Structural insight into HIV-1 reverse transcription initiation in MAL-like templates (CRF01_AE, subtype G and CRF02_AG)

Irina N. Kolomiets*, Margarita I. Zarudnaya*, Andriy L. Potyahaylo* and Dmytro M. Hovorun*a,b,c*

aDepartment of Molecular and Quantum Biophysics, Institute of Molecular Biology and Genetics, National Academy of Sciences of Ukraine, 150 Zabolotnoho Str., 03680 Kyiv, Ukraine; bState Key Laboratory of Molecular and Cell Biology, Research and Educational Center, 150 Zabolotnoho Str., 03680 Kyiv, Ukraine; cDepartment of Molecular Biotechnology and Bioinformatics, Institute of High Technologies, Taras Shevchenko National University of Kyiv, 2-h Hlushkova Ave., 03022 Kyiv, Ukraine

Communicated by Ramaswamy H. Sarma

(Received 25 July 2013; accepted 15 January 2014)

Based on the known structural model for reverse transcription initiation complex of the human immunodeficiency virus type 1 (HIV-1) MAL isolate, we attempted to predict a structural behavior of MAL-like templates (CRF01_AE, subtype G and CRF02_AG) within the initiation complex by in silico experiments. Switches from the D-duplex (dimerization-competent) conformation to the I-duplex (initiation-competent) conformation and then to conformations with an open primer activation signal (PAS) structure have been examined for four fragments of U5 and primer binding site (PBS) region, the minimal fragment (nt 121–243), fragment 1 (nt 110–243), fragment 2 (nt 113–259), and extended fragment 2 (nt 109–261). Switches from the D-duplex conformation to the I-duplex conformation in the minimal fragment or fragment 1 and from the I-duplex conformation to conformations with exposed PAS motif in fragment 1 are similar in all MAL-like templates. A PAS exposure in fragment 2 and extended fragment 2 is supported by PBS stem extension which structure is affected by subtype-specific variations in CRF01 AE (the mutated motif 116GUUAG120) and CRF02 AG (7-nt deletion downstream of the PBS motif and G/C/A insertion at the 3’ end of fragment 2). These switchable conformations contain the established structural elements essential for HIV-1 reverse transcription initiation as well as several elements that may also be relevant to initiation process, namely hairpins with GAAA apical loops, and self-contained motifs of the duplicate insertion and the downstream palindromic sequence. Taken together, our findings suggest a role for the duplicate insertion of MAL-like templates in HIV-1 reverse transcription initiation process and possible mechanisms to realize this role.

Keywords: secondary structure prediction; MAL-like templates; HIV-1 reverse transcription initiation; PBS (primer binding site); PAS (primer activation signal)

1. Introduction

Reverse transcription is a key step in the retroviral replication, at which the single-stranded RNA genome is reverse transcribed into a double-stranded DNA copy with duplicated long terminal repeats. In the human immunodeficiency virus type 1 (HIV-1), initiation of reverse transcription is a complex and highly regulated process that involves the viral RNA (vRNA), the cellular tRNAlys3 as a primer, and the viral enzyme reverse transcriptase (RT) assembling into a productive ternary ribonucleoprotein complex, for reviews, see (Abbink & Berkhout, 2008; Hu & Hughes, 2012; Isel, Ehresmann, & Marquet, 2010). Upon primer annealing, the 18 3’-terminal nucleotides of tRNAlys3 base pair with the highly conserved complementary motif called the primer binding site (PBS) located in the 5’-untranslated region (UTR) of the genomic RNA. This strong tRNA–PBS interaction (the PBS-duplex) is considered as a major determinant for the selection of tRNA primer and an absolute requirement for HIV-1 reverse transcription initiation.

A large volume of early and recent studies has provided clear evidence of additional contacts between the vRNA and tRNAlys3 outside of the PBS, for recent review, see (Sleiman et al., 2012). These studies were conducted on two diverse HIV-1 strains, the MAL isolate (an A/D/K/U recombinant with a PBS domain of subtype A origin) and the Lai/HXB2/NL4.3 isolates (subtype B). The additional contacts involve the A-rich loop upstream of the PBS in vRNA and the U-rich anticodon loop in tRNAlys3 (the A-loop–anticodon interaction), the primer activation signal (PAS) motif of vRNA and the T-arm of the primer (the PAS–anti-PAS interaction), the C-rich motif of vRNA with the 3’ anticodon stem and variable loop of the primer (illustrated in Figure 1(A)). All additional contacts are significantly shorter than the PBS-duplex of 18 base pairs (bps) and may be considered as potential secondary determinants for the selection
of tRNA primer and modulators of HIV-1 reverse transcription initiation.

The additional interactions between vRNA and tRNAlys3 appeared to depend on HIV-1 template used, i.e., the sequence and structure of the regions surrounding the PBS. While the importance of the A-loop–anticodon interaction was clearly demonstrated for reverse transcription initiation in the MAL isolate, there are discrepant data on such interaction in the HXB2/NL4.3 isolates. In the MAL isolate, the A-loop–anticodon interaction is extended through the primer anticodon stem (Bibille et al., 2009; Goldschmidt et al., 2002; Isel, Ehresmann, Keith, Ehresmann, & Marquet, 1995; Isel, Keith, Ehresmann, & Marquet, 1998; Isel et al., 1999). Alternatively, the primer anticodon stem and variable loop are involved in contiguous interaction with the C-rich motif of vRNA as demonstrated for the NL4.3 isolate (Iwatani, Rosen, Guo, Musier-Forsyth, & Levin, 2003; Wilkinson et al., 2008).

For HIV-1 HXB2 initiation complex, the PAS–anti-PAS interaction was proposed to regulate the reverse transcription initiation by occlusion and exposure of the 

Figure 1. Secondary structure of the tRNAlys3 primer with sequence motifs complementary to vRNA (A). Mfold-calculated structures (ΔG in kcal/mol) of U5-PBS fragment nt 121–243 of the MAL isolate (GenBank accession number X04415) – the most favorable conformation with I-duplex (B) and the most favorable conformation with D-duplex (D), the most favorable conformation of U5-PBS fragment nt 123–226 of the NL4.3 isolate (AF070521) (C). PBS motif, PAS motif, and duplicate insertion are marked by green, orange, and blue, respectively. The sites proposed for NC binding in the Lai strain (Damgaard et al., 1998), NC-binding and NC-mediated duplex destabilization in the NL4.3 strain (Wilkinson et al., 2008) and Vif binding in the HXB2 strain (Henriet et al., 2005) are indicated by solid, chain, dashed, and dotted lines, respectively (C).
PAS motif (Abbink, Beerens, & Berkhout, 2004; Beerens & Berkhout, 2002; Beerens, Groot, & Berkhout, 2001; Huthoff, Bugala, Barciszewski, & Berkhout, 2003; Ooms, Cupac, Abbink, Huthoff, & Berkhout, 2007). The initiation efficiency increased for different templates with the ΔΔG_PAS value (ΔΔG value of PAS-masked versus PAS-exposed structures) of about –3 kcal/mol or higher (Ooms et al., 2007). Although a potential to form the PAS–anti-PAS interaction is highly conserved among different HIV-1 strains as well as HIV-2 and SIV strains (Beerens et al., 2001), the rather conflicting data are reported on the involvement of the PAS motif at the initiation stage of reverse transcription. Quite recently, the PAS–anti-PAS interaction in 122-nt fragment of the HXB2 isolate was demonstrated by FRET spectroscopy (Beerens et al., 2013) and in small fragments of the MAL and HXB2 isolates mimicking the natural PAS pairing by NMR spectroscopy (Sleiman, Barraud, Brachet, & Tisne, 2013).

The MAL isolate differs from the Lai/HXB2/NL4.3 isolates by several point mutations upstream of the PBS and a 24-nt duplicate insertion that repeats a portion of the PBS and downstream sequence (Baudin et al., 1993; Goldschmidt et al., 2004). The duplicate insertion is a common sequence signature of the PBS region in all isolates of subtype G, subgroup A2 of subtype A, CRFs (Andersen, Jeeninga, Damgaard, Berkhout, & Kjems, 2003; Chang, Apichartpiyakul, Kuiken, Essex, & Lee, 1999; De Baar et al., 2000; Goldschmidt et al., 2004; Kurosu et al., 2001; Sakuragi, Sakuragi, Ohishi, & Shioda, 2010). Hereinafter, MAL-like templates denote HIV-1 isolates with the duplicate insertion clustering with the MAL isolate in the PBS region (Figure 2).

