*Nucleic Acids Research, 2005, Vol. 33, No. 22* ***7151–7163***
*doi:10.1093/nar/gki1015*

# Structural polymorphism of the HIV-1 leader region explored by computational methods

**Wojciech Kasprzak, Eckart Bindewald and Bruce A. Shapiro[1,*]**

Basic Research Program, SAIC Frederick, NCI-Frederick , Building 469, Room 150, Frederick, MD 21702, USA and [1]Center for Cancer Research Nanobiology Program, National Cancer Institute, Building 469, Room 150, NCI-Frederick, Frederick, MD 21702, USA

Received September 1, 2005; Revised October 24, 2005; Accepted November 28, 2005

## ABSTRACT

**Experimental studies revealed that the elements of the human immunodeficiency virus type 1 (HIV-1) 5′-untranslated leader region (5′-UTR) can fold *in vitro* into two alternative conformations, branched (BMH) and 'linearized' (LDI) and switch between them to achieve different functionality. In this study we computationally explored in detail, with our massively parallel genetic algorithm (MPGAfold), the propensity of 13 HIV-1 5′-UTRs to fold into the BMH and the LDI conformation types. Besides the BMH conformations these results predict the existence of two functionally equivalent types of LDI conformations. One is similar to what has been shown *in vitro* to exist in HIV-1 LAI, the other is a novel conformation exemplified by HIV-1 MAL long-distance interactions. These novel MPGAfold results are further corroborated by a consensus probability matrix algorithm applied to a set of 155 HIV-1 sequences. We also have determined in detail the impact of various strain mutations, domain sizes and folds of elongating sequences simulating folding during transcription on HIV-1 RNA secondary structure folding dynamics.**

## INTRODUCTION

There are several regulatory domains in the 5′-untranslated leader region (5′-UTR) of the HIV-1 genome. They are responsible for regulation of viral life-cycle functions, such as reverse transcription, dimerization and encapsidation, as well as gene expression functions, such as translation, transcription or polyadenylation. The multi-functional character of this region depends on its secondary structure motifs, and it has been postulated to correlate with its structural polymorphism functioning as a switching mechanism in *in vitro* experiments. (1–3). Of particular interest to us were the well-characterized secondary structure motifs of the leader RNA. These include the 5′ R (repeat) region's TAR and poly(A) structures. The TAR–protein complex (with viral Tat and cellular cyclin T) plays a role in activating transcription (4), while the poly(A) stem–loop structure suppresses the 5′ polyadenylation signal (AAUAAA) (5,6). Both of these R region structures are believed to play a role in strand transfer during the reverse transcription (7). Downstream from the R region there is a highly conserved extended structural domain, which includes the reverse transcription primer binding site (PBS) and the primer activation signal (PAS) essential to initiation and elongation of reverse transcription (8,9). Downstream from the PBS/PAS domain there are several stem–loop motifs playing a role in the dimerization and packaging of the HIV-1. Dimerization occurs at a 6 nt long self-complementary sequence (commonly GCGCGC or GUGCAC) in the dimer initiation site (DIS) folding into a hairpin loop structure (SL1/DIS). Two such hairpin structures from two copies of the HIV-1 genome interact by complementary base pairing to create the so-called 'kissing loop' complex (10–18). In most published 5′-UTR secondary structure models the DIS SL is followed by three more stem–loop motifs: the major splice donor (SL2/SD), the packaging signal (SL3/Ψ) and a hairpin structure (SL4) including the *gag* start codon (19–22). However, some *in vitro* results, as well as our folding results for the 368 and 618 nt domains of the HIV-1 LAI, show a possibility that the SL2 and SL4 structures are collapsed into a longer linear motif ending with the SL3 structure—the so-called extended Ψ structure (23).

We studied the structures and folding pathways of the HIV-1 leader region using our massively parallel genetic algorithm (MPGAfold) for RNA structure prediction (24–28) and a consensus probability matrix algorithm that combines thermodynamic predictions of individual sequences according to a sequence alignment (40). We examined effects of MPGAfold population variations, a method for identifying multiple folding states (explained in Materials and Methods). We

*To whom correspondence should be addressed. Tel: +1 301 846 5536; Fax: +1 301 846 5598; Email: bshapiro@ncifcrf.gov

also determined the impact of folding domain sizes on the folding pathways. While the shorter domains (280 and 368 nt) followed the sequence lengths used in the published experimental data, the 618 nt long sequence fragment (in HIV-1 HXB2 coordinates) was selected around the end of a self-contained structural domain, predicted with Mfold 3.1 (29,30) for the entire 9229 nt long genomic RNA of HIV-1 LAI (5′ R to 3′ R) and several other strains. Our folding experiments performed for multiple HIV-1 strains show that two mutually exclusive conformation types, characterized *in vitro* for HIV-1 LAI, can form, but the thermodynamic balance between them varies for different strains as well as with the length of the folding domain for a given strain. One conformation type (with a certain degree of variability), to which we will refer as the branched (B or BMH, for Branched, Multiple Hairpins), is characterized by the presence of the SL1 (DIS) motif and the poly(A) stem–loop. The alternative 'linearized' structure pairs up these two motifs in long-distance interactions (LDI) thus forming a long, linear motif. This LDI structure pairs up the self-complementary bases of the branched structure's DIS hairpin loop with the 3′ side of the poly(A) SL, and exposes the poly(A) signal (AAUAAA) in an internal loop. The existence of the LDI form in domains of varying length was confirmed by *in vitro* experiments for two subtype B strains (1–3,23,31). MPGAfold results show two functionally equivalent LDI conformer types, which occlude the DIS sequence in interactions with two alternative 5′ regions. As this proposed structural switch includes the DIS, which is crucial for retroviral replication, it has been proposed that the BMH conformation is dimerization and encapsidation competent, while the LDI structure is the translation competent form [(18) and references therein]. A choice of a relatively short fragment of 280 nt, not containing what we believe to be a complete self-contained folding domain, leads to a strong preference, based on the free energy calculations, for the linear conformation (2). Our computations show that longer folding domains close the free energy gap between the LDI and the BMH conformations, in some cases reversing the fitness order, and definitely suggest that in the vast majority of cases the two functional states are not separated by large energy differences. A recent study has shown phylogenetic support only for the BMH conformation (31). Our consensus probability matrix algorithm, on the other hand, adds support for the multi-state model and extends the MPGAfold's results to a large set of HIV-1 sequences, indicating the existence of the BMH as well as the two functionally equivalent types of key LDI interactions, one similar to the HIV-1 LAI structure, and the other exemplified by the predicted HIV-1 MAL long-distance interactions. This new finding of an alternate LDI conformation demonstrates flexibility in switching off the dimer competent BMH conformation in different strains and suggests the functional importance of the LDI structure.

