УДК 577.21

DIVERSITY OF DIS, SD AND Ψ HAIRPINS IN HIV-1 ISOLATES OF GROUP M: *IN SILICO* STUDY

M. I. ZARUDNAYA, A. L. POTYAHAYLO, I. N. KOLOMIETS, D. M. HOVORUN

Institute of Molecular Biology and Genetics, National Academy of Sciences of Ukraine, Kyiv;
e-mail: m.i.zarudna@imbg.org.ua

The primary sequence and secondary structure of the region encompassing DIS, SD and Ψ hairpins in HIV-1 genomic RNAs have been analyzed for 731 group M isolates from NCBI database. The secondary structures have been predicted by the mfold program (M. Zuker). Though the primary sequence of the region studied was found to be highly heterogeneous, this region is folded into DIS, SD and Ψ hairpins (DIS-, SD- and Ψ-like hairpins) in 96% of the isolates studied. The phylogenetic analysis showed that the most frequent variants of DIS hairpin (DIS_{Lai}, DIS_{Mal} and DIS_{C}) tolerate certain base changes. Particularly, base changes at stem position 23 occur 5 and 33 times more frequently in DIS_{Lai} than in DIS_{Mal} and DIS_{C}, respectively, while A insertion at the 5'end of apical loop is tolerated in DIS_{Mal} and DIS_{C} but not in DIS_{Lai}. We have revealed that the bottom base pair substitution G-C→A-U in SD hairpin is highly specific for subtype D isolates. All variants of DIS, SD and Ψ hairpins found in our database are discussed, systematized and presented in schemes of hypothetical transitions between variants via a single base change. Most variants of DIS and Ψ hairpins were found to adopt several conformations.

Key words: HIV-1, phylogenetic analysis, RNA folding, DIS hairpin, SD hairpin, Ψ hairpin, genome dimerization, HIV-1 packaging.

The 5′-untranslated region of HIV-1 genomic RNA includes several hairpins that play an important role in the viral replication cycle [1, 2]. Among these are sequentially located DIS hairpin, essential for initiation dimerization of HIV-1 genomic RNAs, SD hairpin containing the major splice donor site and Ψ hairpin, the major packaging signal. The specific structural features of these hairpins are essential for their functioning but the secondary structures reported in literature are rather different [3–6].

To find out frequently occurring base changes in DIS, SD, and Ψ hairpins and how they affect the secondary structure of these hairpins, we have earlier surveyed 350 HIV-1 genomic sequences containing the region of approximately 100 nt preceding the GAG start codon [7]. These sequences were submitted to NCBI database by the beginning of 2004. It was shown that the secondary structure of the region under study is highly specific for all HIV-1 groups (M, O and N) and for subtype C isolates of group M. Besides we have found that the base changes in HIV-1 genome resulting in destruction of DIS, SD or Ψ hairpins are not tolerated.

To periodically review the data on the secondary structures of DIS, SD and Ψ hairpins as newly HIV-1 sequences become available from NCBI database or other sources, we developed the database of the secondary structures of the control elements in HIV-1 genomic RNAs (the database CESSHIV-1) which will be published elsewhere. The database CESSHIV-1 currently includes 757 HIV-1 sequences (731 isolates of group M, 21 isolates of group O and 5 isolates of group N).

The present paper is aimed to find out and analyse all variants of DIS, SD and ψ hairpins for HIV-1 isolates of group M from our database CESSHIV-1 and discuss the variants which occur above the level expected from sequencing and database errors (0.5% as reported in [8]). The phylogenetic analysis allowed us to demonstrate a tolerance to certain base changes in DIS_{Lai}, DIS_{Mal} and DIS_{C}, suggest a specific role of $G^{23}\rightarrow A^{23}$ base change in DIS_{Lai} hairpin in intersubtype recombination, find out country specific base covariations in SD hairpin and propose schemes of hypothetical transitions between different DIS, SD and ψ hairpin variants via a single base change. Based on mfold predictions with parameter *window=0*, we also revealed that both the stem and apical loop of most DIS variants and the stem of most ψ variants adopted several forms, i.e. they were structurally unstable.

The phylogenetic and structural data on DIS, SD and ψ hairpins (and other control elements) in HIV-1 genomic RNA are believed to contribute to clarification of the mechanism of HIV-1 replication, choice of successful primers and probes for HIV-1 diagnostics, tracking the course of the global spread of the virus, etc.

Analysis Procedure

The primary sequence and secondary structure of the region encompassing DIS, SD and Ψ hairpins have been analyzed for 731 HIV-1 isolates of M group from NCBI database for which this region was submitted. Also, the defectless hairpins and linkers have been studied for 43 HIV-1 isolates from this database which lack certain portions at the 5′ end of the region under study or contain symbols r, n, s, etc. for individual nucleotides.

The secondary structures of the region under study were predicted by mfold program of M. Zuker [9]. All predictions were performed with 15% suboptimality parameter. For about 50% of non-recurring primary sequences containing frequent DIS, SD or Ψ variants we also used window parameter 0 (with 10% suboptimality parameter) to obtain possible foldings that may be quite similar to one another. Only the optimal structures were considered, i.e. those with free energy within 5% of the minimal one. The RNA fragment to fold covered the region corresponding to that encompassing DIS, SD, and Ψ hairpins in genomic RNA of HIV-1 isolate with accession number NC_001802 (RefSec in NCBI database) plus one nucleotide located immediately upstream and two nucleotides located immediately downstream.

Accession numbers of the HIV-1 isolates mentioned in the paper are given in Appendix*.

Results and Discussion

DIS_{Lai} and DIS_{Mal} hairpins. The primary sequence of DIS hairpin in HIV-1 isolates of group M is the most variable among the hairpins under study. Within CESSHIV-1 database we found 17 DIS variants above the error level (Figs 1, 2). There are 20 variants of DIS hairpin within the set of 350 isolates [7], these variants include 17 ones found in the present database. Presumably, some infrequent variants can be whether above or below the error level upon data set extension.

Within our CESSHIV-1 database, 119 HIV-1 isolates have DIS hairpin identical to that of HIV-1 isolate LAI (DIS_{Lai}) and 85 isolates – identical to that of HIV-1 isolate MAL (DIS_{Mal}). Most of the isolates with DIS_{Lai} belong to B subtype, 8% of these isolates are B/F intersubtype recombinants (IRs). Most isolates with DIS_{Mal} are intersubtype recombinants containing 2–5 different subtypes. The fragments of A, G, E or K subtypes are included into genomes of 81% (52%), 46%, 29% and 19% of isolates of this class, respectively. The genomic fragments of other subtypes occur rather seldom. Two percentages are given for the fragments of A subtype, because it is still argued whether viruses of E subtype are derived from a recombinant lineage (A/E) or they form an independent monophyletic clade [11].

The secondary structures of the region preceding GAG start codon for LAI and MAL isolates predicted by the mfold program are shown in Fig. 1,*A* and 1,*B*, we pictured a A-U-A base triple platform in SD hairpin reasoning from G. K. Amarasinghe et al. data [12]. The structures of DIS hairpins depicted in Fig. 1,*A* and 1,*B* (right box) are similar to those reported in many works on HIV-1 genomic RNA structure [1, 4, 5].

The alternative structures were reported by J. Greatorex et al. [3] and W. Kasprzak [6]. In these structures the DIS hairpin is elongated by involvement of the GGGC sequence, which is a linker between DIS and SD in case of the non-elongated DIS hairpin. The elongated DIS hairpin contains the additional internal loop (4x2). But the GGGC linker, an essential factor in DIS elongation, is found only in 26% of all HIV-1 isolates from our database. For HIV-1 genomes with the most frequent AGC linker, mfold program predicts the structure containing the elongated DIS hairpin with the additional internal loop (4x4) and the disrupted SD hairpin. So, the structures with the elongated DIS hairpin reported in [3] and [6] can be hardly considered as a common, but specific for the isolates with the GGGC sequence adjacent to the 5′ end of the SD hairpin.

The specific features of HIV-1 DIS hairpin are a 6 nt palindrome in the apical loop, purine residues flanking the palindrome and the stem defect, all of them are essential for dimerization process [2, 8]. Out of all possible 64 self-complementary hexanucleotides, only three 6 nt palindromes (GUGCAC, GCGCGC and GUGCGC) are found in the apical loops of HIV-1 DIS hair-

* Names and accession numbers of the isolates shown in Figures are listed below.
671-99T12 – AY423381, 21301 – AF067156, A050 – AF408631, BREPM108 – AY771589, DH12 – AF069140, DJ263 – AF063223, GHNJ175 – AB231893, HH8793 – AF061640, JRCSF – M38429, LAI/BRU – K02013, MAL – X04415, NDK - M27323, RF – M17451, RefSec – NC_001802, U455 – M62320, US4 – AY173955, X254 – AF423755, 90cr056 – AF005496, 92UG001 – AJ320484, 93br020 – AF005494, 94BR-RJ-41 – AY455781, 94CY017.41 – AF286237, 95TNIH022 – AB032740, 96TZ-BF061 – AF289548, 96TZ-BF110 – AF289550, 98BWMO18.d5 – AF443080, 99ZACM4 – AF411964, 00BW0768.20 – AF443089, 00BW2087.2 – AF443104, 00NE95 – AJ508596, 03ZASK043B1 – AY772700, 03ZASK065B1 – AY772694, 04ZASK174B1 – AY901980.