According to the secondary structure model of the HIV-1 MAL initiation complex (MAL RNA fragments nt 1–311/123–217), RT ultimately recognizes the PBS-duplex (helix 7F) and two intramolecular duplexes, the bottom duplex of the U5-top hairpin (helix 2) and the duplex between an upstream portion of the duplicate insertion, and a PAS motif (helix 1) that locks the model (Goldschmidt et al., 2002, 2004; Isel et al., 1995, 1998, 1999). The sharp kink between the PBS-duplex and helix 2, which is crucial for productive positioning of RT, is imposed by two short junctions: between helices 1 and 2 as well as between helices 8 and 1. These helices and also the duplicate insertion, PBS, and PAS motifs are illustrated in Figure 1(B). In the MAL RNA, extensive interactions with tRNASys3 are required to generate the structural elements recognized by RT, while in the HXB2/NL4.3 RNA, helix 1 and helix 2 are already present prior to tRNASys3 annealing (Goldschmidt et al., 2004). The secondary structure models of MAL and HXB2/NL4.3 initiation complexes imply the same helix 2 for both strains, while the 3'-strand of helix 1 is an upstream portion of the duplicate insertion in the MAL (Goldschmidt et al., 2002, 2004; Isel et al., 1995, 1998, 1999; Uppuludinne, Sonavane, & Joshi, 2013) and the anti-PAS sequence A/GCCAGAG in the HXB2/NL4.3 isolates (Abbink et al., 2004; Beerens et al., 2001, 2013; Goldschmidt et al., 2004; Iwatani et al., 2003; Liu, Harada, Miller, Le Grice, & Zhuang, 2010).

Pfold predictions of the 5' region of 20 divergent HIV-1 sequences demonstrated conservation of helix 1 of HXB2 type (PBS2) and helix 2 (PBS3) with the reliability >90% for all sequences, including MAL and MAL-like isolates (Damgaard, Andersen, Knudsen, Gorodkin, & Kjems, 2004). To distinguish between different names of helices being discussed in the text, the helix locking the MAL initiation complex (helix 1) is hereafter referred to as I-duplex (duplex of initiation-competent structure of the MAL isolate as shown in (Isel et al., 1995)), the helix between the PAS motif and vRNA anti-PAS sequence (A/GCCAGAG in the HXB2/NL4.3 isolates or CCAGAG in MAL-like isolates) as to D-duplex (duplex of dimerization-competent structure as predicted by Pfold in (Damgaard et al., 2004)), and the bottom duplex of U5-top hairpin is denoted as U-duplex (Figure 1(B)–(D)).

The efficiency of reverse transcription is regulated by a number of viral and host proteins, including the nucleocapsid protein (NCp), viral infectivity factor (Vif), Nef, Tat, Vpr, and others, for recent reviews, see (Levin, Mitra, Mascarenhas, & Musier-Forsyth, 2010; Mirambeau, Lyonais, & Gorelick, 2010; Mougel, Houzet, & Darlix, 2009; Sleiman et al., 2012; Warren, Warrilow, Meredith, & Harrich, 2009). In particular, a duplex destabilizing activity of NC was detected at four sites of the PBS subdomain in the NL4.3 isolate (Wilkinson et al., 2008). NC-binding sites were observed in and around the PBS of the Lai/NL4.3 isolates (Damgaard, Dyrh-Mikkelsen, & Kjems, 1998; Wilkinson et al., 2008), and removal of the upper PBS loop structure eliminated ~4 high affinity NC-binding sites but did not impair packaging of the NL4.3 genomic RNA, suggesting their roles in tRNAlys3 binding and/or reverse transcription (Heng et al., 2012). In vitro data evidenced that Vif can partially replace NC in this function(s); the Vif-induced protections against hydrolysis by RNase T1 were also registered in and around the PBS of the HXB2 genome (Henriet et al., 2005, 2007). These overlapping sites for NC and Vif binding are illustrated in Figure 1(C). Among host factors, cellular RNA helicase A (RHA) was reported to bind primarily to the HIV-1 5' UTR, associate with HIV-1 Gag in a RNA-dependent manner and proposed to regulate the interactions of vRNA with tRNAlys3, NC, RT, and other factors such as to form a functional reverse transcription complex (Roy et al., 2006; Xing, Liang, & Kleiman, 2011; Xing, Niu, & Kleiman, 2012).
As compared to the Lai/HXB2/NL4.3 isolates, MAL-like isolates with the duplicate insertion evidently possess a greater number (roughly doubled) of potential sites for binding to NC and/or Vif. This suggests that regulation of the tRNA\text{lys}3 binding and reverse transcription initiation in MAL-like isolates may be different or more complex than that in Lai/HXB2/NL4.3 isolates.

In reviewing the literature on structure of HIV-1 reverse transcription initiation complex, there are some issues that could arouse researchers’ interest. Inasmuch as the duplicate insertion is a common sequence feature for the MAL and MAL-like isolates, whether the structure of MAL-like templates within the initiation complex (the I-duplex conformation) is quite similar to that of the MAL isolate. Whether the entire duplicate sequence and the downstream anti-PAS sequence CCAGAG in MAL-like isolates play a role in the initiation complex structure or are dispensable for this structure. Whether the duplicate sequence in dimerization-competent structure (the D-duplex conformation) is structured or commonly unpaired. Which structural rearrangements and energy input are required to switch from the D-duplex conformation to the I-duplex conformation. Whether the PBS region conformation with an open PAS structure in MAL-like isolates (the PAS conformation) could be adopted with the same difference in energy as it was demonstrated for the HXB2 isolate (Ooms et al., 2007).

To address these issues, the aim of the present work was to computationally explore the propensity of U5-PBS region fragments of MAL-like templates to adopt D-duplex, I-duplex, and an open PAS conformations that are supposed to be relevant for reverse transcription initiation process, the D-duplex conformation is considered as a structural pre-requisite for the I-duplex conformation.

2. Materials and methods

Predictions of RNA secondary structure were carried out using the Mfold web server (RNA Mfold online version 3.5 [http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form] (Zuker, 2003)). Mfold-calculated structures are generated on the principal of free energy minimization using a dynamic programming approach and empirically derived thermodynamic parameters and rules that take into account stability of different secondary structure motifs such as helices, hairpin loops, bulge loops, internal loops, and multibranch loops. Mfold algorithm does not include tertiary interactions or pseudoknots. The optional parameters of window size and percent suboptimality are used to control the number of foldings and the free energy increment for computing suboptimal foldings, respectively. In this study, the window size of 0 and suboptimality range of 10% have been chosen to obtain as many as possible various suboptimal foldings under the upper bound of computed foldings that is fixed at 50 in Mfold version 3.5. This version computes RNA foldings at the fixed temperature and ionic conditions (37°C, 1 M NaCl, no divalent ions).

To model structural determinants of the MAL initiation complex, we applied two folding constraints, the first to prohibit base pairs in 18 nts of the PBS motif (PBS-duplex requirement) and 3 nts immediately upstream of the PBS (requirement for free CAG junction between PBS- and U-duplexes) and the second to force 8 bps in U-duplex (U-duplex requirement) for all folding jobs. In most cases, an initial pool of computed foldings (as a rule, 25–40 structures under the upper bound of 50) contained all three conformations of interest, including conformation(s) of an open PAS structure. From this initial pool, we selected the D-duplex conformation of lowest free energy \(\Delta G\) (the optimal conformation), the I-duplex conformation of lowest \(\Delta G\), and the PAS conformation(s) of lowest \(\Delta G\). To be ascertained that the selected conformations are nonrandom and some conformations are not omitted, we then applied three special folding constraints, the first forcing base pairs in D-duplex to verify the D-duplex conformation, the second forcing base pairs in I-duplex to verify the I-duplex conformation, and the third prohibiting the PAS motif from base pairing to verify the PAS conformation. In most cases, an optimal conformation under each restriction was the same as we selected from an initial pool of computed foldings.

Upon selection of U5-PBS region fragments of MAL-like templates for secondary structure prediction, we based on the chemical (enzymatic) probing and mutational (deletion) data on the sequence requirements for effective tRNA\text{lys}3 annealing and/or reverse transcription initiation complex of the MAL isolate (Goldschmidt et al., 2002, 2004; Isel et al., 1995, 1998, 1999). We also took into consideration studies with additional template bases in the Lai/NL4.3 isolates (upstream/downstream of D-duplex or downstream of the PBS), suggesting their roles in reverse transcription initiation (Beerens et al., 2001; Beerens & Berkhout, 2002; Iwatani et al., 2003; Liang, Rong, Russell, & Wainberg, 2000; Ooms et al., 2007; Xing et al., 2011).

Because reverse transcription is initiates close to the 5′ end of the RNA dimeric genome, we explored different dimerization-competent models of HIV-1 leader RNA, reviewed in (Lu, Heng, & Summers, 2011) to define the 5′ and 3′ boundaries of U5-PBS region fragments for study. Based on general implications that the TAR and poly(A) hairpins can play roles in dimerization and genome packaging, the TAR hairpin structure may contribute to effective reverse transcription initiation (Lalonde et al., 2011), the DIS hairpin being a major determinant of dimerization can also play a role in maintaining proper nucleic acid structures in the reverse transcription complex (Chin et al., 2008), we limited U5-PBS region fragments by the 3′ end.
of the poly(A) hairpin and the 5' end of a common DIS hairpin (the region nt 102 to 260 in MAL coordinates). So, each fragment included the tract nt 123 to 217 ultimately required to form the MAL initiation complex, while its 5' portion varied from the 3' end of the poly(A) hairpin up to nt 123 and 3' portion varied from nt 217 up to the 5' end of a common DIS hairpin.