The first *ex vivo* results appear to support only the BMH conformation, but without the U5–AUG interaction (32). Our results of folding under conditions simulating transcription show a very strong propensity to form the BMH structures, including the U5–AUG interaction, but they also indicate potential for structural rearrangements, which leaves open the possibility of assuming the BMH and LDI conformations by the full-length transcripts.

## MATERIALS AND METHODS

### The massively parallel genetic algorithm

Our massively parallel genetic algorithm for RNA structure prediction (MPGAfold) is based on the principles of such algorithms originally described by Holland (33). It applies the fundamental operators of mutation, recombination and selection to large populations of evolving RNA/DNA secondary structures, and it uses RNA/DNA structure free energy as a fitness criterion (29). Because of MPGAfold's stochastic nature, multiple runs are needed to acquire consensus results. The output data may consist of the end-of-run best fit (lowest free energy) structures of multiple runs; end-of-run (final) population consensus structures; or the maturing RNA structures derived from individual runs. Results presented in this study are based on the predicted consensus structures. When we talk about best fit structures, to keep the text short, we refer to the lowest free energy conformers among the population consensus structures in a set of runs at a given population or in multiple population runs. MPGAfold has been described in detail earlier (24–28), but we will briefly summarize its features below.

One RNA structure evolves in one population element (PE) based on its eight nearest neighbors from a 2D array of PEs. Two parent structures are selected from each eight-neighbor region to produce a new child structure through recombination and mutation. The child structure is placed in the center of the eight-neighbor region. In the actual implementation thousands of structures evolve in parallel at each generation. The algorithm is quite scalable (28), and it is possible to vary the population sizes used, e.g. from 2K to 256K.

The enhancements to the basic algorithm (25–27,34) include the ability to simulate sequential folding of the RNA multiphase runs with population seeding to simulate stages of RNA processing, the ability to predict H-type pseudoknots, and biasing to promote the formation of specific structural motifs. A special form of biasing with stems designated as 'sticky' promotes their passing to future generations once they naturally appear in a structure. A mutation operator incorporating an annealing function permits the statistical determination of convergence criteria. Energy calculations for stem stacking across multibranch loops and free ends (EFN2) in every generation are also included (29). The sequential folding capability used in part of this study simulates the effects of RNA transcription or synthesis on the folding process. It can add or remove a user-specified number of nucleotides at each generation, allowing stems to enter a maturing structure only if they fit the sequence size at the current generation. Based on experience from other studies (34,35) we used the elongation rate of 2 nt per generation. One has to keep in mind, however, that this elongation rate does not necessarily reflect actual RNA synthesis rates.

The algorithm has been shown to focus on a few significant conformations, to be capable of elucidating metastable structures, and it has been successfully applied to determine functional folding intermediates for different RNA sequences (34–37). It should also be noted that MPGAfold initiates folding by concurrent formation of multiple hairpin structures, which most likely simulates the way RNA normally folds. This is a different paradigm than is normally found in dynamic programming algorithms.