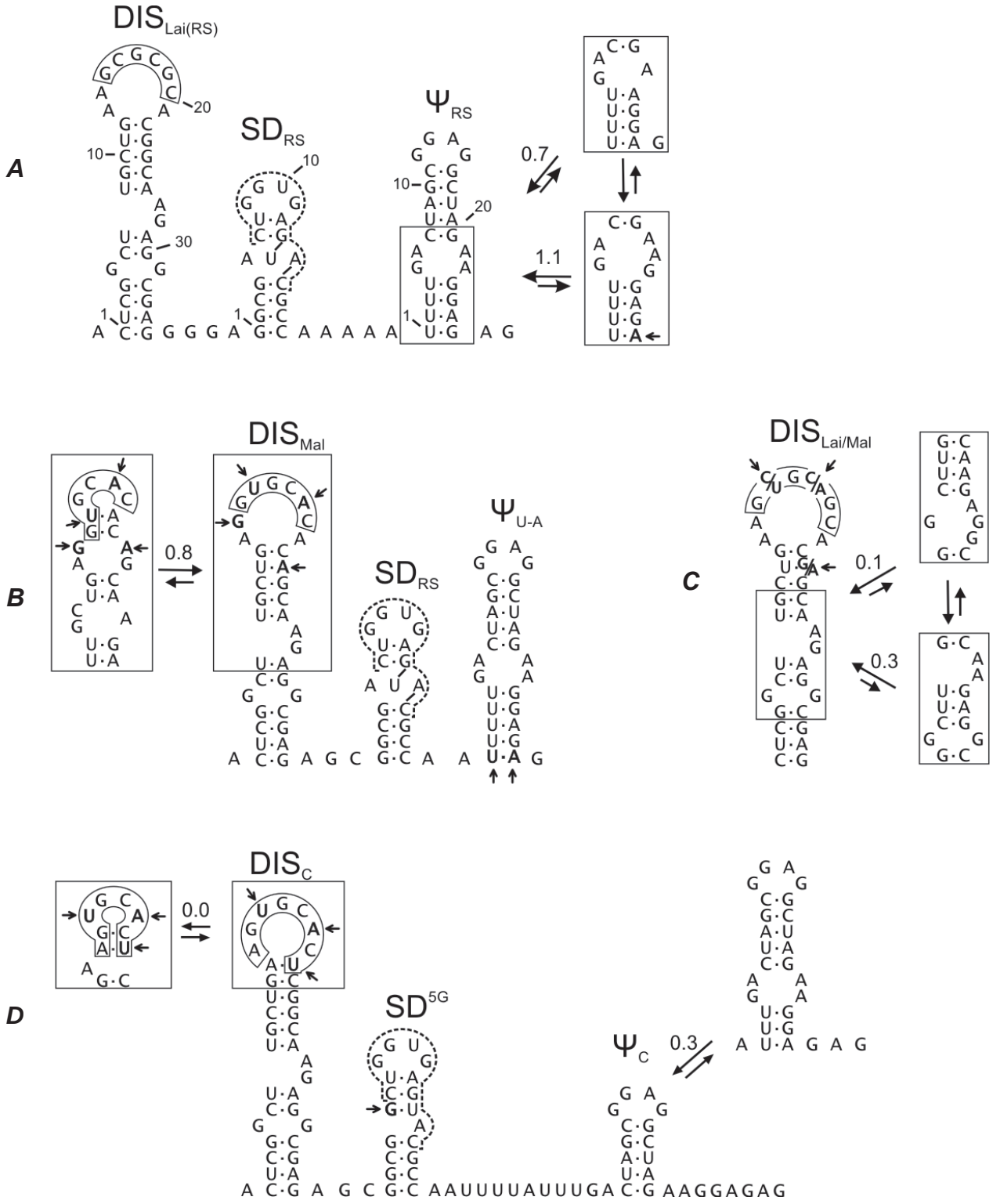

Fig. 1. The secondary structures of the region encompassing DIS, SD and Ψ hairpins in genomic RNAs of HIV-1 isolates LAI (A), MAL (B) and subtype C isolate 00BW2087.2 (D). Nucleotide numeration is given within each hairpin. (C) Stem defect forms in $DIS_{Lai/Mal}$. The palindromes in the upper part of DIS hairpins are boxed. The nucleotides that differ from those in $DIS_{Lai(RS)}$, SD_{RS} and $Ψ_{RS}$ are indicated by arrows. The major splicing donor signal [10] is indicated by dotted line. Hairpin forms are boxed, difference in free energy is given in kcal/mole.

pins [13]. Mfold program predicts three forms of stem defect in DIS_{Lai}, DIS_{Mal} and other DIS hairpins with the stem similar to that of $DIS_{Lai/Mal}$. The first form has AG bulge and GxG internal loop, the second one has asymmetric GxAGG internal loop and the third one has AA bulge and GxG internal loop (Fig. 1,*C*). The second and third forms are less favorable than the first one by 0.1 and 0.3 kcal/mole, respectively. Because of a rapid exchange between stem conformations, it is difficult to determine a structure for DIS stem defect by NMR spectroscopy [3,8].

According to kissing-loop model of HIV-1 dimerization [14, 15] the palindromes of two monomeric RNA molecules interact through Watson-Crick base pairing with formation of a kissing-loop complex. In the presence of nucleocapsid protein (NC) or at high temperature the kissing-loop complex (loose dimer) is converted into more stable form – tight dimer [16–18]. The kissing-loop complexes formed by short RNA constructs derived from HIV-1 DIS sequences are converted to the linear extended duplex ([19] and Refs therein). The tight dimer has been often assumed to be such an extended duplex, however at present this assumption is not experimentally proved [2].

DIS_{Lai} hairpin has GCGCGC palindrome which is flanked by $A^{13}A^{14}$ dinucleotide at the 5′ end, while DIS_{Mal} has GUGCAC palindrome with $A^{13}G^{14}$ flanking at the 5′-end and also U^{11}-$G^{23}\rightarrow U^{11}$-A^{23} substitution in the stem (Fig. 1,*A* and 1,*B*). Of note, base changes in DIS, SD and ψ hairpins are given as compared to RefSec (HIV-1 isolate with accession number NC_001802). The primary sequence of the region under study of RefSec differs from that of LAI isolate only in one position corresponding to the nucleotide adjacent to the 5′-end of the SD hairpin (C in RefSec) and the secondary structures of DIS, SD and ψ hairpins are identical in both isolates.

According to mfold predictions, the upper part of DIS_{Mal} adopts two conformations. The conformation with 9 nt apical loop (Fig. 1,*B*, right box) is more preferable than that with tetraloop (Fig. 1,*B*, left box) by 0.8 kcal/mole.

Dimerization of DIS_{Lai} and DIS_{Mal} hairpins or constructs derived from their sequences has been extensively studied by different biochemical and biophysical methods [13–33]. Unlike DIS_{Lai}, DIS_{Mal} requires magnesium for efficient dimerization [20, 21] which binds to the central part of kissing-loop complex [26, 28, 33]. The structures of the kissing-loop complexes for DIS_{Lai} or DIS_{Mal} constructs reported in literature differ mainly in the location of purines flanking palindrome at the 5′-end. The supposed roles of these purines in dimerization process are various [23, 26, 27, 31]. In particular, M.-R. Mihailescu and J. P. Marino [31] showed that kissing-loop complex displays localized conformational dynamics resulting from protonation of A^{13} at near physiological pH and supposed the non-protonated A^{13} to promote and kinetically trap formation of kissing-loop complex, while the protonated A^{13} to accelerate the rate of DIS dimer structural maturation induced by NC.

Although there is currently no general opinion on the detailed structure of the kissing-loop complex formed by HIV-1 DIS hairpins, it is evident that palindrome sequence and flanking nucleotides play a crucial role in the complex formation. To clarify further the mechanism of dimer formation, it is important to find out other variants of DIS hairpin with tolerable base changes.

DIS hairpin of subtype C isolates. In 287 HIV-1 isolates from our database, the DIS hairpin has three base changes ($C^{16}\rightarrow U^{16}$, $G^{19}\rightarrow A^{19}$ and $A^{21}\rightarrow U^{21}$) in comparison with $DIS_{Lai(RS)}$. Since 97% of the isolates containing this DIS hairpin belong to C subtype, we called it as DIS_C. C subtype is presently the most prevalent among the major circulating HIV-1 subtypes [34]. The rest of isolates with DIS_C are B/C or A/C IRs, except for 94BR-RJ-41 isolate being B/F IR.

Base change $A^{21}\rightarrow U^{21}$ results in palindrome elongation up to 8 nt (AGUGCACU). According to mfold program the upper part of DIS_C adopts two forms and the 8 nt palindrome is completely exposed in none of the forms (Fig. 1,*D*). In the first form (Fig. 1,*D*, right box) the terminal nucleotides in the AAGUGCACU stretch corresponding to the 9 nt apical loop in $DIS_{Lai/Mal}$ form the additional A-U base pair, which leads to elongation of the stem and shortening of the apical loop to 7 nt. The second form exposes the UGCA tetraloop (Fig. 1,*D*, left box). Both forms of DIS_C are equal in free energy, as distinct from DIS_{Mal}, for which the form with tetraloop is much less preferable than that with 9 nt apical loop. DIS_C stem defect is predicted to adopt the same forms as $DIS_{Lai/Mal}$ stem.

Similar to DIS_{Mal}, DIS_C requires magnesium for efficient dimerization [35]. The structure of kissing-loop complex formed by DIS_C hairpins is not reported in literature. The role of DIS_C form with UGCA tetraloop in dimerization process seems intriguing, its existence should be experimentally proved.

According to E. Ennifar et al. [26], the structures of kissing-loop complexes described for DIS_{Lai} and DIS_{Mal} also seem relevant to the class of DISs with seven, rather than nine, bases in the loop, possibly to DIS_C (the first form). On the other hand,

the additional A-U base pair increases the stability of the DIS_C stem, and it may verge towards non-optimal for formation of the extended duplex, since according to K.-I. Takahashi et al. [25] the stem of DIS hairpin is supposed not to be highly stable to make possible the conformation transition from kissing-loop complex to extended duplex. Furthermore, the base pair closing the apical loop in DIS_C is A-U instead of G-C in DIS_{Mal} or DIS_{Lai}. As a formation of the kissing-loop complex requires high structural compatibility between the stem conformation and the apical loop conformation of DIS hairpin [36, 37], the structure of kissing-loop complex formed by DIS_C hairpins may be distinct from that formed by DIS_{Mal} or DIS_{Lai} hairpins.

Other variants of DIS hairpin. Ten DIS variants can be considered as those evolved from DIS_{Lai}, DIS_{Mal} or DIS_C (Fig. 2). The structure of a kissing-loop complex formed by DIS_{Lai} hairpins containing U^{11}-G^{23}→U^{11}-A^{23} base pair substitution (DIS_{Lai}^{23A}, Fig. 2,*A*) probably differs from that of LAI isolate. Using molecular dynamics simulation method to study a DIS_{Lai} construct dimerization, S. Aci et al. demonstrated that G^{12}-C^{22} base pair is equally likely to be found in the open state or in the closed state in the kissing-loop complex [30]. Since the labile U^{11}-G^{23} base pair at the penultimate position of the stem is essential for this equilibrium, its substitution for U^{11}-A^{23} could affect the state of G^{12}-C^{22} base pair. Anyhow the G-U base pair plays a special role in RNA structure, this base pairing is complex and dependent on the secondary and tertiary context [38]. So, the U^{11}-G^{23}→U^{11}-A^{23} base pair substitution may modulate DIS dimer structure.

Remarkably, the occurrence of the substitution at stem position 23 significantly varies in three main DIS variants, DIS_{Lai}, DIS_{Mal} and DIS_C (23%, 4,5% and 0.7%, respectively). These findings are consistent with the idea of F. Kieken et al. [37] of structural compatibility between the stem and apical loop conformations. To support this idea, it would be promising to ascertain how alterations at position 23 affect dimerization efficiency of different DIS variants.

For subtype C isolates, another reason of intolerance of U^{11}-G^{23} base pair substitution for more stable U^{11}-A^{23} may lie in the fact that DIS hairpin stem is already stabilized by the additional A^{13}-U^{21} base pair and the further stabilization impedes the transition to extended duplex.