By the analysis of Mfold secondary structure predictions of these fragments of different lengths for 2 representatives of CRF01_AE, subtype G, and CRF02_AG, we found out that the minimal fragment nt 121–243 (for fragment boundaries, see Figure 2) adopts the D-duplex and I-duplex conformations with a minimal difference in free energy of about $-2 \text{kcal/mol}$. Fragment 1 (nt 110–243) adopts these two conformations with the same $\Delta G$ value and also conformations with an open PAS structure (the PAS1 conformations) without restriction on PAS motif base pairing. Fragment 2 (nt 113–259) and extended fragment 2 (nt 109–261) adopts the D-duplex and the I-duplex conformations with a lower $\Delta \Delta G$ of about $-3 \text{kcal/mol}$, while the conformations with an open PAS structure (the PAS2 conformations) appeared to be more stable than D-duplex and the I-duplex conformations in some MAL-like isolates. Therefore, these four fragments within the U5-PBS region nt 102–260 with the highest $\Delta \Delta G$ between conformations of interest have been chosen for secondary structure prediction of all MAL-like templates studied here.

To examine the conformational changes observed in short fragments for validity in a broader context, we fulfilled Mfold and UNAfold (Markham & Zuker, 2008) predictions with the same chosen parameters and constraints of the fragment nt 102–359 closed by the U5/AUG-duplex (see Section 4). We also constrained the bottom 4-bp duplex of the DIS hairpin structure to model dimeric state of vRNA. Such an approach for assessing a propensity of RNA fragments to adopt stable conformations in short fragments and then in a broader sequence context was previously used in our study of the HIV-1 3' UTR to model a complete poly(A) region structure by Mfold and UNAfold predictions (Zarudnaya, Potyahaylo, Kolomiets, & Hovorun, 2013).

The conformations of four fragments of U5-PBS region and U5/AUG domain have been predicted for all HIV-1 complete genomic sequences of CRF01_AE, CRF02_AG, and subtype G containing the entire U5-PBS region, which were available by the end of 2012 in the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/). For patients with multiple sequences, we selected only one isolate with most common PBS sequence within each intrapatient sequence set. Totally, we examined 52 genomic sequences of CRF01_AE, 18 of subtype G, and 21 of CRF02_AG (for accession numbers and sequence alignment, see Supplement 1, Figure S1).

3. Results and discussion

3.1. CRF01_AE templates

Phylogenetic analysis of U5-PBS region among CRF01_AE, CRF02_AG, and subtype G isolates
demonstrated that the tracts corresponding to I- and U-duplexes in the MAL initiation complex and also D-duplex are highly conservative (Figure 2 and Supplement 1, Figure S1). Second, a portion immediately downstream of the PBS results in the formation of unstable helix 8 (Figure 1(B)), it may be the case that this region of MAL-like templates can fold into other structures than helix 8. Third, the base change A159U found in 100% of CRF01_AE isolates, 76% of CRF02_AG isolates, but none of G subtype isolates is located at the middle position of a 3-bp extension (C58GG160/40CUG42) of an intermolecular A-loop-antisense interaction between vRNA and tRNAlys3 demonstrated for the MAL initiation complex (Goldschmidt et al., 2002, 2004; Isel et al., 1995, 1998, 1999). So, such extension may be valid for subtype G isolates, but not for CRF01_AE and most of CRF02_AG isolates. These phylogenetic observations suggest that formation of functional motifs involved in reverse transcription initiation in the MAL isolate and MAL-like isolates may have common features, but may also vary in some details.

In all CRF01_AE isolates, we observed several base changes located immediately upstream of the PAS, which results in a mutated motif 116GUUAG120 instead of the 116UGUG120 common for all HIV-1 isolates of group M. Infrequent base changes are predominantly located at positions 118 and 119 within this motif, while the extreme positions are not changed (Supplement 1, Figure S1). Upon the PAS-anti-PAS interaction between vRNA and the tRNAlys3 primer, G120 of this mutated motif (98% of CRF01_AE isolates) has the potential for base pairing with C56 of the primer to form the additional base pair G120:C56.

In most CRF01_AE templates, the optimal D-duplex conformation of fragment 1 contains two hairpins with the GAAA apical loops separated by a 9-nt linker (Figure 3(A)). The first hairpin with an irregular stem of two 3-bp duplexes is denoted as 1GAAA(3+3) and the second with a 3-bp stem is denoted as 2GAAA(3). A hairpin with the same GAAA apical loop located downstream of the PBS is depicted in a revised structural model for the NL4.3 leader region with tRNAlys3 annealed onto the PBS sequence (Pollom et al., 2013). Considering the data on NC and Vif binding destabilizing activity around the PBS for the Lai/NL4.3/HXB2 strains (Damgaard et al., 1998; Henriet et al., 2005; Wilkinson et al., 2008), it could conceivably be hypothesized that two hairpins with the GAAA apical loops in MAL-like genomes can also be a potential target for NC and/or Vif binding.

The I-duplex conformation of fragment 1 in most CRF01_AE templates adopts with ΔΔG of ~2.1 kcal/mol (Table 1), and it is accompanied by structural rearrangements of 1GAAA(3+3) to 1GAAA(3), and vice versa, 2GAAA(3) to 2GAAA(3+3) (Figure 3(A) and (B)). Since the D-duplex conformation and the I-duplex conformation of fragment 1 are structurally and energetically equivalent to those of the minimal fragment, hereafter we present the D-duplex and I-duplex conformations for fragment 1 only.

With comparison a structural motif downstream of the PBS-duplex in the D-duplex and I-duplex conformations of CRF01_AE (Figure 3(A) and (B)) with that in the D-duplex conformation of the NL4.3 template (Figure 1(C)), it is seen that this motif is very similar in the I-duplex conformation of CRF01_AE and the D-duplex conformation of the NL4.3. We hypothesize that although the D-duplex conformation of CRF01_AE templates possess all structural requirements for initiation complex formation (the PBS-, D-, U-duplexes and a free

Table 1. The average differences in free energy (ΔΔG values in kcal/mol) between D-duplex conformation and I-duplex conformation or conformations of an open PAS structure obtained by Mfold prediction of U5-PBS region fragments 1 and 2 for the MAL isolate, CRF01_AE, subtype G, and CRF02_AG isolates.

| Subtype          | I-duplex conformation | Conformations of an open PAS structure |
|------------------|-----------------------|----------------------------------------|
| MAL isolate      | Fragment 1            | PAS1a" = 8                             |
| CRF01_AE         | Fragment 1            | PAS1a" = 2.8                           |
| Subtype G        | Fragment 1            | PAS1a" = 9.9                           |
| CRF02_AG         | Fragment 1            | PAS1a" = 7.7                           |
| MAL isolate      | Fragment 2            | PAS2(pal) - 3.4                         |
| CRF01_AE         | Fragment 2            | PAS2(pal) - 3.2                         |
| Subtype G        | Fragment 2            | PAS2(pal) - 3.4                         |
| CRF02_AG         | Ext. fragment 2       | PAS2(pal) - 6.1                         |

*For detailed information on differences in free energy between D-duplex conformation and I-duplex conformation or conformations of an open PAS structure for all MAL-like templates, see Supplement 3, Tables S1–S6.

1-Duplex conformations obtained by forcing I-duplex base pairing and conformations of an open PAS structure obtained by prohibiting PAS motif from base pairing are marked by asterisk.
CAG junction), but a longer by ~20 nts structural motif of two hairpins may be an obstacle to effective initiation.

Without applying a restriction on PAS motif base pairing, fragment 1 of CRF01_AE templates stably adopts several conformations of an open PAS structure (PAS1a conformations). These conformations are very close in free energy (ΔΔG of ~3.0 kcal/mol) but differ in a middle duplex structure, the most favorable PAS1a conformation is shown in Figure 3(C). These conformations are characterized by an opening a wider tract than the PAS motif itself (by 1–5 upstream nucleotides).

Another scheme of PAS exposure in fragment 1 (PAS1b conformation) is accompanied by folding of the duplicate insertion into a self-contained hairpin with the apical loop AAAG, the AAAG(3+5) hairpin (Figure 3(D)). A self-contained motif of the duplicate insertion was earlier predicted to form in the leader region of the CRF01_AE isolate CM240 (Kasprzak, Binnewald, & Shapiro, 2005). Similar to the PAS1a conformation, this scheme also demonstrates a wider exposure than the PAS motif itself, but with a lower ΔΔG value of ~5.2 kcal/mol (Table 1).

Fragment 2 of most CRF01_AE isolates adopts the D-duplex conformation and the I-duplex conformation with ΔΔG of ~3.2 kcal/mol (Figure 3(E) and (F)) that is lower by 1.1 kcal/mol than that of fragment 1 (Table 1). A PAS motif exposure in fragment 2 is accompanied by a 5-bp elongation of the 2GAAA(3+3) hairpin stem involving an upstream portion of the duplicate insertion and the CU-rich tract, the 2GAAA(3+3+5) hairpin. This conformation is denoted as the PAS2(gaaa) conformation (Figure 3(G)).

The PAS2(pal) conformation that is very close in energy to the PAS2(gaaa) conformation (Table 1) demonstrates PAS opening with formation of two self-contained motifs. The first one is the above-mentioned self-contained hairpin of the duplicate insertion and the second is formed by an imperfect palindromic sequence AGAGAUUCCU249, the pal(4) hairpin (Figure 3(H)). Whereas a larger portion of such palindromic sequence (GAGAGAUUCCU233) was demonstrated to contribute to the dimerization process in the HXB2/NL4.3 isolates (Reyes-Darias, Sanchez-Luque, & Berzal-Herranz, 2012; Song, Kafaie, & Laughrea, 2008), here we suppose a role.