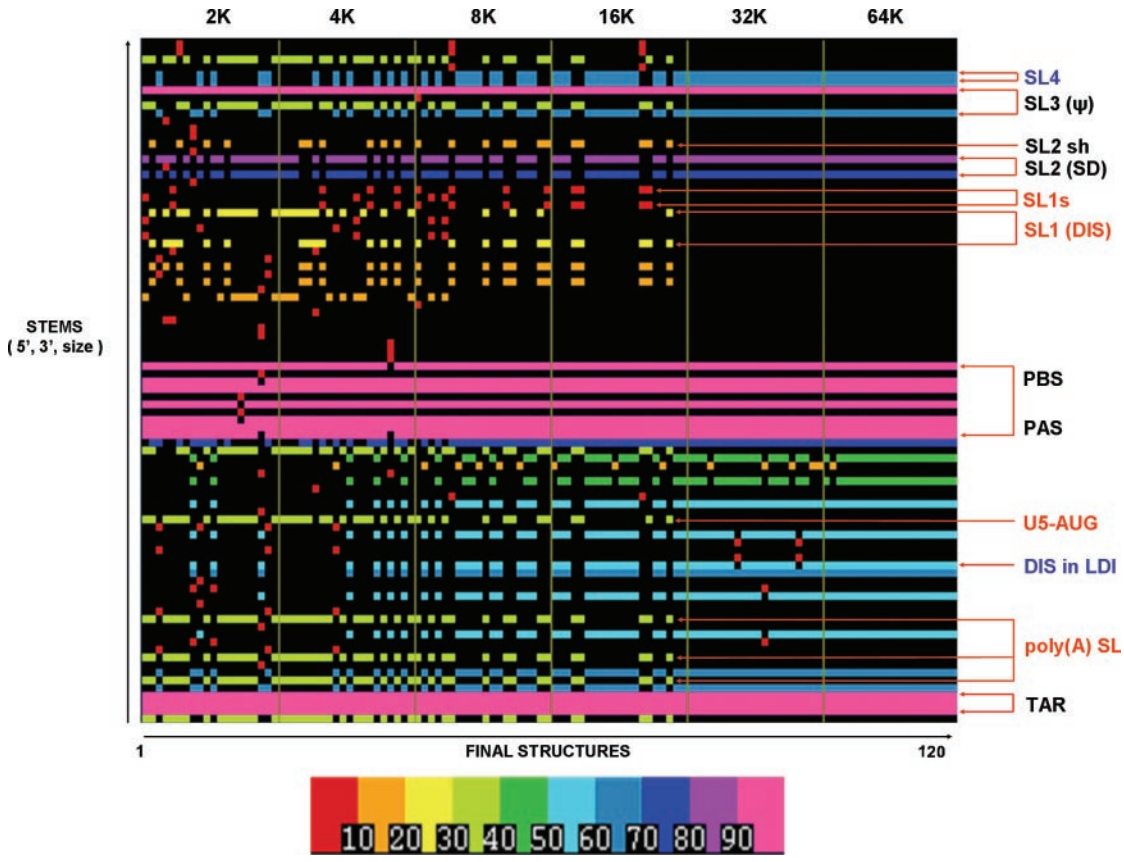


**Figure 1.** Stem trace depicting full transition of HIV-1 MN (366 nt) secondary structures from branched (BMH) to linear (LDI) conformation in the genetic algorithm runs (MPGAfold) with populations varying from 2K to 64K (1K = 1024 structures). Key structural motifs are labeled in red for the BMH conformers and in blue for the LDI structures. Motifs labeled in black are conserved in all conformations. MPGAfold is run 20 times at each population level, and final results of each run are depicted here. Stems, defined as unique triplets (5′ and 3′ size), are plotted along the vertical axis, while consecutive final structures are laid-out along the horizontal axis. Thus a structure contains all stems intersected by a vertical line. Stems are color-coded based on their frequency of occurrence, indicated by the color scale bar below the plot.

## Data analysis

We have employed our own package STRUCTURELAB (38), which includes Stem Trace, a tool helpful in interpreting the results of folding algorithms (39). It presents a 2D, interactive plot in which each position along the *X*-axis represents a new structure as a set of stems, and each position along the *Y*-axis represents an individual helical stem encoded as a unique triplet (5′ start position, 3′ stop position, stem size). Thus, the frequently occurring stems can form horizontal lines in a stem trace plot and are color-coded based on their frequency of occurrence. Stem Trace produces a view orthogonal to the more familiar base pair/stem histogram (also known as a dot plot), and its main advantage is that structural motifs are easy to identify in their original full structure context. Also, importantly, clusters of full structures with a frequency of occurrence lower than 50% can be explicitly identified.

Results of the genetic algorithm's population variation runs can be merged into one normalized stem trace plot, which is a good indicator of alternate structural conformations for a given sequence. Lower population MPGAfold run results usually depict lower fitness states, in which the analyzed sequence may be trapped, and which appear as metastable states early in the high population folding runs. This technique was used in our study to determine existence of both the LDI and the BMH states in multiple sequences and is illustrated in Figure 1.

## The consensus probability matrix algorithm

We also analyzed the HIV-1 LTR region with our version of a probability consensus matrix. This method was developed as part of a machine learning algorithm for RNA secondary structure prediction based on mutual information and thermodynamic considerations (40). In brief it works as follows: for each sequence in an alignment, a probability matrix is computed using RNAfold (41). Each element of this matrix contains RNAfold's estimated probability for that corresponding pair of nucleotides to form a base pair. An average probability matrix is generated by averaging and superposing the individual matrices according to the alignment. Each element of this average probability matrix is multiplied by an element-specific weighting factor. The weighting factor is a non-linear function (the logistic function) of the fraction *f* of sequences that have, according to RNAfold, a greater than zero probability of forming a pair for the nucleotides in question. We found that this non-linear rescaling leads to higher accuracy in predicting consensus base pairs than a simple average (40). Using a test set of RFAM alignments (42,43), we calibrated this method and rescaled the scores of the

consensus probability matrix, so that element (*i*,*j*) in the resulting matrix approximates the probability that positions *i* and *j* correspond to a base pair in the consensus secondary structure of the alignment.

In this study we have used an alignment of 155 HIV-1 sequences obtained from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/). The alignment corresponds to a 368 nt segment of the LTR region (HXB2R positions 455 to 822 according to the Los Alamos Database). Disallowing sequence fragments shorter than the region of interest, we originally obtained 180 HIV-1 sequences from the database (as of August 2, 2005). Removing all redundant sequences, i.e. those whose aligned residues are identical to another sequence in the alignment, led to the 155 sequences of all HIV-1 subtypes.

## RESULTS

We applied MPGAfold to the 280, 368 and 618 nt long fragments of HIV-1 LAI, and to their equivalents in 12 other strains, based on sequence alignments from the Los Alamos National Laboratory HIV sequence database (http://www.hiv.lanl.gov/). The examined strains are LAI, HXB2R, MN (subtype B), 99ZACM9, 93IN101 (C), NDK, ELI (D), CM240 (01_AE), IBNG (02_AG), U455 (A1), MAL (12_BF, a recombinant of subtypes A, D and K), ANT70 (O) and CPZGAB (CPZ). These strains represent the prevalent HIV-1 subtypes, as well as an outlier strain (ANT70) and the closest SIV strain (CPZGAB). The folds were performed with populations ranging from 2K to 64K for the 280 nt domain, and from 4K to 64K for the longer domains.

We will use the nomenclature introduced earlier to describe the secondary structure types; branched (BMH) and linearized (LDI). In the case of the 368 and 618 nt domains we tracked two observed types of the branched topology. One is the B(U5–AUG), which adds the highly conserved U5–AUG duplex to the SL1 (DIS) and poly(A) SL motifs (Figure 2A). The other is the B(LD) conformation, which replaces the U5–AUG, stem with long-distance interactions (Figure 2D). In the linearized (LDI) structures we kept track of the DIS-poly(A) long-distance interactions, as well as the strongly conserved R–Gag duplex (Figure 3A and B). Structures combining motifs from the branched and linear conformations, such as U5–AUG and R–Gag interactions, will be called hybrids. Conformations retaining the poly(A) SL and other key BMH stems, but occluding the DIS sequence, will be called branched alternatives, B(Alt).