Dimerization properties of DIS_{Lai} variant with A^{21}→G^{21} base change at the 3′ end of the apical loop (DIS_{Lai}^{21G}, Fig. 2,*A*) have been investigated by H. Huthoff et al. [39] as a part of the study on unusual DIS hairpin with 15 nt apical loop, this hairpin with long apical loop is discussed below in this section. The A^{21}→G^{21} base change extending the palindrome was found to enhance dimerization markedly [39]. According to mfold program predictions, DIS_{Lai}^{21G} form with minimal free energy is crowned by tetraloop GCGC (Fig. 2,*A*). The second form exposing 6 nt palindrome is significantly less favorable (by 1.5 kcal/mole) than the first one. Therefore DIS_{Lai}^{21G} dimerization is supposed to proceed mainly via the form crowned by the tetraloop. Since the kissing-loop complex formation was registered between DIS hairpins with the apical loops AAGCA [29] or GACG [40] containing only 2 nt palindromes, we suppose that DIS_{Lai}^{21G} may initially kiss through GCGC tetraloops and then through the complete palindromes. Possibly such two-step mechanism of kissing facilitates dimerization of DIS with AA[GCGCGC]G stretch corresponding to the apical loop. However a tetraloop formation in DIS_{Lai}^{21G} has to be experimentally proved.

According to mfold predictions, the upper section of DIS_{Lai} variant with A^{13}→C^{13} (DIS_{Lai}^{13C}) adopts three forms with almost equal free energies (Fig. 2). The first form exposes 9 nt apical loop, the second one has 5 nt apical loop and the third one is crowned by tetraloop. DIS_{Lai}^{13C} is the only DIS variant with pyrimidine flanking palindrome. J.-C. Paillart et al supposed that formation of non-canonical base pair between the first and the last nucleotides in DIS apical loop can facilitate dimerization by structuring the monomeric DIS loop [2]. According to J. S. Lodmel et al., a non-canonical base pair (sheared) can be formed not only between A^{13} and A^{21} (as in $DIS_{Lai/Mal}$) but also between C^{13} and A^{21} [36]. Presumably A^{13}→C^{13} substitution in DIS hairpin of HIV-1 isolates is tolerated because of non-canonical C^{13}·A^{21} base pair formation.

The A deletion at position 13 is also tolerated in DIS_{Lai} ($DIS_{Lai}^{\Delta 13}$, Fig. 2,*A*). This deletion is found in other DIS_{Lai} variants occurring in our database below the level of errors.

As distinct from the alterations at the 5′-end of the DIS_{Lai} apical loop (A^{13}→C^{13} and A^{13} deletion), the A insertion at this end is tolerated in DIS_{Mal} (DIS_{Mal}^{Ains}, Fig. 2,*B*) and DIS_C (DIS_C^{Ains}, Fig. 2,*C*). This insertion occurs 1.5 times more frequently in DIS_{Mal} than in DIS_C.

Since A insertion at the 5′-end of the apical loop is not tolerated in HIV-1 isolates with DIS_{Lai}, such insertion is believed to distort the kissing-loop complex formed by GCGCGC palindromes. Though 4 isolates with A insertion between G^{12} and A^{13} in DIS_{Lai} apical loop are found in our database, three of them have this mutation in combination

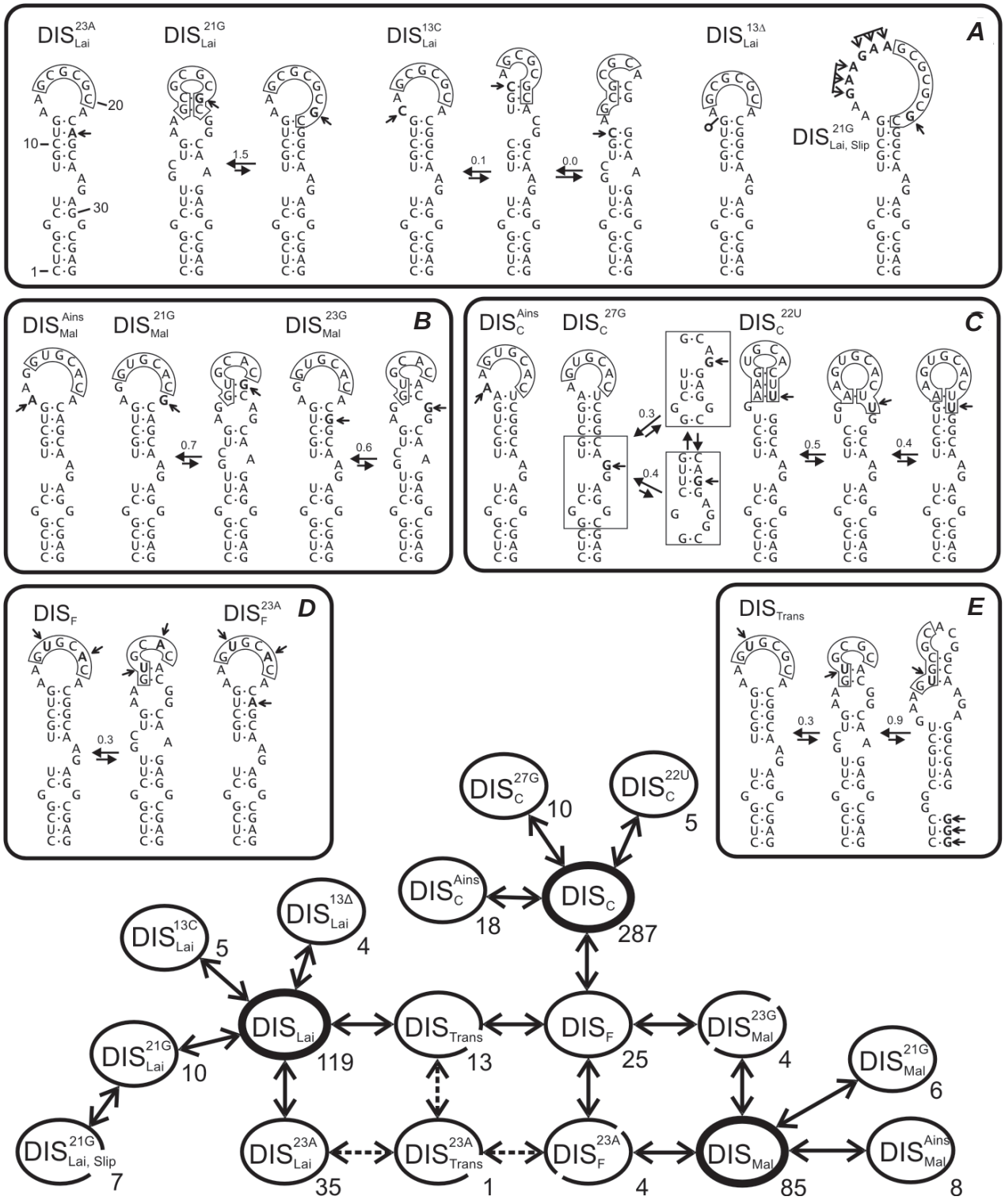

Fig. 2. DIS variants in HIV-1 genomic RNAs and the scheme of hypothetical transitions between them via a single base change. The palindromes in the upper part of DIS hairpins are boxed. The base changes in DIS variants as compared to "parental" DIS (A-C) and as compared to $DIS_{Lai(RS)}$ (D-E) are indicated by arrows. Difference in free energy between hairpin forms is given in kcal/mole. (A) DIS_{Lai}^{23A} (NDK), DIS_{Lai}^{21G} (RF), DIS_{Lai}^{13C} (X254), $DIS_{Lai}^{13\Delta}$ (92UG001), $DIS_{Lai, Slip}^{21G}$(671-99T12); (B) DIS_{Mal}^{Ains} (00NE95), DIS_{Mal}^{21G} (DJ263), DIS_{Mal}^{23G} (94CY017.41); (C) DIS_{C}^{Ains} (98BWMO18-d5), DIS_{C}^{27G} (03ZASK065B1), DIS_{C}^{22U} (00BW0768.20), DIS stem forms are boxed; (D) DIS_{F} (93br020), DIS_{F}^{23A} (GHNJ175); (E) DIS_{Trans} (US4). In the scheme the number of isolates with DIS variant is given.

with $A^{21}\rightarrow U^{21}$ base change, which results in the inserted A elimination from DIS apical loop due to base pair formation between this A and U^{21}.

The upper section of $DIS_{Mal}{}^{Ains}$ has two forms, the first form has 10 nt apical loop with 6 nt palindrome (Fig. 2,*B*), the second one is crowned by GCAC tetraloop (not shown). Since the first form of $DIS_{Mal}{}^{Ains}$ is much more preferable in free energy (by 1.5 kcal/mole) than that with a tetraloop, the first form of $DIS_{Mal}{}^{Ains}$ is likely to play a main role in dimerization process. Dimerization efficiency of the dominant form of $DIS_{Mal}{}^{Ains}$ is supposed to be lower than that of DIS_{Mal}, since according to N. Pattabiraman [27] the occurrence of more than three linker nucleotides with six or seven nucleotides forming the kissing stem is unfavorable for formation of the stable kissing-loop complex.

Like DIS_{Lai}, DIS_{Mal} tolerates the $A^{21}\rightarrow G^{21}$ base change at the 3′-end of its apical loop ($DIS_{Mal}{}^{21G}$, Fig. 2,*B*), this base change rather occurs in DIS_{Mal} with the similar frequency than in DIS_{Lai}. Mfold program predicts two forms of the upper section of $DIS_{Mal}{}^{21G}$, the first form with 9 nt apical loop is more preferable by 0.7 kcal/mole than the second one crowned by GCAC tetraloop as distinct from $DIS_{Lai}{}^{21G}$ for which the form with a tetraloop is dominant.

As mentioned above, DIS_C stem defect adopts three forms. Three forms of the stem defect in DIS_C variant with $A^{27}\rightarrow G^{27}$ base change ($DIS_C{}^{27G}$) are similar to those in common DIS stem, this base change is located in a bulge of the first form or the second form, but in the double-stranded section of the third one (Fig. 2,*C*). The DIS stem moderately interacts with nucleocapsid protein [8, 41]. This interaction may play its role in conversion of the kissing-loop complex into the extended dimer [8, 25]. Y. Yuan [8] supposed that NC bound to the stem defect form with 1x3 internal loop more effectively than to alternative forms. Since the difference in free energy between $DIS_C{}^{27G}$ stem form with GxAGG internal loop and its dominant form (0.4 kcal/mole, Fig. 2,*C*) is greater than that in $DIS_{Lai/Mal}$ (0.1 kcal/mole, Fig. 1,*C*), NC affinity for mutated DIS stem is expected to decrease. But it is not clear how this protein interacts with GG bulge of the dominant form of $DIS_C{}^{27G}$.

The upper part of DIS_C variant with $C^{22}\rightarrow U^{22}$ base change ($DIS_C{}^{22U}$) is predicted to have three forms. The most favorable structure is crowned by UGCA tetraloop, the second form has a 7 nt apical loop and an additional UGxUG internal loop and the third one has 7 nt apical loop (Fig. 2).