Figure 3. Mfold-calculated structures (ΔG in kcal/mol) of U5-PBS region fragment 1 (nt 110–243) (A, B, C, D) and fragment 2 (nt 113–259) (E, F, G, H) in CRF01_AE isolates exemplified by the isolate 90CF402 (GenBank accession number U51188). D-duplex conformations (A, E), I-duplex conformations (B, F), PAS1a conformation (C), PAS1b conformation (D), PAS2(gaaa) conformation (G), and PAS2(pal) conformation (H). PBS motif, PAS motif, and duplicate insertion are marked by green, orange, and blue, respectively. Mutated stretch GUUAG120 is shadowed. PBS stem extension for extended fragment 2 (nt 109–261) is shown in insert (E).
of the palindromic sequence in promoting PAS opening in the PAS2(pal) conformation of MAL-like templates. Irrespective of PAS exposure, a switchable conformation involving the palindromic sequence was recently proposed for 2D model of the Lai 5′ UTR RNA (Stephenson et al., 2013). This switch implies an involvement of a GA-rich portion of the palindromic sequence into the D-duplex and a CU-rich portion into a long-range interaction with GA-rich linker between Psi hairpin and AUG region (the CU/GA-duplex) and alternative interaction between GA-rich and CU-rich portions themselves. It is therefore likely that this switchable conformation may play a role in PAS opening also in Lai-like isolates.

Noteworthy, a PAS exposure in fragment 2 of CRF01_AE genomes is accompanied by opening of additional 2 nts immediately upstream of the PAS motif (Figure 3(G) and (H)), as distinct from other MAL-like genomes (see Sections 2 and 3). This surplus exposure may be in favor of our assumption on potential role of G120 of the mutated motif 116GUUAG120 of CRF01_AE in extension of the PAS–anti-PAS contact by additional G:C base pair.

All conformations of fragment 2 are locked by a 4 (5)-bp duplex formed by a GU-rich tract and 25GACGC255 sequence (shortly the GU/CGC-duplex) (Figure 3(E)–(H)). The extreme nucleotides of the mutated motif 116GUUAG120 are involved in the GU/C GC-duplex and a 3-bp duplex between the D- and GU/C GC-duplexes, while most variable middle nucleotides are located in a 2 × 2 internal loop. Upon extension of fragment 2, the GU/C GC-duplex is elongated by 3 bps (shown as insert, Figure 3(E)). Such an extension stabilizes each conformation by 4.5 kcal/mol and does not change ΔΔG values between conformations. Here, we present the conformations of fragment 2 in MAL-like isolates with implication for the same conformations of extended fragment 2.

The requirement for additional base pairing at the bottom of PBS stem in the NL4.3 isolate was proposed to stabilize the initiation complex in the absence of NC and also contribute to enhanced binding with RT (Iwatani et al., 2003). As distinct from the NL4.3 isolate and other MAL-like isolates studied here, we found that due to the mutated motif 116GUUAG120, PBS stem extension in CRF01_AE isolates results in a regular duplex formation and does not necessarily require the involvement of the 5′-terminal C260 of DIS hairpin. The involvement of C260 into PBS stem extension impairs its regular structure and stabilizes PBS stem only by .3 kcal/mol; its near-optimal conformation possesses a regular duplex and both G110 and C260 ejected from this duplex structure (Supplement 2, Figure S2(A)).

It is worth to note a sequence similarity between the tract 256AGGACU261 in MAL-like isolates (or 239AGGACU244 in the NL4.3 isolate). The latter tract in the NL4.3 isolate was determined as a strong NC-binding site within nucleocapsid interaction domain (Wilkinson et al., 2008). Since the tract 119AGGACU124 in the D-duplex conformation of CRF01_AE fragment 1 shares a NC-binding consensus motif comprised of a purine-rich flexible region of 3–4 nts adjacent to a helix that usually terminates in a G:C base pair, it is therefore possible to speculate that this tract in CRF01_AE templates may serve as an additional target for NC binding that is specific for these isolates.

And finally, frequent and rare base changes occurring in the U5-PBS region nt 109–261 (Supplement 1, Figure S1) are generally well tolerated in conformations of all four fragments of CRF01_AE templates (for details, see Supplement 2, comment on Figure S2; Supplement 3, comments on Tables S1, S4).

### 3.2. Subtype G templates

In most of subtype G isolates, fragment 1 adopts the D-duplex and I-duplex conformations with ΔΔG of ~2.7 kcal/mol (Table 1) and similar conformational rearrangements as observed in CRF01_AE isolates (Figures 4(A), (B) and 3(A), (B)). Two PAS1a conformations are accompanied by a regular middle duplex formation with involvement of an upstream portion of the duplicate insertion. The most favorable PAS1a conformation is shown in Figure 4(C). Similar to the PAS1a conformations of CRF01_AE isolates, a downstream portion of the duplicate insertion of subtype G isolates remains involved into the 2GAAA(3+3) hairpin and PAS exposure is characterized by opening of a wider tract than the PAS motif itself. Similar to CRF01_AE isolates, the PAS1b conformation of G subtype isolates is obtained by restriction on PAS motif base pairing and it is accompanied by folding of the duplicate insertion into a self-contained motif (Figure 4(D)), ΔΔG value is about −6.3 kcal/mol (Table 1).

As distinct from CRF01_AE templates, the folding results of fragment 2 demonstrated a lowest free energy for the PAS2(pal) conformation in 50% of subtype G isolates (Figure 4(E)). For the rest of subtype G isolates, the D-duplex conformation (Figure 4(F)) was found in optimal folding and the PAS2(pal) conformation adopted with ΔΔG of ~2.5 kcal/mol as compared to the D-duplex conformation. Similar to fragment 2 of CRF01_AE, a PAS motif exposure in fragment 2 of subtype G is accompanied by folding of two self-contained motifs or a 5-bp elongation of the 2GAAA(3+3) hairpin (Figure 4(E) and (G)). As distinct from CRF01_AE, both PAS2 conformations are more favorable than the I-duplex conformation and the PAS2(pal) conformation definitely prevails over the PAS2(gaa) conformation in
subtype G isolates (Table 1). Furthermore, PAS exposure in fragment 2 of subtype G isolates is started exactly from the first nucleotide of the PAS sequence (Figure 4(E) and (G)), while it involves extra 2 nts immediately upstream of the PAS motif in CRF01_AE isolates (Figure 3(G) and (H)).

The GU/CGC-duplex of extended fragment 2 is elongated by 4 bps with involvement of the 5'-terminal nucleotide C260 of DIS hairpin (shown as insert, Figure 4(F)). This results in formation of an irregular duplex with a GU bulge and additional stabilization of each conformation by 3.3 kcal/mol (Supplement 2, Figure S2(B)). The same structure of PBS stem extension was demonstrated for the NL4.3 isolate (Iwatani et al., 2003).

Similar to CRF01_AE isolates, rare base changes occurring in the U5-PBS region nt 109–261 (Figure 3(G) and (H)(165,169),(828,769)) are well tolerated in conformations of four fragments of subtype G templates (for details, see Supplement 3, comments on Tables S2, S5).

### 3.3. CRF02_AG templates

Two sequence variations specific for CRF02_AG isolates are observed in PBS region of all CRF02_AG isolates, a 7-nt deletion downstream of PBS and a 1-nt insertion at the 3' end of fragment 2 (Figure 2 andSupplement 1, Figure S1(165,169),(828,769)), which are to affect the conformations of interest. A 7-nt deletion leads to disappearance of the 1GAAA(3+3) hairpin common for CRF01_AE and subtype G isolates. A disruption of the usual PBS substructure caused by this deletion was earlier reported for the CRF02_AG isolate IBNG (Kasprzak et al., 2005).

The optimal D-duplex conformation of fragment 1 in CRF02_AG templates possesses a hairpin containing a larger portion of the duplicate insertion (the AAUAGG(3+4) hairpin), a 1(2)-nt linker and the 2GAAA(3) hairpin (Figure 5(A)). The D-duplex conformation with two short hairpins, a hairpin with the AAGU apical loop and the 2GAAA(3) hairpin separated by a 12-nt linker, was observed in suboptimal folding with ΔΔG value of about −6.1 kcal/mol (Figure 5(B)). This structural motif

---

**Figure 4.** Mfold-calculated structures (ΔG in kcal/mol) of U5-PBS region fragment 1 (nt 110–243) (A, B, C, D) and fragment 2 (nt 113–259) (E, F, G, H) in subtype G isolates exemplified by the isolate DRCBL (GenBank accession number AF084936). D-duplex conformations (A, F), I-duplex conformations (B, H), PAS1a conformation (C), PAS1b conformation (D), PAS2(pal) conformation (E), and PAS2(gaaa) conformation (G). PBS motif, PAS motif, and duplicate insertion are marked by green, orange, and blue, respectively. PBS stem extension for extended fragment 2 (nt 109–261) is shown in insert (F).
is somewhat similar to that of optimal D-duplex conformation in CRF01_AE and subtype G isolates (Figures 3(A) and 4(A)).