The notation we use below to describe particular stems employs the (5′-position 3′-position stem size) triplet convention in the HXB2R sequence coordinates. The following designations were used to denote key structural motifs with observed variations: SL1 (DIS) denotes stems (243 277 4) and (248 270 7) (Figure 2A). SL1s refers to stems (243 277 4), (248 272 2) and (250 268 5) (Figure 2B). SL1c denotes several variations, which partly pair-up the DIS nucleotides resulting in a smaller, compact hairpin loop. In some cases SL1 is outcompeted by an up-stream alternative hairpin (237 252 5). The SL2 (SD) motif refers to stem (282 300 4) with supporting stem (286 295 3), while SL2sh is its 5′-shifted variation (288 301 5). References to the LAI-type and MAL-type LDI stems denote the two major types of the observed linearizing long-distance interactions; (90 260 8) and (102 264 9), shown in Figure 3A and B, respectively, and marked in yellow and purple, respectively, in Figure 4A.

Our folds of the 280–290 nt and the 368 nt fragments of HIV-1 LAI and HXB2R were aimed at detail validation of MPGAfold results based on the reported experimental data for this strain (1–3,23,31,44). In the current study we predicted that 11 of the 13 tested 280 nt long HIV-1 sequences undergo full transition from the less fit BMH conformations in low population runs to the better fit LDI structures in high-population runs. HIV-1 ANT70 and SIV CPZGAB do not fold into the LDI conformations.

Examination of multiple strains and of their longer domains also went beyond the experimental data and was meant to determine the impact of sequence variations and extended sequence context on the motifs crucial to the BMH and LDI conformations [poly(A) SL and SL1]. While working on the 368 nt folds we also examined the 355 nt long domain which ends immediately past the SL4 structure, since some of our explorations of the HIV-1 self-contained folding domains pointed to it as a potential candidate. For this domain nearly deterministic solution spaces at all populations show only the BMH structures. The 13 nts past position 355 are critical to formation of the LDI conformers, providing the 3′ nucleotides in the R–Gag interaction. (60 364 7) On the other hand, in the BMH structures these nucleotides are involved only in short distance interactions in the 368 nt domain folds, and in the 618 nt domain structures they participate in long-distance interactions with downstream nucleotides, thus making them independent of the key BMH motifs (Figures 2A and 3). As a result, a sequence fragment ending at position 355 strongly favors the BMH conformation.

We also studied the influence of mutations in the DIS self-complementary hexamer on the ability of the HIV-1 LAI's 368 nt domain to fold into BMH and LDI conformations. Combining the experimental results on the impact of mutations in this region on dimerization (45) with our folding results, we can see that only one mutated hexamer (GCGCGU) retains a high dimerization rate as well as the transition pattern between the BMH and LDI conformations. Thus it appears that very little can be changed in the wild-type DIS hexamer sequences without losing the ability to effectively dimerize and maintain the two-state folding pattern at the same time.

### Folds of the 5′ 368 nt and 618 nt wild-type domains

In the process of analyzing the predictions for the 5′ fragments of the 13 strains equivalent in length to the HIV-1 HXB2R's 368 and 618 nt long domains we paid special attention to the latest experimental results. For example, studies on the initiation of the reverse transcription (8,46–48), as well as the phylogenetic studies using the program Pfold (49) and enzymatic probing (3,23,31) pointed to the single M-loop topology as the true structure of the PBS substructure region and stressed the importance of the proximity of the PBS and PAS regions. Because of that and since MPGAfold predicted a single M-loop for the majority of sequences across different subtypes, we introduced a 'sticky stem' bias, corresponding to the experimentally verified stem, in the cases where a double