The isolates with variants of DIS_{Lai}, DIS_{Mal} and DIS_C belong to the same subtypes as the isolates with "parental" DIS hairpins, except for isolates with $DIS_{Mal}{}^{Ains}$ and $DIS_{Lai}{}^{23A}$. Almost all isolates with $DIS_{Mal}{}^{Ains}$ are B/G IRs. As many as 20% of isolates with $DIS_{Lai}{}^{23A}$ are C/D, B/C or A/C/D recombinants. The high percentage of C-containing recombinants among the isolates with $DIS_{Lai}{}^{23A}$ and a very low percentage of such recombinants among isolates with DIS_{Lai} (0.8%) give an evidence of importance of $G^{23}\rightarrow A^{23}$ base change in intersubtype recombination between viruses of B, D subtypes (with DIS_{Lai}) and viruses of subtype C (with DIS_C). As the rate of intersubtype recombination between subtype B virus and subtype C virus is low due to the difference in DIS sequences [42], $G^{23}\rightarrow A^{23}$ is believed to facilitate a formation of dimer between DIS_{Lai} and DIS_C. As we mentioned in the beginning of the section this base change might modulate the structure of kissing-loop complex.

Three DIS variants seem to be aside from the main variants. Among them, there is the unusual DIS hairpin with long apical loop AAGAAGAA[GCGCGC]G of a country specific origin (Netherlands) The long apical loop has GAAGAA insertion in the 5′-terminal part and A→G substitution at the 3′-end as compared to DIS_{Lai} apical loop sequence (Fig. 2,*A*). H. Huthoff et al. [39] reported that the insertion abolishes dimerization, while both mutations decrease dimerization slightly, i.e. a detrimental effect of the large insertion is compensated by A→G base change at the 3′-end of the apical loop. According to the authors [39], the GAAGAA insert being double repeat of the preceding GAA triplet may have occurred through slippage during reverse transcription, so we called the DIS hairpin with long apical loop as $DIS_{Lai}{}^{21G}{}_{Slip}$ (Fig. 2,*A*). Interestingly DIS hairpin of isolate BREPM108 from our database has AAGAA[GUGCGC]A apical loop with a single GAA repeat, which implies that the occurrence of GAA repeat(s) in DIS hairpin apical loop is not an accidental event.

The DIS hairpin with AA[GUGCAC]A apical loop and the stem like in DIS_{Lai} is reported in literature as DIS_F [24, 33]. However our phylogenetic data evidence that this designation could be considered as conditional, since only 36% of isolates with DIS_F are of subtype F or recombinants containing this subtype, while the same percentage of isolates with DIS_F belong to C subtype. By mfold predictions, DIS_F exists in two energetically close forms (Fig. 2,D). According to M. Laughrea [22] DIS_F has dimerization properties similar to those of MAL isolate. The structures of the kissing-loop complex formed by DIS_F constructs have been studied by NMR [43] and X-ray crystallography [33]. The reported structures are similar except for A^{13} and A^{14} location. The NMR struc-

ture [43] showed a bulged-in conformation of A^{13} and A^{14}, while two crystal forms of DIS_F dimer [33] demonstrated a bulged-out conformation of these adenines. According to E. Ennifar [33] one of the crystal forms of DIS_F kissing-loop complex is similar to that of DIS_{Lai} or DIS_{Mal}.

The third "alien" DIS variant has the AA[GUGCGC]A apical loop and the stem as in DIS_{Lai}. M. Laughrea et al. [13] consider the GUGCGC palindrome to be a transition sequence linking GCGCGC to GUGCAC via single point-mutations, so we called DIS with GUGCGC palindrome as DIS_{Trans}. Like DIS_F the DIS_{Trans} has dimerization properties similar to those of MAL isolate [22].

By mfold predictions, DIS_{Trans} adopts three forms, the most favorable form has 9 nt apical loop, the second form crowned by GCGC tetraloop is energetically close to the first one, difference in free energy is 0.3 kcal/mole (Fig. 2,*E*). Of note, the second form of DIS_{Trans} is similar to the dominant form of $DIS_{Lai}{}^{21G}$ (Fig. 2,*A*), these forms differ only in the base pair in the upper section of the stem (U^{16}—A^{21} and C^{16}—G^{21}, respectively). Presumably the two-step mechanism of kissing, which we proposed for $DIS_{Lai}{}^{21G}$, is also relevant for DIS_{Trans}. The third form of DIS_{Trans} exposes GCACG pentaloop and involves the GGG(C) linker between DIS and SD hairpins (Fig. 2). This form is by 1.2 kcal/mole less favorable than the first one. The conformation similar to the third form of DIS_{Trans} is also formed in HIV-1 genomes with DIS_F and GGG(C) linker (not shown), but this linker occurs much more rarely in isolates with DIS_F than in those with DIS_{Trans}.

Similar to HIV-1 isolates with DIS_{Lai}, most of the isolates with DIS_{Trans} have the GGGC linker between DIS and SD and belong to B subtype or are B-containing IRs (though B-containing IRs are prevailing among the isolates with DIS_{Trans}, while subtype B isolates are prevailing among those with DIS_{Lai}). Hence we may suppose that DIS_{Trans} has evolved from DIS_{Lai} through a single base change. In turn DIS_F, which is close to DIS_{Trans} both in secondary structure and dimerization properties, is supposed to evolve from DIS_{Trans}. All DIS variants are arranged in the scheme of hypothetical transitions between them via a single base change (the bottom part of Fig. 2). DIS_{Mal} is supposed to evolve from either $DIS_F{}^{23A}$ or $DIS_F{}^{14G}$ (otherwise they could be called as $DIS_{Mal}{}^{14A}$ and $DIS_{Mal}{}^{23G}$, respectively). All transition directions via $DIS_{Trans}{}^{23A}$ variant are believed to be less probable since only one isolate with $DIS_{Trans}{}^{23A}$ was found in our database. As seen from the scheme, every of three main DIS variants (DIS_{Lai}, DIS_{Mal} and DIS_C) can be the starting point of transitions between these variants, for instance, DIS_F may be mutated from DIS_C via $U^{21}\rightarrow A^{21}$ base change.

Variants of SD hairpin. SD and Ψ hairpins are the most preferable sites for nucleocapsid protein binding in the 5′-terminal region of HIV-1 genomic RNA [41, 44]. Based on NMR structures of NC bound to complete SD [45] and Ψ construct [46], G. K. Amarasinghe et al. [45] supposed that NC protein similarly interacted with tetraloops of both hairpins but differently with their stems, in particular, it interacted with A-U-A base triple platform in the SD stem. The authors [45] supposed that NC binds to SD and Ψ hairpins in an adaptive manner via different subsets of inter and intramolecular interactions, which had been supported by the study of NC complexes with complete SD hairpin and Ψ construct by linear-scaling quantum methods [47]. In the light of the idea of adaptive NC-RNA binding, it is reasonable to suppose that interaction of nucleocapsid protein with SD and Ψ variants discussed below is also adaptive.

In CESSHIV-1 database there are eight SD hairpin variants above the error level vice 14 ones in the set of 350 isolates, new variants did not appear. About 60% of HIV-1 isolates in our database have SD hairpin identical to that in LAI/MAL isolates or in the isolate with accession number NC_001802 (SD_{RS}, Fig. 1,*A* and 1,*B*). Exactly this SD variant has been studied by G. K. Amarasinghe et al. [45] and J. Khandogin et al. [47]. In 49 isolates SD_{RS} is elongated by C-G base pair ($SD_{C\text{-}G}$) due to linkers involvement (Fig. 3,*A*), most isolates with $SD_{C\text{-}G}$ belong to C subtype.

In 196 HIV-1 isolates from our database, A residue at position 5 of SD stem is substituted for G (SD^{5G}), this base change prevents the formation of the A-U-A base triple platform (Fig. 1,*D*). About 90% of these isolates belong to C subtype. Though $A^5\rightarrow G^5$ base change in SD hairpin occurs primarily in C subtype HIV-1 isolates, it cannot be considered as a hallmark for this subtype, since about a half of C subtype isolates have a common SD hairpin without this substitution.

Unlike rather frequent SD_{RS} elongation by C-G base pair, such elongation is found only in two isolates with SD^{5G} (i.e. below the error level).

Other SD variants are much less frequent than SD_{RS}, SD^{5G} and $SD_{C\text{-}G}$. Among them there is the variant with double base change $C^6U^7\rightarrow A^6C^7$ in the stem (SD^{AC}). According to mfold program predictions, this double base change leads to elongation of SD apical loop and appearance of AxA internal loop (Fig. 3,*B*). Almost all isolates with SD^{AC} belong to B subtype and they originate from Canada or USA.

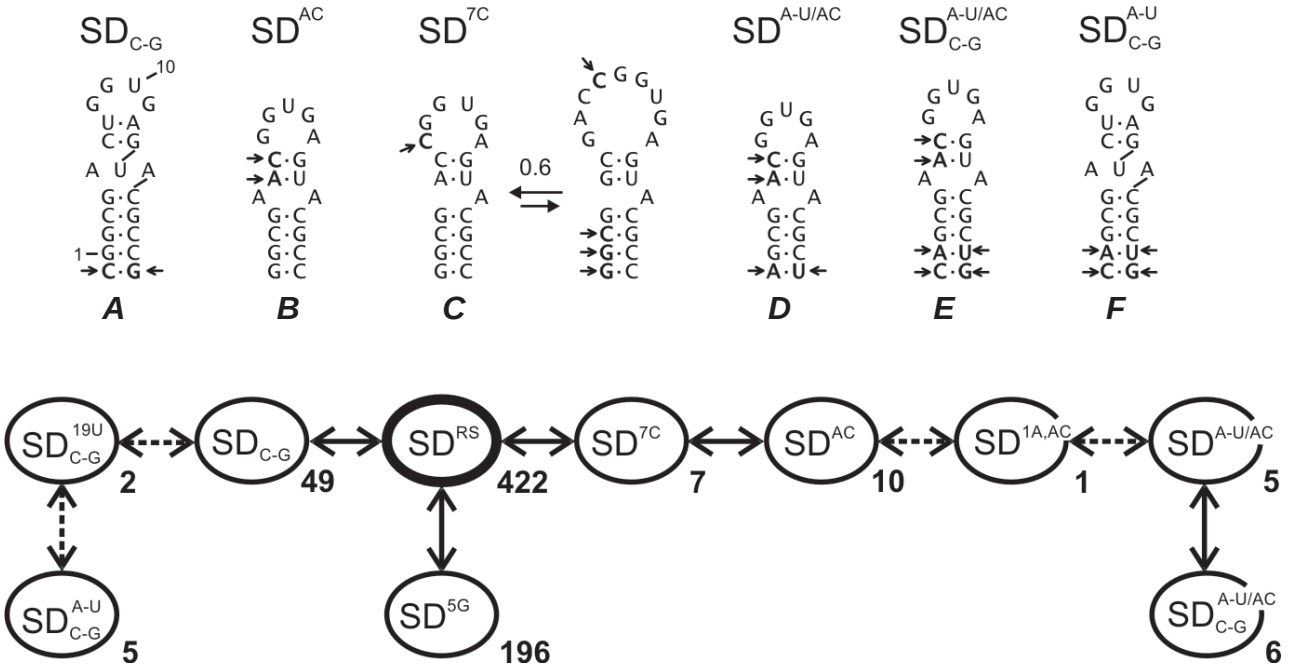

Fig. 3. SD hairpin variants in HIV-1 genomic RNAs and the scheme of hypothetical transitions between them via a single base change. The nucleotides that differ from those in SD_{RS} are indicated by arrows. Difference in free energy between hairpin forms is given in kcal/mole. (A) SD_{C-G} (JRCSF), (B) SD^{AC} (DH12), (C) SD^{7C} (A050), (D) $SD^{A-U/AC}$ (96TZ-BF061), (E) $SD^{A-U/AC}{}_{C-G}$ (96TZ-BF110), (F) $SD^{A-U}{}_{C-G}$ (99ZACM4). In the scheme the number of isolates with SD variant is given.