The I-duplex conformation with the AAGU(2) hairpin instead of 1GAAA(3) and the 2GAAA(3+3) hairpin was obtained as optimal structure by forcing I-duplex formation (Figure 5(C)). ΔΔG value between the optimal D-duplex conformation and the I-duplex conformation was about −7.7 kcal/mol (Table 1). Three PAS1 conformations were also obtained by restriction on PAS base pairing: the optimal PAS1 conformation with the same structural motif downstream of PBS-duplex as the optimal D-duplex conformation denoted as PAS1-specific conformation (Figure 5(D)), the PAS1b conformation (Figure 5(E)), and the PAS1a conformation (Figure 5(F)). All three PAS1 conformations of CRF02_AG lack the 1GAAA(3) hairpin.

A conformational change from D-duplex conformation to I-duplex conformation of the minimal fragment or fragment 1 in CRF02_AG isolates requires a much more energy input (about 7.7 kcal/mol) than that in CRF01_AE and subtype G isolates (about 2–3 kcal/mol). A PAS motif exposure is also significantly impeded in CRF02_AG templates. These results may be interpreted in two ways. First, a switch from the D-duplex conformation to the I-duplex conformation in CRF02_AG isolates does require a high energy input, which can imply the presence of some initiation factors, possibly specific for CRF02_AG. In this case, a switch from the I-duplex conformation to the PAS1a conformation (Figure 5(F)) proceeds with energy input of about 2.2 kcal/mol (Table 1), which is comparable with those in CRF01_AE and subtype G isolates, 2.8 and 1.7 kcal/mol, respectively.

Second, the optimal D-duplex conformation with a compactly folded 32-nt sequence between the PBS duplex and D-duplex represents a functional structure that is highly specific for CRF02_AG isolates and could be accepted by a tight structure of the initiation complex to start effective initiation bypassing the I-duplex conformation. In this case, a PAS exposure may be realized via the PAS1-specific conformation (Figure 5(D)) with

![Figure 5. Mfold-calculated structures (ΔG in kcal/mol) of U5-PBS region fragment 1 (nt 110–243) in CRF02_AG isolates exemplified by the isolate 99CMBD6 (GenBank accession number AY271690). D-duplex conformations (A, B), I-duplex conformation (C), PAS1-specific conformation (D), PAS1b conformation (E), and PAS1a conformation (F). PBS motif, PAS motif, and duplicate insertion are marked by green, orange, and blue, respectively. A 7-nt deletion is indicated by triangle.](image-url)
energy input of about 7.8 kcal/mol (Table 1). Notewor-
thy, the PAS1-specific conformation contains the same
two hairpins separated by a 1(2)-nt linker as the optimal
D-duplex conformation (Figure 5(A)). As distinct from
PAS1a conformation, the bottom duplex of this confor-
mation does not involve the duplicate insertion.

Upon folding of fragment 2 in CRF02_AG isolates,
we observed a great structural variability in the bottom
duplex of the calculated conformations and an incom-
plete PAS exposure. This variability is provoked by a
1-nt insertion between positions 256 and 257 found in
all CRF02_AG isolates (Supplement 1, Figure S1).
However, folding jobs of extended fragment 2 yielded
the D-duplex conformation and the PAS2(pal) and PAS2
(gaaa) conformations closed by the extended PBS stem
without restriction on PAS base pairing in most of
CRF02_AG isolates (Figure 6(A)–(D)).

In 76% of CRF02_AG isolates, the folding results of
extended fragment 2 demonstrated a lowest free energy
for the D-duplex conformation (Figure 6(A)) and in 24%
of CRF02_AG isolates, the PAS2(pal) conformation
adopted with lowest free energy. Similar to subtype G
isolates, the PAS2(pal) conformation is more favorable
by about −2.0 kcal/mol than the PAS2(gaaa) confor-
mation (Figure 6(B) and (D)) and obviously prevails in
CRF02_AG isolates. Both PAS2 conformations are much

Figure 6. Mfold-calculated structures (ΔG in kcal/mol) of U5-PBS region extended fragment 2 (nt 109–261) in CRF02_AG isolates exemplified by the isolate 99CMBD6 (GenBank accession number AY271690). D-duplex conformation (A), PAS2(pal) conformation (B), PAS2specific conformation (C), PAS2(gaaa) conformation (D), and I-duplex conformation (E). PBS motif, PAS motif, and duplicate insertion are marked by green, orange, and blue, respectively. PBS stem for fragment 2 (nt 113–259) is shown in insert (A). A 7-nt deletion downstream of PBS and 1-nt insertion between positions 256 and 257 is indicated by triangle and by plus, respectively.
more favorable than the I-duplex conformation which was obtained by forcing I-duplex formation (Figure 6(E)).

A 1-nucleotide insertion at the 3' end of fragment 2 results in additional base pair upon PBS stem extension and thus a G or U bulge instead of a GU bulge that is common for subtype G isolates (Supplement 2, Figure S2(C)–(E)). These and above-mentioned results demonstrated that PBS stem extension in MAL-like isolates possesses subtype-specific features: a 8-bp regular duplex in CRF01_AE isolates, a 9-bp irregular duplex with a GU bulge in subtype G, which is similar to that in Lai-like isolates, and a 10-bp irregular duplex with a single-nucleotide bulge (G or U) in CRF02_AG isolates (Supplement 2, Figure S2(A)–(E)).

3.4. Switchable conformations in the context of U5/AUG domain and dimer linkage structure (DLS) model with G-quadruplexes downstream of the DIS hairpin

To our knowledge, the structural rearrangements within the 5’ UTR of HIV-1 genome upon reverse transcription initiation (if any) are not reported. But it is known that the HIV-1 dimeric RNA structure is a key element in efficient template switching between two RNA monomers during reverse transcription (Andersen et al., 2003; Paillart, Shehu-Xhilaga, Marquet, & Mak, 2004; Sakuragi et al., 2010) and a base pairing at the DIS likely has an important function in maintaining nucleic acid structure in the reverse transcription complex (Chin et al., 2008).

The U5/AUG domain conformation with the CU/GA-duplex locking the DIS, SD, and Psi hairpins was experimentally demonstrated for the NL4.3 dimer of different lengths: nt 1–413 (Kenyon, Prestwood, Le Grice, & Lever, 2013), nt 1–712 (Lu et al., 2011), nt 1–972 (Wilkinson et al., 2008), the entire genome (Pollom et al., 2013; Watts et al., 2009) and nt 1–406 of the Lai isolate (Seif, Niu, & Kleiman, 2013). So, we explored this conformation to examine the conformational changes observed in short fragments of MAL-like templates for validity in a broader context. Such an approach can be supported by evidence that tRNAlys3 in virions exists in two forms, unextended or extended, by the first two DNA bases incorporated (Huang et al., 1997), i.e., reverse transcription may start on the condensed RNA dimer within virion particles.

However, it should be noted that the tract nt 101–120 of the Lai isolate encompassing the 5’ arm of the U5/AUG-duplex is critical for obtaining the RHA-induced conformational switch of the tRNAlys3-vRNA binary complex (slow-migrating band) and substitution mutation of this tract results in a significant decrease in the ability to initiate reverse transcription (Xing et al., 2011). These authors supposed that additional changes in vRNA not seen in U5/AUG conformation (middle-migrating band) may be required for this switch.

In the context of U5/AUG domain, the four chosen fragments of MAL-like templates correspond to PBS subdomain structure between the U5/AUG- and CU/GA-duplexes (minimal fragment and fragment 1) or between the U5/AUG-duplex and the DIS hairpin 5’ end (fragment 2 and extended fragment 2). Since both the minimal fragment and fragment 1 are sufficient for switching from the D-duplex conformation to the I-duplex conformation with minimal energy input, it can thus be suggested that this switch can occur within PBS subdomain structure between the U5/AUG- and CU/GA-duplexes. PAS2(pal) and PAS2(gaaa) conformations infer a disruption of the CU/GA-duplex that may be structurally compensated by Psi stem elongation and breaking of CGC/GCG-duplex at the bottom of elongated DIS stem that may be counterbalanced by formation of the GU/GCG-duplex. Both PBS stem extension by the bottom duplex of 3–4 bps (extended fragment 2) and a nonspecific opening of the PAS motif (fragment 1) imply a disruption of 2–3 upper bps of the U5/AUG-duplex. In vivo conditions, such a disruption may be facilitated by duplex-destabilizing activity of NC as it was demonstrated for the NL4.3 isolate by SHAPE (Wilkinson et al., 2008) and NMR (Spriggs, Garyu, Connor, & Summers, 2008).