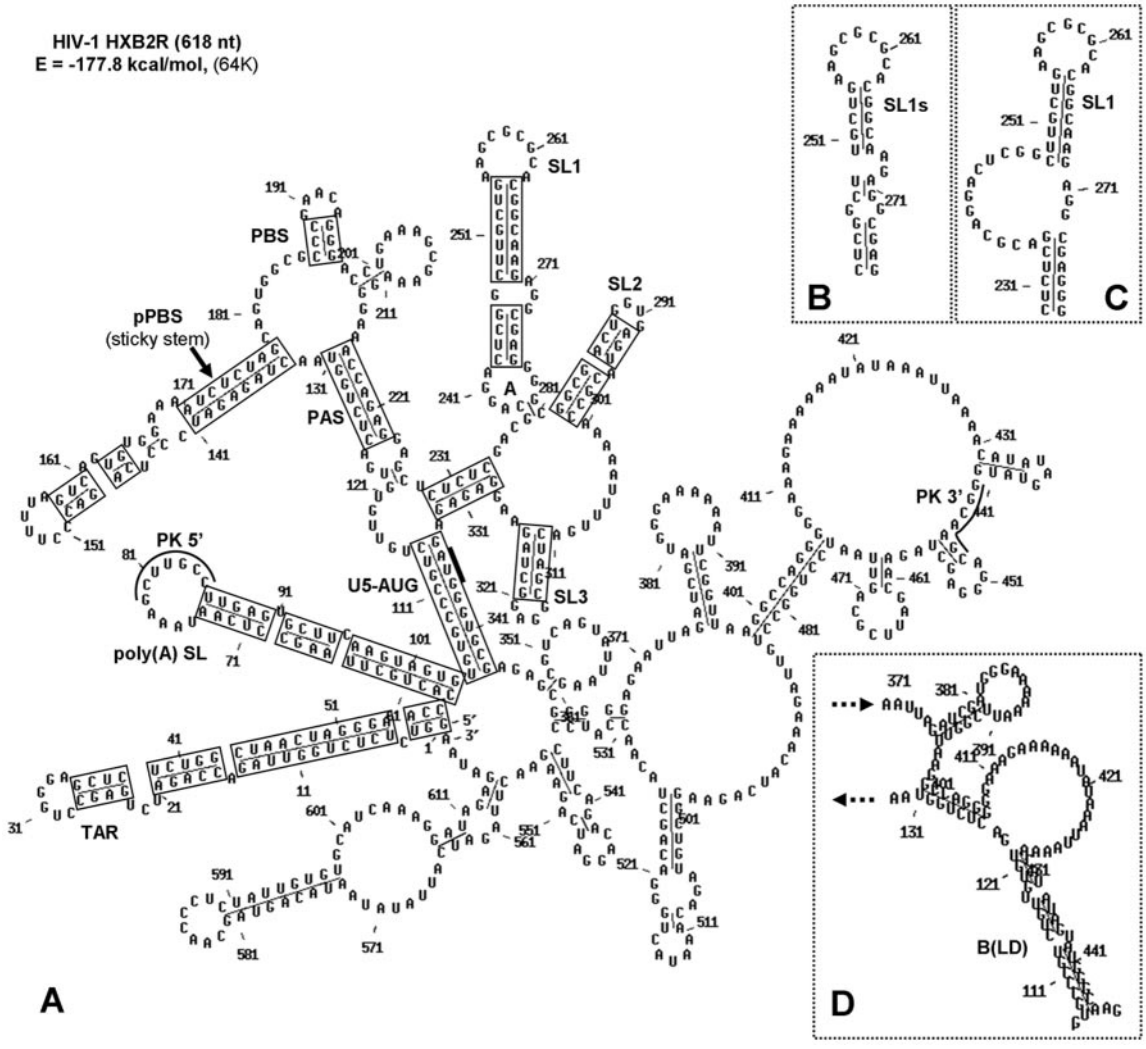


**Figure 2.** A representative BMH conformation of the B(U5–AUG) type. (**A**) predicted for the 618 nt domain of the HIV-1 HXB2R strain. Boldface labels indicate key structural motifs discussed in the text. Loop A at the bottom of the SL1 motif is predicted with high frequency for the strains HXB2R, MN and ELI (see text). Dotted-line insets show two SL1 variations; slightly better fit SL1s (**B**), and phylogenetically supported, less fit and transitional motif (**C**). Solid line boxes superimposed on the structure drawing indicate interactions supported by the consensus probability matrix algorithm in the 368 nt domain, to which the BMH and LDI alternatives are confined, based on a set of 155 HIV-1 sequences. Nucleotides open to participate in a proposed pseudoknot interaction are labeled as PK 5′ and PK 3′. Bottom inset (**D**) illustrates the key long-distance interactions of the branched conformation B(LD), which was predicted as a significant alternative to the B(U5–AUG) for some strains (see text). One side effect of the B(LD) interactions, not shown in the inset, is the presence of the SL4 motif, normally associated with the LDI conformers.

M-loop PBS substructure was dominant in solution spaces. The same bias was also applied in strains with another structural motif separating the PBS and the PAS motifs (HIV-1 MAL, IBNG, U455 and CM240). The 'sticky stem' was defined as (134 178 8), in HXB2R coordinates, and we will refer to it as the pPBS (for pre-PBS) stem. Use of this bias yielded topological proximity of the PBS and PAS motifs in the predicted structures, lowered motif variability in the solution spaces of some strains, and decreased the number of alternative branched and hybrid structures (U455, IBNG, MAL). While the free energies of the structures predicted in the pPBS biased runs were generally 0.5–4.0% higher for different strains, compared with unbiased runs, the two conformation type transitions were not impacted negatively. Thus the BMH and LDI structures were stressed as the two main conformation types shared by nearly all the tested strains. Specific biases applied are indicated in Tables 1 and 2.

Since the published LDI conformations for the 368 nt HIV-1 LAI domain suggested the existence of the extended Ψ structure motif, which collapses the SL2 and SL4 stem–loops into a longer linear motif with the SL3 (Ψ) stem preserved at the top of it (23), we tracked this motif in our results. In brief, in the HIV-1 LAI's 368 nt domain, the extended Ψ motif was predicted with high frequency (70% of LDI structures in high-population runs). However, in the 618 nt domain it was part of the LDI conformers only twice in all the population runs. Only a few less fit HIV-1 MN, 93IN101 and IBNG LDI structures included this motif as well.

Our folds yielded a novel alternative branched conformation B(LD) in the 618 nt domain folds. In it the U5–AUG stem, (105 344 10/11), is replaced with longer distance interaction stems (107 446 7) or (108 445 6), extended by stems (114 438 4) and (119 434 4), which also produce a large A and U rich bulge loop (Figure 2D). The B(LD) conformation