The SD^{AC} may be evolved from the SD variant with a single base change $U^7 \rightarrow C^7$ (SD^{7C}, Fig. 3,*C*), since the base change $U^7 \rightarrow C^7$ is much more frequent than $C^6 \rightarrow A^6$. The SD^{7C} adopts two conformations with 6 and 9 nt apical loops, respectively (Fig. 3,*C*). The second form partly involves GGGC linker and it is less stable by 0.6 kcal/mol than the first one. HIV-1 isolates with SD^{7C} mainly belong to B subtype or are B-containing IRs. Almost all these isolates originate from Northern or Southern America.

The double base change $C^6U^7 \rightarrow A^6C^7$ also occurs in combination with G^1-$C^{19} \rightarrow A^1$-U^{19} base pair substitution ($SD^{A-U/AC}$, Fig. 3,*D*). This combination is also found in SD hairpin elongated by C-G base pair ($SD^{A-U/AC}{}_{C-G}$, Fig. 3,*E*). Almost all isolates with $SD^{A-U/AC}$ or $SD^{A-U/AC}{}_{C-G}$ are D-containing IRs and originated from Tansania and Kenia.

In the last SD variant, the stem has G^1-$C^{19} \rightarrow A^1$-U^{19} base pair substitution and is elongated by C-G ($SD^{A-U}{}_{C-G}$, Fig. 3,*F*). These isolates mostly belong to D subtype (or D-containing IRs) and they originated from Central and Southern Africa.

It is worth noting that almost all HIV-1 isolates in our database with G^1-$C^{19} \rightarrow A^1$-U^{19} base pair substitution in SD (SD^{A-U}, $SD^{A-U/AC}$ and $SD^{A-U}{}_{C-G}$) are of D subtype or D-containing IRs. Since each of 9 isolates of D subtype from the database CESSHIV-1 has SD hairpin with this base pair substitution, it is believed to be specific for D subtype. The subtype specificity for G^1-$C^{19} \rightarrow A^1$-U^{19} base pair substitution is firstly reported in this work, it is expected to be valid for a greater set of HIV-1 isolates.

All SD variants found are arranged in the scheme of hypothetical transitions between them via a single base change. In this scheme, the transitions to SD variants elongated by C-G base pair are presented as one step, because in 88% of HIV-1 isolates the nucleotide adjacent to 5′-end of SD hairpin is already C. Of note, upon extension of HIV-1 isolates database other transition ways to $SD^{A-U}{}_{C-G}$ or $SD^{A-U/AC}{}_{C-G}$ could emerge.

Variants of Ψ hairpin. Within the present database we found 9 Ψ hairpin variants above the error level vice 10 ones within the set of 350 isolates [7]. Genomes of 182 HIV-1 isolates studied here have Ψ hairpin sequence identical to that in LAI isolate or in the isolate with accession number NC_001802 (Ψ_{RS}, Fig. 1,*A*).

The long Ψ_{RS} was reported in some works (for example [4, 41]), while most authors depicted truncated Ψ hairpin without any internal loop (like that

shown in Fig. 1,*D*, the first form). The open-closed state of the nucleotides corresponding to the bottom section of the long Ψ hairpin (Fig. 1,*A*) is difficult to identify because they interact weakly both with single-strand-specific and double-strand-specific agents [4, 5, 48]. Chemical modifications assay directly in infected cells and virions demonstrated that A^{25} in the bottom part of Ψ hairpin highly reacted with Me_2SO_4 [49]. However this fact does not evidence unambiguously that this nucleotide is unpaired because chemically modified nucleotides are allowed in or adjacent to G-U pairs anywhere [50]. Since the bottom section of the long Ψ hairpin is formed by low stable U-G and U-A base pairs, we suppose that they can be easily melted during the helix "breathing" and temporarily exposed to chemical agents and nucleases.

The Ψ_{RS} is predicted to have three forms, which differ from each other in internal loops only (Fig. 1,*A*). The most favorable structure has GAxAA internal loop. The second form with GAxA internal loop and the third form with GAxAAG internal loop are less favorable than the first one by 0.7 and 1.1 kcal/mole, respectively. So, similar to DIS stem, Ψ stem is also structurally unstable. Possibly this flexibility is important for interaction between NC protein and Ψ (DIS) hairpins. NC-Ψ interaction has been investigated mainly on Ψ construct that is short Ψ hairpin (7–20 nt) elongated by one A-U and two G-C base pairs [41, 46, 47, 51]. The evidence of specific interaction between NC and the stem defect of ψ hairpin follows from enzymatic probing of HIV-1 untranslated leader [51]. This study shows that NC contacts the nucleotides corresponding to long Ψ hairpin internal loop and the adjacent nucleotides.

Almost all HIV-1 isolates from our database have AG residues adjacent to 3′-end of the stretch corresponding to Ψ_{RS}. One or both of these nucleotides are involved in Ψ_{RS} elongation in 152 isolates which have U, UU or CU residues adjacent to 5′-end of Ψ_{RS} ($\Psi_{U\text{-}A}$, $\Psi_{UU\text{-}AG}$ or $\Psi_{CU\text{-}AG}$). Most of these isolates have Ψ_{RS} elongated by U-A base pair, for example, MAL isolate ($\Psi_{U\text{-}A}$, Fig. 1,*B*). $\Psi_{U\text{-}A}$ stem has the second form with GAxA internal loop (not shown) which is less favorable by 1.4 kcal/mole than the first one. For $\Psi_{UU\text{-}AG}$, mfold program predicts two forms: with GAxAA internal loop and GAxA internal loop (Fig. 4,*A*). Ψ_{RS} elongated by U-A and C-G base pairs has only one form ($\Psi_{CU\text{-}AG}$, Fig. 4,*B*). Many isolates with elongated Ψ hairpin belong to B, A, D subtypes or are B/F, B/G IRs.

As many as 271 isolates in our database have Ψ hairpin with $U^1 \rightarrow A^1$ base change, 96% of these isolates belong to C subtype and the rest (except for one isolate) are C-containing IRs. On the other hand, 93% of subtype C isolates have such substitution. Most of subtype C isolates without this substitution have A residue immediately upstream of the Ψ hairpin, which may result from A displacement into the linker sequence because of U insertion at the first position of the hairpin. Therefore the $U^1 \rightarrow A^1$ base change in Ψ hairpin can be considered as a hallmark for subtype C isolates, and we called this hairpin as Ψ_C.

The Ψ_C hairpin adopts two main forms, the truncated one and the shortened by one base pair (Fig. 1,*D*). Since the truncated form is more favorable than the shortened one by 0.3 kcal/mole only, these forms are supposed to co-exist.

Most isolates with Ψ_C hairpin have oligo(U) tract adjacent to the 5′-end of the sequence corresponding to Ψ_{RS}. When the length of this tract is equal or above 4 nucleotides, the stem of Ψ_C hairpin adopts additional conformations involving the tract (not shown).

The Ψ hairpin with the base change $A^{21} \rightarrow G^{21}$ combined with the $U^1 \rightarrow A^1$ base change (Ψ_C^{21G}, Fig. 4,*C*) is a new variant found in 16 HIV-1 isolates in the present database, in the previous database it was found in only one isolate that was below the rate of sequencing and database errors. By mfold program predictions, the $A^{21} \rightarrow G^{21}$ base change results in the highly stable Ψ hairpin with GAxGA internal loop, its stability is higher by 2 kcal/mole than the second form of Ψ_C (Fig. 1,*D*). It should be noted that the isolates with Ψ_C^{21G} have been collected during the period of 2000–2004. Though the date of collection of one isolate with this hairpin is not indicated, it was also submitted to NCBI in this period. So this Ψ hairpin variant may evolve very recently.

The base change $G^{12} \rightarrow A^{12}$ is found in 10 isolates with Ψ_{RS} (Ψ^{12A}), 9 isolates with Ψ_C (Ψ_C^{12A}) and 6 isolates with $\Psi_{U\text{-}A}$ ($\Psi_{U\text{-}A}^{12A}$). In all these Ψ variants the base change $G^{12} \rightarrow A^{12}$ results in substitution of GGAG apical loop for AGAG and does not alter stem conformations (Fig. 4,*D*-*F*). A. C. Paoletti et al. supposed that $G^{12} \rightarrow A^{12}$ in Ψ hairpin is tolerated because A^{12} in AGAG loop and G^{12} in GGAG loop similarly stack upon the base-paired stem and do not make specific H-bonds with NC protein [51]. Of note, within our database Ψ hairpin with the base change $G^{12} \rightarrow A^{12}$ is found in the isolates with DIS_C, DIS_{Mal}, DIS_F or their variants only, but not with DIS_{Lai} (B subtype). The deficiency of the base change $G^{12} \rightarrow A^{12}$ in Ψ hairpin of subtype B isolates implies that NC protein of these isolates interacts with the AGAG apical loop with lower efficiency than that of subtypes C, F isolates or A-containing IRs (with DIS_{Mal}).