As compared to the above-mentioned conformation of the NL4.3/Lai dimer, the BMH structure model for the Lai isolate first proposed by Berkhout’s group (Abbink & Berkhout, 2003) depicts a shortened DIS hairpin stem, but elongated Psi stem within a domain containing PBS, DIS, SD, and Psi hairpins separated by a purine-rich ring structure. We earlier demonstrated that 91% isolates of group M, including MAL-like isolates, can form an elongated Psi stem (Zarudnaya, Kolomiets, Potyahaylo, & Hovorun, 2006; Zarudnaya, Potyahaylo, Kolomiets, & Hovorun, 2007). With implication for the present study on MAL-like templates, a purine-rich ring structure may be considered as a more favorable for switching to PAS2 conformations, since a CU-rich portion of the palindromic sequence is looped out, Psi stem is already elongated and the stretch GACGC is partly single stranded. Alternatively, the GACGC sequence was recently proposed to interact with the GCGUC sequence located 10 nts downstream of Gag start codon forming a pseudoknot-like contact, the GACGC-GCGUC duplex, essential for HIV-1 RNA dimerization (Sakuragi, Ode, Sakuragi, Shioda, & Sato, 2012).

To test the above implications for structural rearrangements within U5/AUG domain, we fulfilled Mfold and UNAFold (Markham & Zuker, 2008) predictions of this domain of MAL-like templates and analyzed the most favorable structures with the CU/GA-duplex and PBS region in D-duplex, I-duplex, or PAS1 conforma-
tions and the most favorable structures with a purine-rich ring and PBS region in D-duplex, I-duplex, or PAS2 conformations. Mfold predictions of U5/AUG domain of MAL-like templates showed a preference of purine-rich ring structures with elongated Psi hairpin, while UNAfold results demonstrated lowest energies for structures with the CU/GA-duplex. In Supplement 2 (Figures S3–5), we present the most favorable U5/AUG domain structures obtained by UNAfold predictions for the CRF01_AE isolate 90CF402, the subtype G isolate DRCBL and the CRF02_AG isolate 99CMBD6. Since we were unable to discriminate between the resulting PAS2(gaaa) and PAS2(pal) conformations of PBS region and propose a functional meaning for none of them, both resulting conformations have been considered.

For CRF01_AE isolate 90CF402, ΔΔG values of four switchable conformations of U5/AUG domain with the resulting PAS2(gaaa) conformation of PBS region (Figure S3(A)–(C) and (F)) are within the range of 0–6.2 kcal/mol, those with the resulting PAS2(pal) conformation (Figure S3(A), (B), (E) and (H)) are within the range of 0–6.7 kcal/mol. For the subtype G isolate DRCBL, these values appeared somewhat lower: those with the resulting PAS2(pal) conformation (Figure S4(A)–(C) and (E)) – within the range of 0–4.7 kcal/mol and those with the resulting PAS2(gaaa) conformation (Figure S4(A)–(C) and (G)) – within the range of 0–4.9 kcal/mol. For the CRF02_AG isolate 99CMBD6, ΔΔG values of three structures of U5/AUG domain with the resulting PAS2(pal) conformation (Figure S5(A), (B) and (E)) and PAS2(gaaa) conformation (Figure S5(A), (D) and (F)) are within the range of 0–8.6 and 0–9.3 kcal/mol, respectively. The latter scheme includes the I-duplex conformation (Figure S5(D)), while the former may bypass it.

We earlier proposed a DLS model of DIS–DIS interaction (either kissing loop complex or extended duplex formation) stabilized by intermolecular quadruplexes (G- and mixed tetrads) located downstream of the DIS hairpin for 350 HIV-1 isolates of different subtypes, including MAL-like isolates (Zarudnaya et al., 2005). Such stable interactions can keep dimers in proximity over a long range during reverse transcription. G-rich tracts involved into intermolecular quadruplexes formation encompass a sequence corresponding to SD, Psi hairpins, the 3' arm of the U5/AUG-duplex, and about 30 nts downstream of AUG start codon, but do not include a sequence upstream of the DIS hairpin.

Within this model, the U5-PBS region sequence corresponds to the extended fragment 2 in MAL-like isolates and can easily adopt the conformations with PBS stem extension (the GU/CGC-duplex plus the bottom duplex of 3–4 bps). To illustrate this approach, we present structural models of the U5-PBS region extended fragment 2 and the DLS region encompassing DIS–DIS interaction and intermolecular quadruplexes (nt 109–384, MAL coordinates) for three MAL-like templates (Supplement 2, Figures S6–8). Such a structure of the U5-PBS region with D-duplex stabilized by PBS stem extension is commonly presented for the NL4.3/Lai isolates, e.g., (Iwatani et al., 2003; Liu et al., 2010; Ooms et al., 2007; Xing et al., 2011).

Thus, the conformational switches observed in short fragments of MAL-like templates can be potentially applicable to U5/AUG domain conformation and/or DLS model of DIS–DIS interaction stabilized by intermolecular quadruplexes; however, it must not be ruled out that in the context of the entire HIV-1 sequence with multiple cellular and viral factors, either conformation may be driven by the presence of these factors or alternative explanations are possible.

4. Conclusions

Switches from the D-duplex conformation to the I-duplex conformation in the minimal fragment or fragment 1 are similar in all MAL-like templates studied here, but require much more energy input in CRF02_AG. A PAS exposure in fragment 1 is nonspecific and accompanied by opening of a broader tract than PAS motif itself in all MAL-like templates. A PAS exposure in fragment 2 (or extended fragment 2) is energetically facilitated (Table 1) and affected by subtype-specific variations in U5-PBS region of CRF01_AE (the mutated motif G116 GUUAG120 upstream of the PAS motif) and CRF02_AG (a 7-nt deletion downstream of the PBS motif and 256_257 insG/C/A at the 3' end of fragment 2):

- a PAS exposure in fragment 2 of CRF01_AE templates is accompanied by opening of 2 nts immediately upstream of the PAS motif, which may be favorable for possible extension of the PAS–anti-PAS contact between vRNA and the trNAlys3 primer by the additional base pair G120:C56;
- the PAS2(gaaa) and PAS2(pal) conformations are very close in energy in CRF01_AE isolates, while the PAS2(pal) conformation prevails over the PAS2(gaaa) conformation in subtype G and CRF02_AG isolates;
- a 7-nt deletion in CRF02_AG isolates results in specific structural motif between the PBS- and D-duplexes and the PAS2(pal) and PAS2(gaaa) conformations lacking the 1GAAA(3) hairpin;
- PBS stem extension (the GU/CGC-duplex and the bottom duplex of 3–4 bps) has different structures in CRF01_AE, subtype G, and CRF02_AG isolates.

A switching between alternative conformations with D-duplex, I-duplex, and an open PAS structure observed in MAL-like templates may be supported by two types
of evidence. First, they contain the established structural elements essential for HIV-1 reverse transcription initiation (D-duplex as a structural pre-requisite for I-duplex, I-duplex, U-duplex, and PBS stem extension) as well as several elements which are not widely discussed in literature and not particularly referred to initiation process, namely hairpins with GAAA apical loops, self-contained motifs of the duplicate insertion (the AAAG(3+5) hairpin) and the downstream palindromic sequence (the pal(4) hairpin). And second, most frequent and rare base changes occurring in U5-PBS region of MAL-like templates are well tolerated by structural elements widely known and also discussed here (Supplement 3, comments on Tables S1–6).

Taken together, our findings suggest a role for the duplicate insertion of MAL-like templates in HIV-1 reverse transcription initiation process and possible mechanisms for its realization. This role is likely of two kinds. On the one hand, a switch from the dimerization-competent conformation (D-duplex) to the reverse transcription initiation-competent conformation (I-duplex) may be considered as an additional tool to prevent a premature initiation of reverse transcription in MAL-like isolates. On the other hand, an opening of PAS motif is promoted by refolding of the duplicate insertion and the palindromic sequence into two self-contained motifs (the PAS2(pal) conformation) or involvement of the duplicate insertion into the 2GAAA(3+3+5) hairpin stem elongation (the PAS2(gaata) conformation). The duplicate insertion involvement into structural motifs supporting PAS opening may be also considered as an additional tool to PAS exposure in MAL-like isolates. In Lai-like isolates, which do not possess the duplicate insertion, an opening of PAS motif is possibly promoted only by switchable conformation of the palindromic sequence.

And finally, we realize that switching from the D-duplex conformation to the I-duplex conformation and then to the conformation of an open PAS structure in MAL-like templates observed in in silico system has to be validated by in vitro and in vivo experiments. On the other part, these switchable conformations may provide a modeling basis for experimental validation and testing. Evidently, such conformational changes might be affected by the presence of RT, tRNAlys3, and other factors modulating reverse transcription initiation. However, an intrinsic propensity of MAL-like templates to adopt the I-duplex conformation and to expose the PAS motif is clearly demonstrated by our findings.

Supplementary materials
Supplementary materials dealing with sequence alignment of all MAL-like templates studied in the present work (Supplement 1, Figure S1), PBS stem extension of five MAL-like templates (Supplement 2, Figure S2), U5/AUG domain structures relevant to initiation process (Supplement 2, Figures S3–5) and structural models of U5-PBS region and DLS region containing intermolecular quadruplexes downstream of the DIS (Supplement 2, Figures S6–8) for CRF01_AE isolate 90CF402, subtype G isolate DRCBL and CRF02_AG isolate 99CMBD6. Mfold-calculated differences in free energy between D-duplex and I-duplex or PAS conformations of fragments 1 and 2 for 91 MAL-like templates (Supplement 3, Tables S1–S6) are available from the authors’ server at the URL http://cesshiv1.org/supplementary1/.