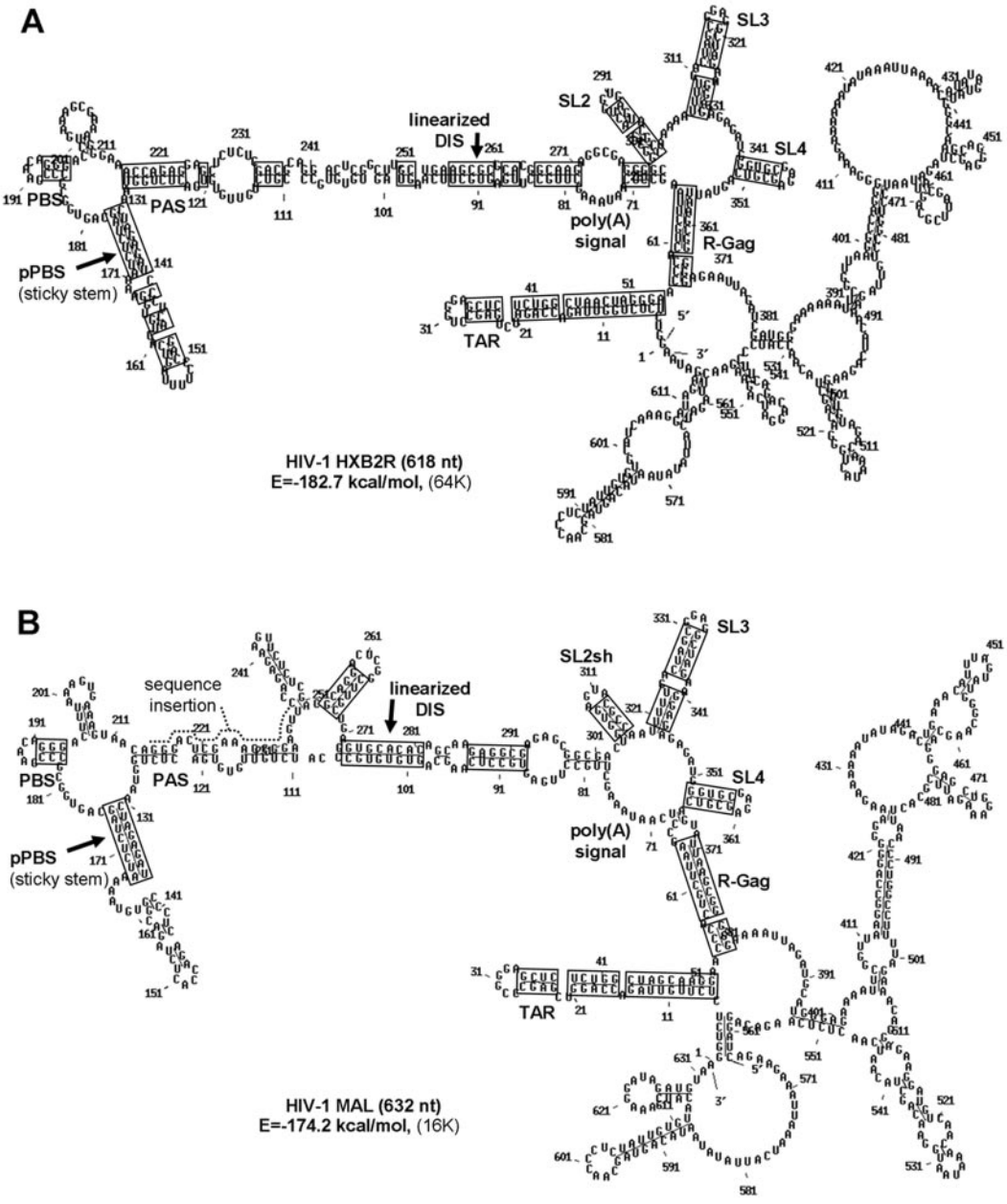


**Figure 3.** Representative LDI conformations with the LAI-type [shown for HIV-1 HXB2R in (**A**)] and the MAL-type (**B**) DIS interactions predicted by MPGAfold for the 618 nt domain. Key structural features discussed in the text are labeled in boldface. Solid line boxes denote interactions supported by the consensus probability matrix algorithm in the 368 nt domain where the BMH and LDI conformations are present, based on a set of 155 HIV-1 sequences.

is energetically the best and the dominant solution for the HIV-1 LAI and showed a strong presence in HXB2R final folds. In several other strains (MN, CM240 and ANT70) a small fraction of the final MPGAfold solutions also shows this topology. Our results showed that while the stem (107 446 7) forms at different stages of folding runs in all the strains, it has to be supported by the other two stems in order to form an energetically viable structure.

In the cases where the LDI conformers were not predicted, or predicted in a very low percentage of runs, we also employed 'sticky stems' biasing towards one of the two LDI types observed in the unbiased runs of other strains (Figure 4A). Such biasing helped us to establish if poor fitness or other factors were responsible for the low propensity to fold LDI structures. Detailed results for all the strains are listed in Tables 1 and 2. Below we discuss the most important trends observed in the MPGAfold solution spaces.

### Subtypes B (LAI, HXB2R, MN) and D (NDK, ELI)

In the 368 nt domain all the subtype B and D strains, are predicted to undergo full transitions from the strongly dominant B(U5–AUG) type conformation to the strongly dominant and energetically best fit LAI-type LDI structure (Figure 1 and Table 1).