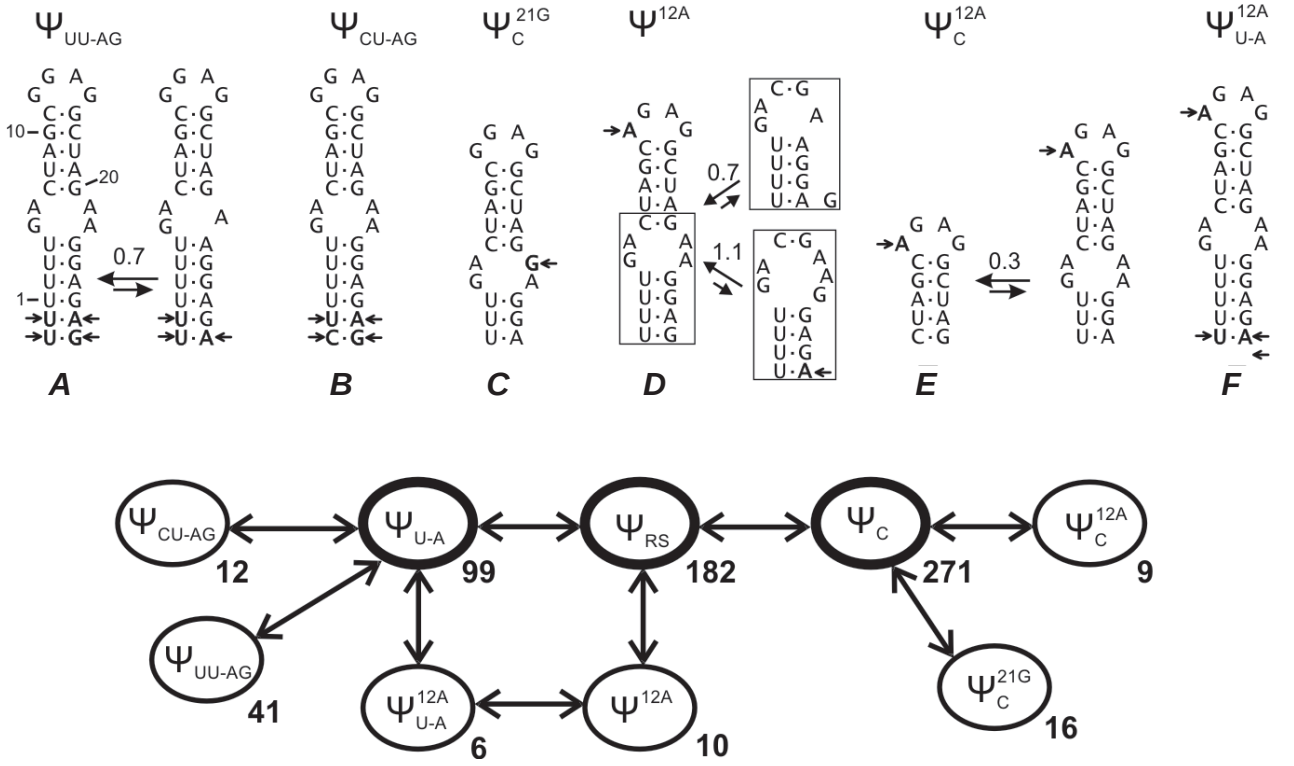

Fig. 4. Ψ hairpin variants in HIV-1 genomic RNAs and the scheme of hypothetical transitions between them via a single base change. The nucleotides that differ from those in Ψ_{RS} are indicated by arrows. Stem defect forms are boxed. Difference in free energy between hairpin forms is given in kcal/mole. (A) $\Psi_{UU\text{-}AG}$ (U455), (B) $\Psi_{CU\text{-}AG}$ (HH8793), (C) Ψ_C^{21G} (04ZASK174B1), (D) Ψ^{12A} (21301), (E) Ψ_C^{12A} (94CY017.41), (F) $\Psi^{12A}_{U\text{-}A}$ (95TNIH022). In the scheme the number of isolates with Ψ variant is given.

All Ψ variants are arranged in the scheme of hypothetical transitions between them via a single base change (the bottom part of Fig. 4) as done for DIS and SD variants. As seen from this scheme, variants Ψ_{RS}, Ψ_C and $\Psi_{U\text{-}A}$ can be considered as "parental" ones for other Ψ variants.

Linkers. In our previous work we formally reported the linkers between DIS and SD hairpins or SD and Ψ hairpins as the stretches corresponding to these linkers in HIV-1 isolate with accession number NC_001802 (Tables 3 and 4 [7]) and indicated that in some isolates one or both of linkers are shortened/extended because of some base changes in the region under study. In the present paper the actual linkers' sequences occurring above the level expected from sequencing and database errors are listed in Tables 1 and 2.

The linker between DIS and SD hairpins is rather conservative, about 70% of HIV-1 isolates from our database are found to contain AGC or GGGC as DIS/SD linker (Table 1). The AGC linker occurs in isolates of different subtypes, while the GGGC linker occurs mainly in isolates of subtype B. The conservation of DIS/SD linker can be explained by its involvement in the G-quadruplex structure. We earlier reported a model of dimer linkage structure in HIV-1 that includes both duplex formed by DIS hairpins and quadruplex domains formed by conservative G-rich stretches located immediately downstream of DIS hairpin, including the DIS/SD linker portion [53, 54].

The linker between the SD and Ψ hairpins is much more heterogeneous both in length and sequence (Table 2) than DIS/SD linker. In most isolates of B subtype the SD/Ψ linker is oligo(A) tract of 3–5 nt in length. A majority of subtype C isolates have SD/Ψ linker consisting of A and U residues, for example AAUUUUA(UUUGA) (isolate 00BW2087.2, Fig. 1,*D*), AAUUUUUA(UUUGA) or AAAUUUUA(UUUGA).

A high variability of the SD/Ψ linker and a high conservation of the stretch UUUUGA, which forms the bottom part of Ψ_{RS} hairpin, may further evidence that this stretch is rather a part of long Ψ hairpin as predicted by mfold program (Fig. 1,*A*) than a part of the linker.

Summary data on the secondary structure of the region encompassing DIS, SD and Ψ hairpins

*Table 1. Linker sequences between DIS and SD hairpins in HIV-1 isolates**

No	Linker sequences	The number of isolates	Isolate subtype
1	AGC	337	Mainly of C subtype, rather frequently of A, E (A/E), G subtypes or B/G, A/G IRs
2	GGGC	201	Mainly of B subtype, rather frequently B/F IRs or of C subtype
3	AGG	48	Mainly of C subtype
4	GGG	44	Rather frequently of C subtype
5	AG	41	Mainly of C subtype
6	AGA	9	Mainly of C subtype
7	AGGC	7	Mainly of C subtype
8	GGGAC	6	Mainly unknown or of B, C subtypes
9	G	5	Mainly of B subtype
10	GGC	4	Mainly of C subtype

* Linker sequence is given for the most energetically favorable structure of the region under study.

*Table 2. Linker sequences between SD and Ψ hairpins in HIV-1 isolates **

No	Linker sequences	The number of isolates	Isolate subtype
1	AAAUUUUA(UUUGA)	115	Almost all of C subtype
2	AAA	88	Rather frequently of E (A/E), B subtype or B/F, B/G IRs
3	AAAAA	85	Mainly of B subtype
4	AAUUUUA(UUUGA)	82	Almost all of C subtype
5	AAAA	77	About a half of isolates belong to B subtype, rather frequently B/F IRs
6	AAUUUUUA(UUUGA)	64	All of C subtype
7	AA	40	Rather frequently of C subtype or A-containing IRs
8	AAAAAA	18	Rather frequently of B subtype
9	AUUUUUA(UUUGA)	15	Almost all of C subtype
10	AUUUUA(UUUGA)	13	Almost all of C subtype
11	UAAAA	13	Mainly unknown, rather frequently of A subtype
12	A	10	Mainly of A subtype or A-containing IRs
13	AUAAA	8	Mainly B/G IRs or of G subtype
14	AAAAUUUUA(UUUGA)	7	All of C subtype
15	AUA	6	A half of isolates are A/F/G/K/U IRs
16	AAAG	5	All of B subtype
17	AAAUUUA(UUUGA)	4	All of C subtype

* Linker sequence is given for the most energetically favorable structure of the region under study. For isolates with Ψ_C or Ψ_C^{12A}, the UUUGA tract is given in brackets, since this tract is a part of the SD/Ψ linker in the first structure, while it is involved in Ψ hairpin stem in the energetically close second structure.

Table 3. HIV-1 subtype specificity of the genomic region encompassing DIS, SD and Ψ hairpin

Subtype	Number of isolates[a]	DIS	Linker DIS/SD	SD	Linker SD/Ψ	Ψ
C	360 (359)	DIS_C – 78% DIS_C variants – 17% DIS_F – 2%	AGC – 56% AGG, AG – 22% GGG, GGGC – 9%	SD_{RS} – 35% SD^{5G} – 48% $SD_{C\text{-}G}$ – 9%	(AA)AUUUU(U)A[b] – 70% A(U)UUUUA – 7% AA – 3%	Ψ_C – 76% Ψ_C^{21G} – 4% Ψ_{RS} – 4% Ψ_C^{12A} – 2%
B	149 (127)	DIS_{Lai} – 70% DIS_{Lai}^{23A} – 15% DIS_{Lai}^{21G} – 2% DIS_{Trans} – 2%	GGGC – 85%	SD_{RS} – 79% SD^{7C}, SD^{AC} – 8% SD^{5G} – 4% $SD_{C\text{-}G}$ – 3%	(AA)AAA – 83% AAAAAA – 4% AAAG – 3%	Ψ_{RS} – 58% $\Psi_{U\text{-}A}$ – 26% $\Psi_{UU\text{-}AG}$ – 3%
E (A/E)	25	DIS_{Mal} – 80% DIS_{Mal} variants – 20%	AGC – 84%	SD_{RS} – 96%	AAA – 92%	Ψ_{RS} – 84%
A	21 (20)	DIS_{Mal} – 80% DIS_{Mal} variants – 20%	AGC – 81%	SD_{RS} – 86%	(A)A – 38% UAAAA – 19% AAAAAA – 9% UAAGGA – 9%	$\Psi_{UU\text{-}AG}$ – 57% $\Psi_{U\text{-}A}$ – 24% $\Psi_{U\text{-}A}^{12A}$ – 9%
D	13 (9)	DIS_{Lai} – 33% DIS_{Lai}^{23A} – 33% Other DIS_{Lai} variants – 34%	GGGC – 56% (G)GG – 44%	SD variants with A^1-U^{19} base pair substitution – 100%	(AA)AAA – 83%	$\Psi_{U\text{-}A}$ variants – 73% Ψ_{RS} – 23%
G	11 (10)	DIS_{Mal} – 50% DIS_{Mal} variants – 50%	AGC – 100%	SD_{RS} – 91%	(A)AAA – 55% AUAAA – 18%	$\Psi_{U\text{-}A}$, $\Psi_{U\text{-}A}$ variants – 73% Ψ_{RS} – 18%
F	7	DIS_F – 86% DIS_F variants – 14%	AGC – 57% A(G)G – 29%	SD_{RS} – 71% SD^{5G} – 14% $SD_{C\text{-}G}$ – 14%	(AA)AA – 86%	Ψ_{RS}, Ψ^{12A} – 43% $\Psi_{U\text{-}A}$ – 29% Ψ_C^{12A} – 14%
B/F	43 (38)	DIS_{Lai} variants – 33% DIS_{Lai} – 26% DIS_{Trans} – 15% DIS_F, DIS_F variants – 8%	GGGC – 63% AGC – 15% GGG – 7%	SD_{RS} – 68% $SD_{C\text{-}G}$ – 7% SD^{5G} – 5% SD^{7C}, – 5%	(AAA)AA –73%	$\Psi_{U\text{-}A}$ – 40% Ψ_{RS} – 19% Ψ^{12A} – 16% $\Psi_{UU\text{-}AG}$ – 12%
B/G	17 (16)	DIS_{Mal} – 53% DIS_{Mal}^{Ains} – 35% DIS_{Lai}^{13C} – 6%	AGC – 76% GGGC – 6%	SD_{RS} – 71% No – 18%	AAA – 50% AUAAA – 25% AAAA 6%	$\Psi_{CU\text{-}AG}$ – 50% Ψ_{RS} – 50%
A/G	11	DIS_{Mal} – 73% DIS_{Mal} variants – 18%	AGC – 100%	SD_{RS} – 100%	(A)AA – 91%	$\Psi_{U\text{-}A}$ – 45% Ψ_{RS} – 36%

The continuation of Table 3.