References
Abbink, T. E. M., Beerens, N., & Berkhout, B. (2004). Forced selection of a human immunodeficiency virus type 1 variant that uses a non-self tRNA primer for reverse transcription: involvement of viral RNA sequences and the reverse transcriptase enzyme. *Journal of Virology*, 78, 10706–10714. doi:10.1128/JVI.78.19.10706–10714.2004
Abbink, T. E. M., & Berkhout, B. (2003). A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. *Journal of Biological Chemistry*, 278, 11601–11611. doi:10.1074/jbc.M210291200
Abbink, T. E. M., & Berkhout, B. (2008). HIV-1 reverse transcription initiation: A potential target for novel antivirals? *Virus Research*, 134, 4–18. doi:10.1016/j.virusres.2007.12.009
Andersen, E. S., Jeeninga, R. E., Damgaard, C. K., Berkhout, B., & Kjems, J. (2003). Dimerization and template switching of the primer activation signal in tRNA annealing onto the HIV type 1 reverse transcriptase enzyme. *Journal of Virology*, 77, 3020–3030. doi:10.1128/JVI.77.5.3020–3030.2003
De Baar, M. P., De Ronde, A., Berkhout, B., Cornelissen, M., Van Der Horn, K. H. M., Van Der Schoot, A. M., … Goudsmit, J. (2000). Subtype-specific sequence variation of the HIV type 1 long terminal repeat and primer-binding site. *AIDS Research and Human Retroviruses*, 16, 499–504. doi:10.1089/08922200309160
Baudin, F., Marquet, R., Isel, C., Darlix, J. L., Ehresmann, B., & Ehresmann, C. (1993). Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. *Journal of Molecular Biology*, 229, 382–397. doi:10.1006/jmbi.1993.1041
Beerens, N., & Berkhout, B. (2002). The tRNA primer activation signal in the human immunodeficiency virus type 1 genome is important for initiation and processive elongation of reverse transcription. *Journal of Virology*, 76, 2329–2339. doi:10.1128/JVI.76.5.2329–2339.2002
Beerens, N., Groot, F., & Berkhout, B. (2001). Initiation of HIV-1 reverse transcription is regulated by a primer activation signal. *Journal of Biological Chemistry*, 276, 31247–31256. doi:10.1074/jbc.M102441200
Beerens, N., Jepsen, M. D. E., Nechyporuk-Zloy, V., Kruger, A. C., Darlix, J. L., Kjems, J., & Birkedal, V. (2013). Role of the primer activation signal in the human immunodeficiency virus type 1 reverse transcriptase enzyme. *Journal of Virology*, 76, 2329–2339. doi:10.1128/JVI.76.5.2329–2339.2002
Baudin, F., Marquet, R., Isel, C., Darlix, J. L., Ehresmann, B., & Ehresmann, C. (1993). Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. *Journal of Molecular Biology*, 229, 382–397. doi:10.1006/jmbi.1993.1041
Beerens, N., & Berkhout, B. (2002). The tRNA primer activation signal in the human immunodeficiency virus type 1 genome is important for initiation and processive elongation of reverse transcription. *Journal of Virology*, 76, 2329–2339. doi:10.1128/JVI.76.5.2329–2339.2002
Beerens, N., Groot, F., & Berkhout, B. (2001). Initiation of HIV-1 reverse transcription is regulated by a primer activation signal. *Journal of Biological Chemistry*, 276, 31247–31256. doi:10.1074/jbc.M102441200
Beerens, N., Jepsen, M. D. E., Nechyporuk-Zloy, V., Kruger, A. C., Darlix, J. L., Kjems, J., & Birkedal, V. (2013). Role of the primer activation signal in the human immunodeficiency virus type 1 reverse transcriptase enzyme. *Journal of Virology*, 76, 2329–2339. doi:10.1128/JVI.76.5.2329–2339.2002
Baudin, F., Marquet, R., Isel, C., Darlix, J. L., Ehresmann, B., & Ehresmann, C. (1993). Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. *Journal of Molecular Biology*, 229, 382–397. doi:10.1006/jmbi.1993.1041
Beerens, N., & Berkhout, B. (2002). The tRNA primer activation signal in the human immunodeficiency virus type 1 genome is important for initiation and processive elongation of reverse transcription. *Journal of Virology*, 76, 2329–2339. doi:10.1128/JVI.76.5.2329–2339.2002

mismatch base pairs. *Nucleic Acids Research*, 37, 3342–3353. doi:10.1093/nar/gkp187

Chang, S. Y., Apichartpyakul, C. A., Kuiken, C. L., Essex, M., & Lee, T. H. (1999). Sequence note sequence features downstream of the primer-binding site of HIV type 1 subtype E shared by subtype G and a subset of subtype A. *AIDS Research and Human Retroviruses*, 15, 1703–1706. doi:10.1089/08892299309748

Chin, M. P. S., Lee, S.-K., Chen, J., Nikolaitchik, O. A., Powell, D. A., Fivash, M. J., Jr., & Hu, W.-S. (2008). Long-range recombination gradient between HIV-1 subtypes B and C variants caused by sequence differences in the dimerization initiation signal region. *Journal of Molecular Biology*, 377, 1324–1333. doi:10.1016/j.jmb.2008.02.003

Damgaard, C. K., Andersen, E. S., Knudsen, B., Gorodkin, J., & Kjems, J. (2004). RNA interactions in the 5′ region of the HIV-1 genome. *Journal of Molecular Biology*, 336, 369–379. doi:10.1016/j.jmb.2003.12.010

Damgaard, C. K., Dyhr-Mikkelsen, H., & Kjems, J. (1998). Mapping the RNA binding sites for human immunodeficiency virus type-1 Gag and NC proteins within the complete HIV-1 and -2 untranslated leader regions. *Nucleic Acids Research*, 26, 3667–3676. doi:10.1093/nar/26.16.3667

Goldschmidt, V., Paillart, J. C., Rigourd, M., Ehresmann, B., Aubertin, A. M., Ehresmann, C., & Marquet, R. (2004). Structural variability of the initiation complex of HIV-1 reverse transcription. *Journal of Biological Chemistry*, 279, 35923–35931. doi:10.1074/jbc.M404473200

Goldschmidt, V., Rigourd, M., Ehresmann, C., Le Grice, S. F. J., Ehresmann, C., & Marquet, R. (2002). Direct and indirect contributions of RNA secondary structure elements to the initiation of HIV-1 reverse transcription. *Journal of Biological Chemistry*, 277, 43233–43424. doi:10.1074/jbc.M205295200

Heng, X., Kharytonchyk, S., Garcia, E. L., Lu, K., Divakaruni, S. S., LaCotti, C., … Summers, M. F. (2012). Identification of a minimal region of the HIV-1 5′-leader required for RNA dimerization, NC binding, and packaging. *Journal of Molecular Biology*, 417, 224–239. doi:10.1016/j.jmb.2012.01.033

Henriet, S., Richer, D., Bernacchi, S., Decroly, E., Vigne, R., Ehresmann, B., … Marquet, R. (2005). Cooperative and specific binding of Vif to the 5′ region of HIV-1 genomic RNA. *Journal of Molecular Biology*, 354, 55–72. doi:10.1016/j.jmb.2005.09.025

Henriet, S., Sicin, L., Bec, G., Gorelick, R. J., Marquet, R., & Paillart, J. C. (2007). Vif is a RNA chaperone that could temporally regulate RNA dimerization and the early steps of HIV-1 reverse transcription. *Nucleic Acids Research*, 35, 5141–5153. doi:10.1093/nar/gkm542

Hu, W. S., & Hughes, S. H. (2012). HIV-1 reverse transcription. *Cold Spring Harbor Perspectives in Biology*, 2, a006882. doi:10.1101/cshperspect.a006882

Huang, Y., Wang, J., Shalom, A., Li, Z., Khorchid, A., Wainberg, M. A., & Kleinman, L. (1997). Primer tRNA3Lys on the viral genome exists in unextended and two-base extended forms within mature human immunodeficiency virus type 1. *Journal of Virology*, 71, 726–728. Retrieved from http://jvi.asm.org/content/71/1/726

Huthoff, H., Bugala, K., Barciszewski, J., & Berkhout, B. (2003). On the importance of the primer activation signal for initiation of tRNA(lys3)-primed reverse transcription of the HIV-1 RNA genome. *Nucleic Acids Research*, 31, 5186–5194. doi:10.1093/nar/gkg714

Isel, C., Ehresmann, C., Keith, G., Ehresmann, B., & Marquet, R. (1995). Initiation of reverse transcription of HIV-1: Secondary structure of the HIV-1 RNA/tRNA3Lys (template/primer). *Journal of Molecular Biology*, 247, 236–250. doi:10.1006/jmbi.1994.0136

Isel, C., Ehresmann, C., & Marquet, R. (2010). *Initiation of HIV reverse transcription*. *Viruses*, 2, 213–243. doi:10.3390/v20102123

Isel, C., Keith, G., Ehresmann, B., Ehresmann, C., & Marquet, R. (1998). Mutational analysis of the tRNA3Lys/HIV-1 RNA (primer/template) complex. *Nucleic Acids Research*, 26, 1198–1204. doi:10.1093/nar/26.5.1198