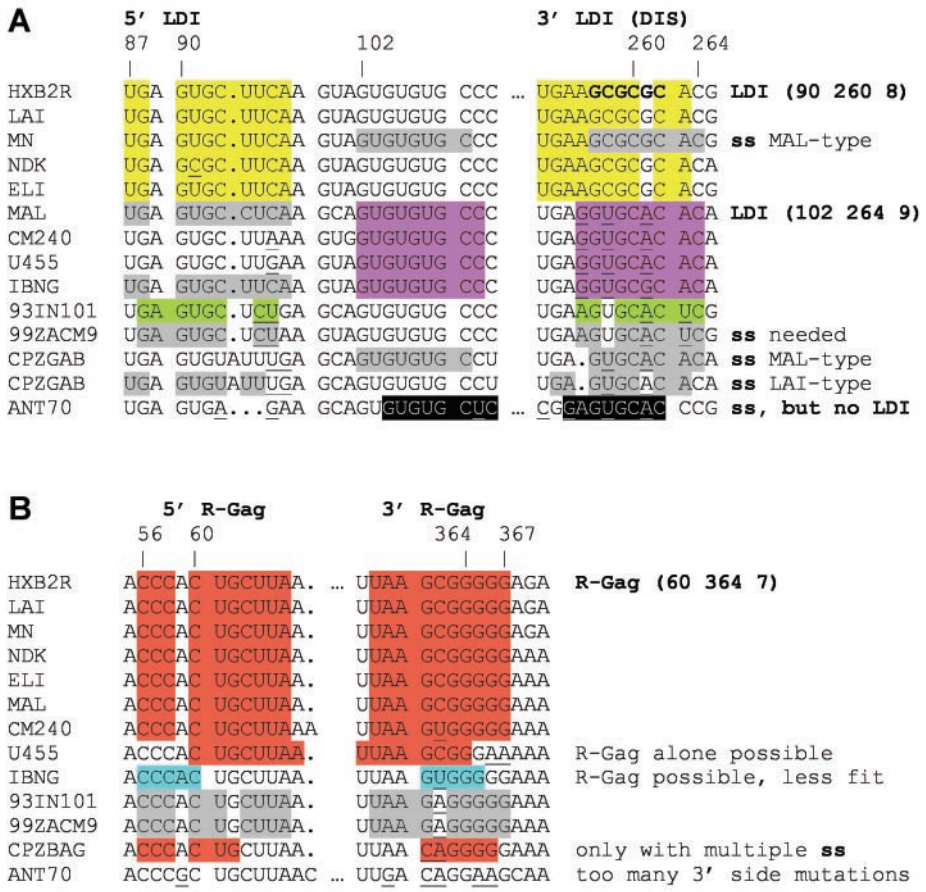


**Figure 4.** The LDI conformation motifs as a function of sequences. (**A**) The long-distance interaction types predicted in unbiased MPGAfold runs are highlighted in three colors corresponding to the observed variations. Yellow denotes the LDI folds of the subtype B and D sequences, corresponding to stems (87 263 2) and (90 260 8) in HXB2R coordinates. Highlighted in purple are the LDI structures best represented by strains MAL (subtype ADK) and U455 (subtype A1) and corresponding to stem (102 264 9) in HXB2R coordinates. Green highlights subtype C sequences folding into a variation of the LAI-type LDI. Highlighted in gray are stems designated as 'sticky' (ss) in biased MPGAfold runs, in which the full structures were predicted to be less fit than both the BMHs and LDIs, with exception of 99ZACM9 where other factors played a role (see text). Highlighted in black is a 'sticky stem' applied to HIV-1 ANT70 (type O), which occludes the DIS site, but does not disrupt the 5′ poly(A) SL motif, resulting in an alternative branched structure. Mutations relative to the HXB2R sequence are underlined. (**B**) The R–Gag interactions observed in the MPGAfold results. Red denotes stems (60 364 7) and its companion (56 367 3) in HXB2R coordinates. Blue highlights show a motif dominant in the best fit LDI conformers of the IBNG strain, in which the regular R–Gag interactions are possible, but the overall fitness of linear structures containing them is low. Stems highlighted in gray had to be designated as 'sticky' (ss) in biased GA runs, in order for them to be included in the LDI conformers. There are too many mutations in the strain ANT70 (type O) for the R–Gag to form. Mutations relative to the HXB2R sequence are underlined.

The 618 nt domain folds of all the subtype B and D strains produce predominantly branched conformers of type B(U5–AUG) and B(LD). The LDI structures range from 15% to 45% of all solutions. A small number of them exhibited the alternative, MAL-type LDI stem (102 264 9) in the 2.2% less fit linear conformations of the LAI and HXB2R sequences (Figure 4A). While in the shorter folding domain the LDI conformers are the best fit, in the longer domain the best fit structures belong to all the topology types shown in Tables 1 and 2, underscoring the delicate free energy balance between them.

### Subtype A1DK (MAL)

In the 382 nt domain the BMH structures and the branched-linear hybrids are always in the minority of all results (up to 40%). The LDI conformations are dominant in 16K and higher population runs (up to 90% in the 64K runs). The MAL's LDI stem (102 264 9) is distinct from that predicted for subtypes B and D (Figures 3B and 4A).

In the 632 nt domain folds it is the LDI and the hybrid conformers that are in the minority of solutions (up to 15% in the 16K runs), while the vast majority of solutions is of the BMH type (up to 95% in 64K). Alternatives to the SL1 motif dominate the predicted branched conformations, with a few less fit structures containing the SL1 motif (reflected in Table 1 entries). We have to keep in mind, however, that the NC protein can help stabilize the SL1 motif [see (1) and references therein]. Thus in the case of HIV-1 MAL we can see the BMH and the LDI conformations, but the transition from one to other is not as striking as in the subtype B and D sequences.

### Subtypes 01_AE (CM240), A1 (U455) and 02_AG (IBNG)

Despite some key differences in their sequences, the subtype A strains consistently form MAL-type LDIs (Figures 3B and 4A). They also exhibit a strong tendency to fold into branched

**Table 1.** Free energies of major conformations predicted by MPGAfold for the HIV-1 and SIV strains listed in 368 nt long 5′ domains

| | BMH (U5–AUG) (kcal/mol) | B(LD), B(Alt.) and Hybrids (kcal/mol) | LDI (kcal/mol) | LDI vs. BMH free energy difference (%) |
|---|---|---|---|---|
| **LAI**[a] | −130.1 | | **−134.1** | 3.0 |
| **HXB2R**[a] | −133.0 | | **−137.6** | 3.3 |
| **MN** | −136.4 | | **−141.9** | 3.9 |
| **NDK** | −124.8 | | **−129.2** | 3.4 |
| **ELI** | −137.7 | | **−143.3** | 3.9 |
| **MAL**[a] | −126.1 | −129.4 | **−130.8** | 3.6 |
| **CM240**[a] | −136.0 | **−123.4** | **−138.4** | 1.7 |
| **U455**[a] | −117.6 | −126.9 | −116.8 | 0.7 |
| **IBNG**[a] | **−128.9** | | −120.0 | 7.0 |
| **93IN101**[b] | −131.5 | **−133.7** | −131.7 | 0.1 |
| **99ZACM9**[b] | −134.0 | **−138.0** | −131.8 | 1.6 |
| **ANT70** | **−153.7** | −141.2 | None | N/A |
| **CPZGAB**[b] | **−140.0** | | −119.0 | 15.0 |
| [c] | | | −121.4 | 13.2 |

[a]Results from runs biased with the pPBS stem (134 178 8).
[b]LDI results obtained in runs biased with the LAI-type stem (90 260 8).
[c]LDI results obtained in runs biased with the MAL-type stem (102 264 9), in HXB2R/LAI coordinates. Boldfaces indicate the best fit results, while the underlined values show the better fit of the BMH and LDI conformers in the cases where an alternate branched or a hybrid structure had the lowest free energy.

