window sizes of four deep troughs (A−D) are 30, 72, 82 and 88 respectively. The significant folding regions represented by these four valleys are 3450−3479(A) 3406−3477(B), 3396−3477(C), and 3391−3478(D) in *gag-pro* domain. The significant folding region D fully encompasses the other three regions. The stem-loop structure with the lowest free energy predicted in the regions D is identical with that predicted in region C (Fig.4a). The stem-loop structure presented in Fig.4a also fully contains the other two structures predicted in region A and B. Similarly, the significant folding region of trough A in Fig.3b is 4216−4299 (window size: 84) in *pro-pol* region and the stem-loop structure with the lowest free energy is presented in Fig.4b. The two secondary structures in Fig.4 are both extremely significant (their segment scores are at least less than −5.0 SD units).

Thermodynamic Stability and Statistical Significance of Stem-loop Structures in the gag-pol and pro-pol Domains of MMTV

In order to assess the thermodynamic stability and statistical significance of these distinct secondary structures located at the frameshift sites of MMTV *gag-pro*, a window size of 84 was chosen (i.e., two bases upstream from the proposed frameshift signal were added to the 5′ of the segment sequence plotted in Fig.4a). Similarly, in the simulation of *pro-pol* domain, a window size of 94 was chosen for which 11 nucleotides (fully encompass proposed frameshift signal of *pro-pol* upstream from the 5′ of the segment in Fig.4b were enclosed and one nucleotide in the 3′ of the segment was neglected in the calculation. The distributions of free energies and segment scores of the secondary structures with the lowest free energy predicted in *gag-pro* and *pro-pol* domains are displayed in Fig.5. These plots explicitly indicate that the remarkable minima of both segment scores and free energies in these two domains are located at their frameshift signals or close to 3′ of their frameshift sites. Clearly, these two predicted secondary structures situated just to the frameshift sites of *gag-pro* and *pro-pol* of MMTV are most stable and statistically significant in relation to others in these two domains.

Figure 5. The distributions of thermodynamic stability and statistical significance of stem-loop structures in MMTV. (a) The stability distribution of stem-loop structures in the *gag-pro* domain and b) in the *pro-pol* domain of MMTV. (c) The segment score distribution of stem-loop structures in the *gag-pro* region and (d) in the *pro-pol* region. The window sizes were taken as 84-base in (a) and (c), as well as 94-base in (b) and (d). For further details see the caption to Figure 2.

Thermodynamic Stability and Statistical Significance of Stem-loop Structures Predicted in the gag-pol or gag-pro and pro-pol Junction Domains of RSV, BLV, and HTLV-II
Following the same procedure the potentially highly significant folding regions and their possible secondary structures in *gag-pol* of RSV, as well as in *gag-pro* and *pro-pol* of BLV and HTLV-II, were predicted (the data are not present here). The distributions of free energies and segment scores of the secondary structures with the lowest free energy predicted in *gag-pol* of RSV, and *gag-pro* and *pro-pol* domains of HTLV-II and BLV, are calculated. Similar to the features presented in Figs. 2 and 5, These plots (not shown here) explicitly indicate that these secondary structures that follow close to the frameshift sites of RSV *gag-pro*, and HTLV-II *gag-pro* and *pro-pol* are both the most stable and statistically significant in relation to others in these domains. These structures that downstream from possible frameshift sites of RSV and HTLV-II all display a level of remarkable stability which is highly statistically significant (score less than −5.0 SD units from the mean). The global minimum segment score of the distinct stem-loop structure located at the 3′ of frameshift site of BLV *gag-pro* is −5.13 SD and the other of *pro-pol* is −3.52 SD. Although the global minimum of free energies in BLV *gag-pro* isn't at the neighborhood of the frameshift site, there is a local minimum in the distribution located at the frameshift site for the region. One can notice that those more stable folding regions are far from the frameshift site and are not overlapping with the significant folding region. Thus, the stem-loop structure predicted at downstream from 3′ of frameshift site of BLV *gag-pro* is also both highly stable and statistically significant in relation to the others in the neighborhood of the frameshift site. This is similar to the results in the BLV *pro-pol* and other frameshift regions of HIV-1, RSV, MMTV, and HTLV-II mentioned above.