Iwatai, Y., Sato, H., & Shindo, T. (2001). Variable sequences in the long terminal repeat and its downstream region of some HIV Type 1 CRF01_AE recently distributing among Thai carriers. *AIDS Research and Human Retroviruses*, 17, 863–866. doi:10.1089/088922201750252061

Kasprzak, W., Bindewald, E., & Shapiro, B. A. (2005). Structural polymorphism of the HIV-1 leader region explored by computational methods. *Nucleic Acids Research*, 33, 7151–7163. doi:10.1093/gkar1015

Kenyon, J. C., Prestwood, L. J., Le Grice, S. F. J., & Lever, A. M. B. (2013). In-gel digestion of individual RNA conformers within a mixed population reveals a dimerization structural switch in the HIV-1 leader. *Nucleic Acids Research*, 41, e174. doi:10.1093/nar/gkt690

Kurosu, T., Mukai, T., Auwanit, W., Ayuthaya, P. L., Saeng-Aroon, S., & Ikuta, K. (2001). Variable sequences in the long terminal repeat and its downstream region of some HIV Type 1 CRF01_AE recently distributing among Thai carriers. *AIDS Research and Human Retroviruses*, 17, 863–866. doi:10.1089/088922201750252061

Lalonde, M. S., Lobritz, M. A., Ratcliff, A., Chamanian, M., Athanassiou, Z., Tyagi, M., … Arts, E. J. (2011). Inhibition of both HIV-1 reverse transcription and gene expression by a cyclic peptide that binds the Tat-transactivating response element (TAR) RNA. *PLoS Pathogens*, 7(11), e1002308. doi:10.1371/journal.ppat.1002308

Levin, J. G., Mitra, M., Mascarenhas, A., & Musier-Forsyth, K. (2010). Role of HIV-1 nucleocapsid protein in HIV-1 reverse transcription. *RNA Biology*, 7, 754–774. doi:10.4161/rna.7.6.14115

Li, C., Rong, L., Russell, R. S., & Wainberg, M. A. (2000). Deletion mutagenesis downstream of the 5′ long terminal repeat of human immunodeficiency virus type 1 is compensated for by point mutations in both the US region and gag gene. *Journal of Virology*, 74, 6251–6261. doi:10.1128/JVI.74.14.6251-6261.2000

Liu, S., Harada, B. T., Miller, J. T., Le Grice, S. F. J., & Zhuang, X. (2010). Initiation complex dynamics direct the transitions between distinct phases of early HIV reverse transcription. *Nature Structural & Molecular Biology*, 17, 1453–1460. doi:10.1038/nsmb.1937

Lu, K., Heng, X., Garry, L., Monti, S., Garcia, E. L., Kharytonchyk, S., Dorjsuren, B., … Summers, M. F. (2011). NMR detection of structures in the HIV-1 5′-leader RNA that regulate genome packaging. *Science*, 334, 242–245. doi:10.1126/science.1210460
Lu, K., Heng, X., & Summers, M. F. (2011). Structural determinants and mechanism of HIV-1 genome packaging. *Journal of Molecular Biology*, 410, 609–633. doi:10.1016/j.jmb.2011.04.029

Markham, N. R., & Zuker, M. (2008). UNAFold: Software for nucleic acid folding and hybridization. *Methods in Molecular Biology*, 453, 3–31. doi:10.1007/978-1-60327-429-6_1

Mirambeau, G., Lyonnais, S., & Gorelick, R. J. (2010). Features, processing states and heterologous protein interactions in the modulation of the retroviral nucleocapsid protein function. *RNA Biology*, 7, 724–734. doi:10.4161/rna.7.6.13777

Mougel, M., Houzet, L., & Darlix, J. L. (2009). When is it time for reverse transcription to start and go? *Retrovirology*, 6, 24. doi:10.1186/1742-4690-6-24

Ooms, M., Cupac, D., Abbinck, T. E. M., Huthoff, H., & Berkhout, B. (2007). The availability of the primer activation signal (PAS) affects the efficiency of HIV-1 reverse transcription initiation. *Nucleic Acids Research*, 35, 1649–1659. doi:10.1093/nar/gkm046

Paillart, J. C., Shehu-Xhilaga, M., Marquet, R., & Mak, J. (2010). Features, processing states and heterologous protein interactions in the modulation of the retroviral nucleocapsid protein function. *RNA Biology*, 7, 724–734. doi:10.4161/rna.7.6.13777

Pollom, E., Dang, K. K., Potter, E. L., Gorelick, R. J., Burch, C. L., Weeks, K. M., & Swanstrom, R. (2013). Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-conserved structural motifs. *PLoS Pathogens*, 9, e1003294. doi:10.1371/journal.ppat.1003294

Reyes-Darias, J. A., Sanchez-Luque, F. J., & Berzal-Herranz, A. (2012). HIV RNA dimerisation interference by antisense oligonucleotides targeted to the 5′ UTR structural elements. *Virus Research*, 169, 63–71. doi:10.1016/j.viruses.2012.07.004

Roy, B. B., Hu, J., Guo, X., Russell, R. S., Guo, F., Kleiman, L., & Liang, C. (2006). Association of RNA helicase A with human immunodeficiency virus type 1 particles. *Journal of Biological Chemistry*, 281, 12625–12635. doi:10.1074/jbc.M510596200

Sakuragi, J., Ode, H., Sakuragi, S., Shioda, T., & Sato, H. (2012). A proposal for a new HIV-1 DLS structural model. *Nucleic Acids Research*, 40, 5012–5022. doi:10.1093/nar/gks156

Sakuragi, J., Sakuragi, S., Ohishi, M., & Shioda, T. (2010). Direct correlation between genome dimerization and recombination efficiency of HIV-1. *Microbes and Infection*, 12, 1002–1011. doi:10.1016/j.micinf.2010.06.012

Seif, E., Niu, M., & Kleiman, L. (2013). Annealing to sequences within the primer binding site loop promotes an HIV-1 RNA conformation favoring RNA dimerization and packaging. *RNA*, 19, 1384–1393. doi:10.1261/rna.038497.113

Slezman, D., Govers, R., & Maks, J. (2004). Dimerization of retroviral RNA genomes: an inseparable pair. *Nature Reviews Microbiology*, 2, 461–472. doi:10.1038/nrmicro903

Stephenson, J. D., Li, H., Kenyon, J. C., Symmons, M., Kleeman, D., & Lever, A. M. L. (2013). Three-dimensional RNA structure of the major HIV-1 packaging signal region. *Structure*, 21, 951–962. doi:10.1016/j.str.2013.04.008

Uppaladine, M. V., Sonavane, U. B., & Joshi, R. R. (2013). MD simulations of HIV-1 RT primer–template complex: effect of modified nucleosides and antisense PNA oligomer. *Journal of Biomolecular Structure and Dynamics, 31*, 539–560. doi:10.1080/07391102.2012.706076

Warren, K., Warrillow, D., Meredith, L., & Harrich, D. (2009). Reverse transcriptase and cellular factors: Regulators of HIV-1 reverse transcription. *Virus Rep., 1*, 873–894. doi:10.1039/v1030873

Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Jr., Swanstrom, R., … Weeks, K. M. (2009). Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 460, 711–716. doi:10.1038/nature08237

Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H., … Weeks, K. M. (2008). High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biology*, 6, e96. doi:10.1371/journal.pbio.0060096

Xing, L., Liang, C., & Kleiman, L. (2011). Coordinate roles of Gag and RNA helicase A in promoting the annealing of tRNAsLys to HIV-1 RNA. *Journal of Virology*, 85, 1847–1860. doi:10.1128/JVI.02010-10

Zarudnaya, M. I., Kolomiets, I. M., Potyahaylo, A. L., & Hovorun, D. M. (2005). Dimer linkage structure in retroviruses: models that include both duplex and quadruplex domains. *Ukr Biokhim Zh.*, 77, 5–15. Retrieved from http://www.biochemistry.org.ua/images/stories/pdf/2005/UBJ_N2_2005/Zarudnaya.pdf

Zarudnaya, M. I., Kolomiets, I. M., Potyahaylo, A. L., & Hovorun, D. M. (2006). The primary and secondary structures of HIV-1 genomic RNA region encompassing DIS, SD and Psi hairpins: in silico study. In P. A. McNamara (Ed.), *Trends in RNA research* (pp. 159–189). NY: Nova Science Publishers.

Zarudnaya, M. I., Potyahaylo, A. L., Kolomiets, I. N., & Hovorun, D. M. (2007). Diversity of DIS, SD and Psi hairpins in HIV-1 isolates of group M: in silico study. *Ukr Biokhim Zh.*, 79, 68–84. Retrieved from http://ubj.biochemistry.org.ua/index.php?content=archive&article=225

Zarudnaya, M. I., Potyahaylo, A. L., Kolomiets, I. M., & Hovorun, D. M. (2013). Structural model of the complete poly(A) region of HIV-1 pre-mRNA. *Journal of Biomolecular Structure and Dynamics, 31*, 1044–1056. doi:10.1080/07391102.2012.718530

Zuker, M. (2003). MFold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31, 3406–3415. doi:10.1093/nar/gkg595