Subtype	Number of isolates[a]	DIS	Linker DIS/SD	SD	Linker SD/Ψ	Ψ
A/G/ J/K	11 (10)	DIS_{Mal} – 60% DIS_F – 30% DIS_{Mal} variants – 10%	AGC – 90%	SD_{RS} – 91%	(AA)A – 36% (A)AAAAA – 27% AUA – 18%	$\Psi_{U\text{-}A}$, $\Psi_{UU\text{-}AG}$ – 36% $\Psi_{U\text{-}A}{}^{12A}$ – 18%, $\Psi_{UU\text{-}AG}{}^{12A}$ – 9% $\Psi_C{}^{12A}$ – 12%
A/C	8	DIS_C – 38% DIS_C variants – 25% $DIS_{Mal}{}^{13C}$ – 12% $DIS_F{}^{23A}$ – 12%	AGC – 75%	SD_{RS} – 50% SD^{5G} – 38%	(AA)AUUUUA[b] – 50% A(A) – 25%	Ψ_C – 50% $\Psi_{UU\text{-}AG}$ – 12% $\Psi_C{}^{12A}$ – 12%
B/C	6	DIS_C – 67% $DIS_{Lai}{}^{23A}$ – 33%	AGC – 83%	SD_{RS} – 83%	(A)AUUUAUA[b] – 33% AAUUAUA[b] – 33% (A)AAAAA – 33%	Ψ_C – 67% $\Psi_{U\text{-}A}$ – 17%
C/D	6	$DIS_{Lai}{}^{23A}$ –50% DIS_{Trans} – 33% $DIS_C{}^{27G}$ – 17%	GGGC – 50% AGC – 17% GGG – 17%	$SD^{AC,\ A\text{-}U\ (C\text{-}G)}$ – 83% SD^{5G} – 17%	(AAA)AAA – 83%	$\Psi_{U\text{-}A}$, $\Psi_{UU\text{-}AG}$ – 50% Ψ^C – 17% Ψ^{20A} – 17%

[a] The number of isolates with defectless hairpin and linkers is given in brackets (see ANALYSIS PROCEDURE).
[b] The UUUGA tract which is a part of the SD/Ψ linker in the first structure is omitted.

for different subtypes and intersubtype recombinants are presented in Table 3. In our database only the groups of isolates of C and B subtypes and B/F IRs are rather representative, nevertheless we included small/scanty groups of isolates of other subtypes and IRs in this table to illustrate common trends in subtype specificity of the region under study.

For example, the region encompassing DIS and SD hairpins is very similar for HIV-1 isolates of subtypes E (A/E), A and G – they have invariably DIS_{Mal} or DIS_{Mal} variants and very frequently AGC linker between DIS and SD and SD_{RS} (Table 1). On the other hand, these isolates differ in Ψ hairpin structure: 84% of subtype E (A/E) isolates have Ψ_{RS}, while 90% of subtype A and 73% of subtype G isolates have elongated Ψ hairpin. As for intersubtype recombinants, most of B/F IRs have DIS_{Lai} variants, DIS_{Lai} or DIS_{Trans} and GGGC linker between DIS and SD, while only about 10% of these IRs have DIS_F or DIS_F variants and AGC linker DIS/SD, which may evidence for the prevalence of B subtype specificity in the region encompassing DIS and SD hairpins in B/F IRs. Whereas most of B/G IRs have DIS_{Mal} or DIS_{Mal} variants and AGC linker DIS/SD, i.e. this region is subtype G specific in B/G IRs. In particular, prevalence of B, F or G subtypes in the region under study may be caused by different affinity of NC protein from these subtypes isolates for DIS_{Lai}, DIS_F or DIS_{Mal} and Ψ_{RS} or $\Psi_{U\text{-}A}$ variants.

Conservation of DIS, SD and Ψ hairpins. Our phylogenetic findings proved that the primary sequence of the region encompassing DIS, SD and Ψ hairpins in HIV-1 isolates is highly heterogeneous. This sequence is not recurrent in 62% genomes from our database. The base changes in DIS, SD and Ψ variants described in previous sections mainly lead to moderate alterations of hairpins' structures, and these base changes are believed to affect moderately the functioning of these hairpins if any.

Most base changes occurring below the level of errors seem also to influence moderately the hairpins' efficiency. Among them there are the base changes leading to co-existence between the common structure (containing DIS, SD and Ψ hairpins) of the region preceding GAG start codon and alternative structure(s) lacking some of the hairpins. Actually, a co-existence of common and alternative structures can be considered as a decrease in concentrations of the functional hairpins. The pattern of co-existence between common and alternative structures is given in Fig. 5, in this case the presence of alternative (dominant) structure is due to G residue adjacent to the 5′-end

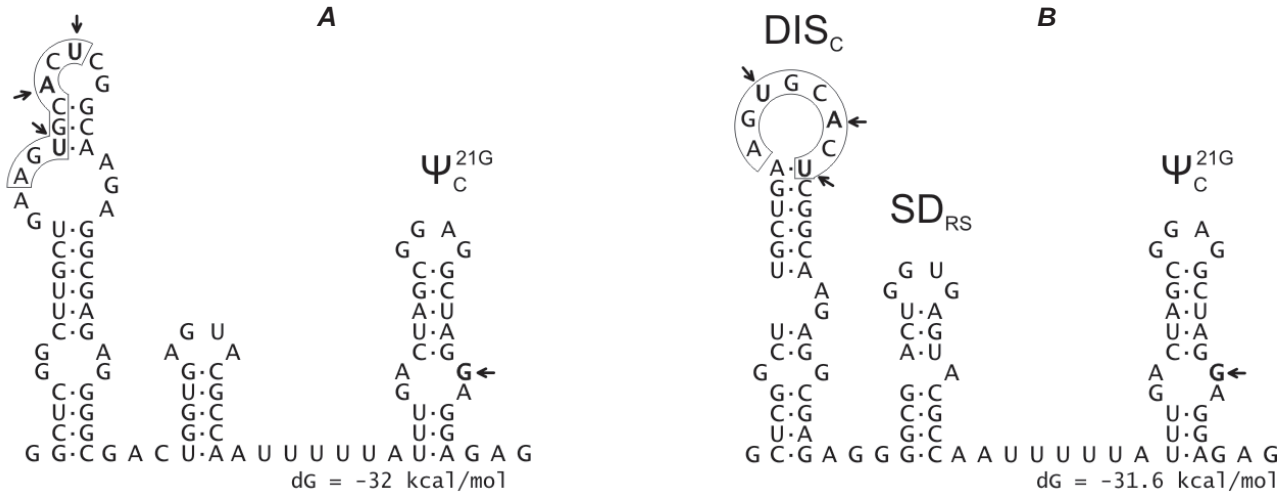

Fig. 5. The secondary structures of the region encompassing DIS, SD and Ψ hairpins in genomic RNA of isolate 02ZAPS015MB1 (accession number – DQ369995). The palindrome in the upper part of DIS hairpin (or in the sequence corresponding to it) is boxed. The nucleotides that differ from those in $DIS_{Lai(RS)}$, SD_{RS} and Ψ_{RS} (or in the sequences corresponding to these hairpins) are indicated by arrows.

of DIS hairpin (A residue commonly flanks DIS at this end).

About 4% of HIV-1 isolates from our database have the critical base changes which lead to significant prevalence (with difference in free energy above 2 kcal/mole) of alternative structures of the region under study or greatly disturb the structure of DIS, SD or Ψ hairpins. Almost all critical base changes were found only once among the isolates studied, and most likely they were caused by sequencing or database errors.

So, earlier phylogenetic studies on small sets (less than 50) of HIV-1 isolates [55, 56], recent works on large pools of several hundreds of isolates [7, 8, 51] and also the findings of the present paper show that notwithstanding a great diversity of HIV-1 genome region under study, DIS, SD and Ψ hairpins (or DIS-, SD- and Ψ-like structures) are formed in most HIV-1 isolates, i.e. these elements are essential for HIV-1 replication cycle.

Concluding remarks

As compared to the set of 350 HIV-1 isolates previously reported [7] the investigation of 757 HIV1 genomes performed in this work has not revealed new variants of DIS, SD and Ψ hairpins (except for Ψ_C^{27G}). This can imply that most variants with the tolerable mutations have been already registered. The functional properties of newly revealed variants are the challenge for experimentalists.

Both the findings on DIS, SD and Ψ hairpins diversity reported in this paper and the CESSHIV-1 database which will be published soon (and include also other control elements) could be considered as a guide to variants of HIV-1 control elements, their secondary structures, subtype specificity and geographical distribution. Also the knowledge of all tolerable mutations in control elements is important for clarification of mechanism of viral replication. For example, our findings on A insertion intolerance in DIS_{Lai}, but not in DIS_{Mal}, are perspective in NMR and X-ray studies to model structures of kissing-loop complexes.

МНОГООБРАЗИЕ ШПИЛЕК DIS, SD И Ψ В ИЗОЛЯТАХ ВИЧ-1 ГРУППЫ М: ИССЛЕДОВАНИЕ *IN SILICO*

М. И. Зарудная, А. Л. Потягайло, И. Н. Коломиец, Д. Н. Говорун

Институт молекулярной биологии и генетики НАН Украины, Киев;
e-mail: m.i.zarudna@imbg.org.ua

Для 731 ВИЧ-1 изолята группы М из базы данных NCBI проанализированы первичная последовательность и вторичная структура области генома, охватывающей шпильки DIS, SD и Ψ. Предсказание вторичной структуры проводилось с помощью программы mfold М. Зукера (M. Zuker). Было установлено, что несмотря на высокую гетерогенность первичной последовательности изучаемой области, она сворачивается в шпильки DIS, SD и Ψ (DIS-, SD- и Ψ-подобные шпильки) в 96% исследованных геномах. Филогенетический анализ показал, что в наиболее часто встречающихся вариантах шпильки DIS (DIS_{Lai}, DIS_{Mal} и DIS_C) допускаются замены оснований лишь в определенных положениях. В частности, за-

мена основания в положении 23 стебля происходит в 5 и 33 раза чаще в DIS_{Lai}, чем в DIS_{Mal} и DIS_C соответственно, в то время как вставка A на 5′-конце апикальной петли допускается в DIS_{Mal} и DIS_C, но не в DIS_{Lai}. Нами обнаружено, что замена нижней пары G-C на A-U в шпильке SD является высокоспецифичной для изолятов субтипа D. В статье описаны и систематизированы все варианты шпилек DIS, SD и Ψ и предложены схемы переходов между вариантами этих шпилек через единичные замены оснований. Обнаружено, что большинство вариантов шпилек DIS и Ψ существуют в нескольких конформациях.