The *pol* gene is generally the best conserved in the retroviral genes (23). HTLV has more sequence similarity with MMTV and RSV than BLV does (13). The detailed comparisons of these *gag-pol* junction domains of RSV and HIV-1, or *gag-pro* and *pro-pol* junction domains of BLV, HTLV-II and MMTV did not reveal greater sequence similarity than that of other parts in these viruses. Although there is sequence similarity in the neighborhood of the frameshift sites of *gag-pro* for HTLV-II and BLV, as well as, in the sites of *pro-pol* for MMTV and BLV, no other distinct conserved sequence in these three retroviruses were detected except for frameshift signals (the data are not presented here). When the nucleotide patterns of the potential signals, AAAAAAC and UUUA were searched in the *gag-pol* or *gag-pro* and *pro-pol* junction domains of these retroviruses of HIV-1, RSV, BLV, HTLV-2 and MMTV, only one site of the oligonucleotide pattern AAAAAAC was detected in the junction domain for each of these retroviruses. However, the pattern UUUA occurs at several sites in these *gag-pol* or *pro-pol* junction domains of HIV-1, BLV, and MMTV. The sites of the nucleotide A in the pattern UUUA found in HIV-1 locate at the positions of 1334, 1482, 1634 and 1826. Among them, the position 1634 is a real frameshift site of *gag-pol* of HIV-1 (4). We have indicated that the stem-loop structure located just 3′ of UUUA (position 1634) is the most stable and statistically significant in the *gag-pol* junction domain of HIV-1 in Fig.2. The secondary structures situated just 3′ of other patterns of UUUA (positions 1334, 1482, and 1826) are less stable and significant relative to others in the domain (the position 1634 was transformed to the position zero in Fig.2, thus, the position 1334 was transformed to position −300, 1482 to −152, and 1826 to +192). Similarly, the sites of the nucleotide A in the pattern UUUA occur at the positions of 1928 (transformed to −196), 1948(−176), 1978(−146) and

2124(0) for BLV and the positions of 3909(−301), 4030(−180), 4210(0) and 4320(+110) for MMTV. Among them, except of the site 2124(0) of BLV and the site 4210(0) of MMTV, no other sites of these UUUA patterns are followed by more stable and statistically significant secondary structures.

An interesting question raised by the study is why the frameshift sites are closely followed by highly stable and statistically significant stem-loop structures in these virion RNAs. The occurrence of these secondary structures is not random. Except for the stem-loop structure downstream of BLV *pro-pol* junction region (the segment score is −3.52 SD units from the mean), the segment scores of these distinct stem-loop structures by chance are all less than −5.0 SD units. These structures, being both the most stable and significant in the *gag-pol* or *gag-pro* and *pro-pol* domains of these retroviruses, are apparently relevant to some biologic function of these viruses. One possibility is that such extremely stable stem-loop structures may act by stalling translating ribosomes, where they promote the RNA to slip back one nucleotide and pair with the codon in the −1 frame(1). Our results strongly support the possible control mechanism for frameshifting proposed by Jack and Varmus et al(1,4,5) and Moore et al.(2).

In the absence of confirming experimental data, the prediction of RNA secondary structure has some limitation. The secondary structure with the lowest free energy obtained using current dynamic programming algorithm and energy rules (Tinoco energy rules (18,19) and Turner energy rules (20,21)) often differ from those observed in solution or in crystal form. In this study, one may notice that we have paid much attention to detecting the statistically non-random and thermodynamically stable folding regions instead of to specific predicted structure. We previously reported that a different set of energy rules does not change the prediction of the segment score level for folding regions (16), even though detailed predicted structures may vary. Although these predicted secondary structures may have some defects, the highly stable and significant folding regions detected in these junction domains are nevertheless noteworthy. This represents yet another case where the domain of a biologically interesting phenomenon correlates with the location of non-random structure, as was the case for *rev* and *tat* of HIV-1 (6, 8). The computer simulation presented here is a powerful tool for detecting statistically significant folding regions in RNAs.

ACKNOWLEDGMENTS

Research sponsored, at least in part, by the National Cancer Institute, DHHS, under contract N01-C0-74102 with Program Resources, Incorporated.

REFERENCES

1. Jacks, T., Townsley, K., Varmus, H. E. and Majors, J. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 4298−4302.
2. Moore, R., Dixon, M., Smith, R., Peters, G. and Dickson, C. (1987) J. Virol. **61**, 480−490.
3. Rice, N. R., Stephens, R. M., Burny, A. and Gilden, R. V. (1985) Virology **142**, 357−377.
4. Jacks, T., Power, M. D., Masiarz, F. R., Luciw, P. A., Barr, P. J. and Varmus, H. E. (1988) Nature **331**, 280−283; Jack, T., Madhani, H. D., Masiarz, F. R. and Varmus, H. E. (1988) Cell **55**, 447−458.
5. Varmus, H. E. (1988) Science **240**, 1427−1435.
6. Le, S.-Y., Chen, Jih-H., Braun, M.J., Gonda, M.A. and Maizel, J.V.Jr. (1988) Nucl. Acids Res. **16**, 5153−5168.
7. Le, S.-Y., Chen, J.-H., Currey, K.M., and Maizel, J.V.Jr. (1988) Computer Applications in the Biosciences **4**, 153−159.
8. Malim, M.H., Hauber, J., Le, S.-Y., Maizel, J.V. and Cullen, B.R. (1989) Nature in press.
9. Nussinov, R. and Jacobson, A.B. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 6309−6313.

10. Zuker, M. and Stiegler, P.(1981) Nucl. Acids Res. **9**,133–148.
11. Schwartz, D.E., Tizard, R. and Gilbert, W. (1983) Cell, **32**, 853–869.
12. Ratner, L. et al. (1985) Nature **313**, 277–283.
13. Sagata, M., Yasunaga, T., Tsuzuku-Kawamura, J., Ohishi, K., Ogawa, Y. & Ikawa, Y. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 677–681.
14. Shimotohno, K., Takahashi, Y., Shimizu, N., Gojobori, T., Golde, D.W., Chen, I.S.Y., Miwa, Masanao and Sugimura, T. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 3101–3105.
15. Chen, Jih-H., Le, S.Y., Shapiro, B., Currey, K.M. and Maizel, J.V., Jr. (1988) submitted.
16. Le, S.-Y. and Maizel, J.V. (1989) J. of Theor. Biol. in press.
17. Jacks, T. and Varmus, H. E. (1985) Science **230**, 1237–1242.
18. Cech, T.R., Tanner, N.K., Tinoco, I.Jr., Weir, B.R., Zuker, M. & Perlman, P.S. (1983) Proc. Natl. Acad. Sci. USA **80**, 3903–3907.
19. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. **42**, 985–1002.
20. Turner, D.H., Sugimoto, N. and Freier, S.M. (1988) Ann. Rev. Biophys. Chem. **17**, 167–192.
21. Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T. and Turner, D.H. (1986) Proc. Natl. Acad. Sci. USA **83**, 9373–9377.
22. Shapiro, B.A., Maizel, J.V., Lipkin, L.E., Currey, K. and Whitney, C. (1984) Nucleic Acids Res. **12**, 75–88.
23. Chiu, I.-M., Callahan, R., Tronick, S. R., Schlom, J. & Aaronson, S. A. (1984) Science **233**, 364–370.
24. Le, S.-Y, Chen, Jih-H., Chatterjee, D. and Maizel, J.V. Jr. (1989) Nucl. Acids Res. **17**, 3275–3288.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.