Ключевые слова: ВИЧ-1, филогенетический анализ, свертывание РНК, шпилька DIS, шпилька SD, шпилька Ψ, димеризация генома, упаковка ВИЧ-1.

РІЗНОМАНІТНІСТЬ ШПИЛЬОК DIS, SD ТА Ψ У ІЗОЛЯТАХ ВІЛ-1 ГРУПИ М: ДОСЛІДЖЕННЯ *IN SILICO*

М. І. Зарудна, А. Л. Потягайло, І. М. Коломієць, Д. М. Говорун

Інститут молекулярної біології і генетики НАН України, Київ;
e-mail: m.i.zarudna@imbg.org.ua

Для 731 ізоляту ВІЛ-1 групи М з бази даних NCBI проаналізовано первинну послідовність та вторинну структуру ділянки геному, що охоплює шпильки DIS, SD та Ψ. Вторинні структури передбачували за допомогою програми mfold М. Зукера (M. Zuker). Показано, що попри високу гетерогенність первинної послідовності досліджуваної ділянки вона згортається у шпильки DIS, SD та Ψ (DIS-, SD- та Ψ-подібні шпильки) у 96% геномів. Філогенетичний аналіз свідчить, що у варіантах шпильки DIS, які зустрічаються найчастіше (DIS_{Lai}, DIS_{Mal} та DIS_C), заміни нуклеотидних основ допустимі лише в певних положеннях. Так, основи в положенні 23 стебла замінюються у 5 та 33 рази частіше у DIS_{Lai}, ніж у DIS_{Mal} та DIS_C відповідно, у той час як вставка A на 5′-кінці апікальної петлі можлива лише в DIS_{Mal} та DIS_C, але не у DIS_{Lai}. Виявлено, що заміна нижньої пари G-C на A-U у шпильці SD є високоспецифічною для ізолятів субтипу D. Описано та систематизовано усі варіанти шпильок DIS, SD и Ψ і запропоновано схеми переходів між варіантами цих шпильок через поодинокі заміни основ. Встановлено, що шпильки DIS та Ψ існують здебільшого в кількох конформаціях.

Ключові слова: ВІЛ-1, філогенетичний аналіз, згортання РНК, шпилька DIS, шпилька SD, шпилька Ψ, димеризація геному, пакування ВІЛ-1.

1. *Russell R. S., Liang C., Wainberg M. A.* // Retrovirology. – 2004. – **1**. – P. 23.
2. *Paillart J.-C., Shehu-Xhilaga M., Marquet R., Mak J.* // Nat. Rev. Microbiol. – 2004. – **2**, N 6. – P. 461–472.
3. *Greatorex J., Gallego J., Varani G., Lever A.* // J. Mol. Biol. – 2002. – **322**, N 3. – P. 543–557.
4. *Abbink T. E. M., Berkhout B.* // J. Biol. Chem. – 2003. – **278**, N 13. – P. 11601–11611.
5. *Damgaard C. K., Andersen E. S., Knudsen B. et al.* // J. Mol. Biol. – 2004. – **336**, N 2. – P. 369–379.
6. *Kasprzak W., Bindewald E., Shapiro B. A.* // Nucl. Acids Res. – 2005. – **33**, N 22. – P. 7151–7163.
7. *Zarudnaya M. I., Kolomiets I. M., Potyahaylo A. L., Hovorun D. M.* // Trends in RNA Research / Ed. P. A. McNamara. – New York: Nova Science Publishers, 2006. – P. 159–189.
8. *Yuan Y., Kerwood D. J., Paoletti A. C. et al.* // Biochemistry. – 2003. – **42**, N 18. – P. 5259–5269.
9. *Zuker M.* // Nucleic Acids Res. – 2003. – **31**, N 13. – P. 3406–3415.
10. *Ashe M. P., Pearson L. H., Proudfoot N. J.* // EMBO J. – 1997. – **16**, N 18. – P. 5752–5763.
11. *Anderson J. P., Rodrigo A. G., Learn G. H. et al.* // J. Virol. – 2000. – **74**, N 22. – P. 10752–10765.
12. *Amarasinghe G. K., De Guzman R. N., Turner R. B., Summers M. F.* // J. Mol. Biol. – 2000. – **299**, N 1. – P. 145–156.
13. *Laughrea M., Shen N., Jetté L., Wainberg M. A.* // Biochemistry. – 1999. – **38**, N 1. – P. 226–234.
14. *Skripkin E., Paillart J.-C., Marquet R. et al.* // Proc. Natl. Acad. Sci. USA. – 1994. – **91**, N 11. – P. 4945–4949.
15. *Laughrea M., Jetté L.* // Biochemistry. – 1994. – **33**, N 45. – P. 13464–13474.
16. *Muriaux D., De Rocquigny H., Roques B.-P., Paoletti J.* // J. Biol. Chem. – 1996. – **271**, N 52. – P. 33686–33692.
17. *Laughrea M. and Jetté L.* // Biochemistry. – 1996. – **35**, N 5. – P. 1589–1598.
18. *Muriaux D., Fossé P., Paoletti J.* // Ibid. – N 15. – P. 5075–5082.
19. *Ulyanov N. B., Mujeeb A., Du Z. et al.* // J. Biol. Chem. – 2006. – **281**, N 23. – P. 16168–16177.

20. *Marquet R., Paillart J.-C., Skripkin E. et al.* // Nucleic Acids Res. – 1994. – **22**, N 2. – P. 145–151.
21. *Muriaux D., Girard P.-M., Bonnet-Mathonière B., Paoletti J.* // J. Biol. Chem. – 1995. – **270**, N 14. – P. 8209–8216.
22. *Laughrea M., Jette L.* // Biochemistry. – 1997. – **36**, N 31. – P. 9501–9508.
23. *Mujeeb A., Clever J. L., Billeci T. M. et al.* // Nat. Struct. Biol. – 1998. – **5**, N 6. – P. 432–436.
24. *Jossinet F., Paillart J.-C., Westhof E. et al.* // RNA – 1999. – **5**, N 9. – P. 1222–1234.
25. *Takahashi K.-I., Baba S., Chattopadhyay P. et al.* // Ibid. – 2000. – **6**, N 1. – P. 96–102.
26. *Ennifar E., Walter P., Ehresmann B. et al.* // Nat. Struct. Biol. – 2001. – **8**, N 12. – P. 1064–1068.
27. *Pattabiraman N., Martinez H. M., Shapiro B. A.* // J. Biomol. Struct. Dyn. – 2002. – **20**, N 3. – P. 397–411.
28. *Réblová K., Špačková N., Šponer J. E. et al.* // Nucleic Acids Res. – 2003. - **31**, № 23. – P. 6942–6952.
29. *Windbichler N., Werner M., Schroeder R.* // Ibid. – N 22. – P. 6419–6427.
30. *Aci S., Ramstein J., Genest D.* // J. Biomol. Struct. Dyn. – 2004. – **21**, N 6. – P. 833–840.
31. *Mihailescu M.-R., Marino J. P.* // Proc. Natl. Acad. Sci. USA. – 2004. – **101**, N 5. – P. 1189–1194.
32. *Kieken F., Paquet F., Brulé F. et al.* // Nucleic Acids Res. – 2006. – **34**, N 1. – P. 343–352.
33. *Ennifar E., Dumas P.* // J. Mol. Biol. – 2006. – **356**, N 3. – P. 771–782.
34. *Esparza J., Bhamarapravati N.* // Lancet. – 2000. – **355**, N 9220. – P. 2061–2066.
35. *Andersen E. S., Jeeninga R. E., Damgaard C. K. et al.* // J. Virol. – 2003. – **77**, N 5. – P. 3020–3030.
36. *Lodmell J. S., Ehresmann C., Ehresmann B., Marquet R.* // J. Mol. Biol. – 2001. – **311**, N 3. – P. 475–490.
37. *Kieken F., Arnoult E., Barbault F. et al.* // Eur. Bioph. J. – 2002. – **31**, N 7. – P. 521–531.
38. *Chen X., McDowell J. A., Kierzek R. et al.* // Biochemistry. – 2000. – **39**, N 30. – P. 8970–8982.
39. *Huthoff H., Das A. T., Vink M. et al.* // J. Virol. – 2004. – **78**, N 9. – P. 4907–4913.
40. *Kim C.-H., Tinoco I.,Jr.* // Proc. Natl. Acad. Sci. USA. – 2000. – **97**, N 17. – P. 9396–9401.
41. *Shubsda M. F., Paoletti A. C., Hudson B. S., Borer P. N.* // Biochemistry. – 2002. – **41**, N 16. – P. 5276–5282.
42. *Chin M. P. S., Rhodes T. D., Chen J.* // Proc. Natl. Acad. Sci. USA. – 2005. – **102**, N 25. – P. 9002–9007.
43. *Baba S., Takahashi K.-i., Noguchi S. et al.* // J. Biochem. – 2005. – **138**, N 5. – P. 583–592.
44. *Clever J. L., Miranda D. Jr., Parslow T. G.* // J. Virol. – 2002. – **76**, N 23. – P. 12381–12387.
45. *Amarasinghe G. K., De Guzman R. N., Turner R. B. et al.* // J. Mol. Biol. – 2000. – **301**, N 2. – P. 491–511.
46. *De Guzman R. N., Wu Z. R., Stalling C. C. et al.* // Science. – 1998. – **279**, N 5349. – P. 384–388.
47. *Khandogin J., Musier-Forsyth K., York D. M.* // J. Mol. Biol. – 2003. – **330**, N 5. – P. 993–1004.
48. *Henriet S., Richer D., Bernacchi S. et al.* // Ibid. – 2005. – **354**, N 1. – P. 55–72.
49. *Paillart J.-C., Dettenhofer M., Yu X.-f. et al.* // J. Biol. Chem. – 2004. – **279**, N 46. – P. 48397–48403.
50. *Mathews D. H., Disney M. D., Childs J. L. et al.* // Proc. Nat. Acad. Sci. USA. – 2004. – **101**, N 19. – P. 7287–7292.
51. *Paoletti A. C., Shubsda M. F., Hudson B. S., Borer P. N.* // Biochemistry. – 2002. – **41**, N 51. – P. 15423–15428.
52. *Damgaard C. K., Dyhr-Mikkelsen H., Kjems J.* // Nucleic Acids Res. – 1998. – **26**, N 16. – P. 3667–3676.
53. *Зарудная М. И., Потягайло А. Л., Говорун Д. Н.* // Биополимеры и клетка. – 2003. – **19**, № 1. – P. 37–42 (in Russian).
54. *Zarudnaya M. I., Kolomiets I. M., Potyahaylo A. L., Hovorun D. M.* // Укр. біохім. журн. – 2005. – **77**, № 2. – P. 5–15. (in Engl.)
55. *Berkhout B., van Wamel J. L. B.* // J. Virol. – 1996. – **70**, N 10. – P. 6723–6732.
56. *Laughrea M., Jette L., Mak J. et al.* // Ibid. – 1997. – **71**, N 5. – P. 3397–3406.

Reseived 01.11.2006