**Table 2.** Free energies of major conformations predicted by MPGAfold for the HIV-1 and SIV strains listed in 618 nt long 5′ domains

| | BMH (U5–AUG) (kcal/mol) | B(LD), B(Alt.) and Hybrids (kcal/mol) | LDI (kcal/mol) | LDI versus BMH free energy difference (%) |
|---|---|---|---|---|
| **LAI**[a] | −177.0 | **−182.8** | −179.1 | 1.2 |
| **HXB2R**[a] | −179.8 | **−184.6** | −182.7 | 1.6 |
| **MN** | −185.8 | −186.4 | **−187.9** | 1.1 |
| **NDK** | −176.0 | | −174.6 | 0.8 |
| **ELI** | −180.3 | | **−183.4** | 1.7 |
| **MAL**[a] | −171.3 | **−176.4** | −176.0 | 2.7 |
| **CM240**[a] | −179.6 | **−182.6** | −178.5 | 0.6 |
| **U455**[a] | −167.2 | **−175.4** | −164.1 | 1.9 |
| **IBNG*** | −185.1 | −182.3 | −172.9 | 6.6 |
| **93IN101**[b] | −183.8 | **−190.8** | −183.3 | 0.3 |
| **99ZACM9**[b] | −193.8 | **−195.4** | −186.1 | 4.0 |
| **ANT70** | **−217.8** | −216.9 | None | N/A |
| **CPZGAB**[b] | **−209.0** | −190.3 | −189.0 | 9.6 |
| [c] | | | −192.5 | 7.9 |

[a]Results from runs biased with the pPBS stem (134 178 8).
[b]LDI results obtained in runs biased with the LAI-type stem (90 260 8).
[c]LDI results obtained in runs biased with the MAL-type stem (102 264 9), in HXB2R/LAI coordinates. Boldfaces indicate the best fit results, while the underlined values show the better fit of the BMH and LDI conformers in the cases where an alternate branched or a hybrid structure had the lowest free energy.

conformations and a greater structural variability among them than the subtype B and D sequences.

The 25 nt long sequence insertion close to the end of the PBS in CM240 was predicted to form a self-contained motif. In the IBNG folds, the 17 nt sequence insertion at the same position disrupted the usual PBS substructure topology in the best fit branched structures, while in the LDI conformers the extra sequence tended to refold into a slightly modified, single M-loop PBS substructure. With the stems PAS and pPBS designated as 'sticky' MPGAfold predicted the BMH and LDI conformers with a constant PBS structure. The extra nucleotides in IBNG appeared to have an indirect impact, together with mutations preceding the SL2 motif, on the R–Gag motif, predicted only in the low fitness LDI structures.

In the 386 nt domain folds, CM240 was the only strain in this group undergoing a full transition from the BMH to the LDI. In the IBNG (379 nt) folds, the B(U5–AUG) conformers dominated solution spaces in all populations owing to their much better fitness than that of the LDI structures. The U455 (364 nt) folds predicted LDIs in a quarter of the low population runs, while the high-population runs were strongly dominated by the branched conformers, most of them with the U5–AUG stem alternative (also occluding the *gag* start codon). The most likely cause is the mutation C331 to A, which shortens the U5–AUG stem by 2 bp and lowers fitness of the B(U5–AUG) structures.

The 636 nt CM240 folds predicted LDIs in a third of the low population runs, the B(U5–AUG) type conformations in up to a third of the medium population runs, while the branched structures with several U5–AUG motif alternatives were in the majority of higher population runs. The 629 nt IBNG and 614 nt U455 folding results followed patterns established in their shorter domain folds.

### Subtype C (99ZACM9, 93IN101)

The 368 nt domain folds predicted only two distinct branched conformations for the subtype C strains. The B(U5–AUG) type was prominent in lower and medium population runs. The best fit branched structures contained long-distance interactions between the U5–AUG's 5′ side and the R–Gag's 3′ side nucleotides, thus mixing the typical BMH and LDI motifs. Some of the 93IN101 strain folding pathways included the LAI-type LDI conformers (but without the R–Gag stem) as transitional structures. Both subtype C sequences, however, with the LAI-type LDI and R–Gag stems designated as 'sticky,' were able to form the LDI conformers only slightly less fit than the best B(U5–AUG). It appears that the structural motifs of the best fit branched structures interfere with the development of the LDI conformers by out-competing the R–Gag stem in the early folding stages.

The 618 nt domain folds produced exclusively branched conformations. The B(U5–AUG) conformations were dominant in the low population solution spaces, and the U5–AUG stem alternatives dominated high-population runs. Linear conformers energetically close to the B(U5–AUG) structures were produced in biased runs with the LAI-type LDI stem designated as 'sticky.'

### Type O (ANT70)

The 385 nt long domain was predicted to fold exclusively into B(U5–AUG) with the best fitness in all the tested strains (Table 1). The long sequence insertion past the SL3 motif folds into a self-contained prominent stem–loop structure.

Mutations in this strain are too disruptive for it to form a viable LAI-type LDI. We also tested the MAL-type interactions in biased runs. However, disruptive mutations forced a 2 nt downstream shift in the 5′ side of this long-distance interaction relative to the full MAL-type LDI (Figure 4A). As a result the poly(A) SL was preserved, and, instead of an LDI structure, an alternative branched conformation with

the DIS region paired up was predicted. It was 8.1% less fit then the best BMH structure.

The 630 nt domain also folded exclusively into branched conformations, with the best and the dominant B(U5–AUG) conformation reaching the lowest free energy level of all structures predicted in this domain (Table 2).

### Subtype CPZ (SIV CPZGAB)

The 385 nt domain was predicted to fold exclusively into the branched conformations of the B(U5–AUG) type. In high-population runs the SL1 motif is replaced with a 5′-shifted and slightly better fit SL1c branch. The prominent sequence insertion following the SL3 motif forms a stem–loop structure nearly identical to that in HIV-1 ANT70.

Based on the sequence analysis, we have designated three stems, which could produce LAI-type and MAL-type LDIs and folded the entire domain in biased runs. The resulting linear structures were much less fit than the BMH conformers (Figure 4A and Table 1).

The 635 nt solution spaces retain the characteristics described for the 385 nt domain, adding to them strongly conserved motifs between the positions 385 and 635.

Biased folds with the 'sticky' LDI stems (LAI-type and MAL-type) yielded linear conformations without the R–Gag motif, both much less fit than the BMH conformers. Adding 'sticky stems' to bias runs towards R–Gag motif resulted in linear structure with even lower fitness (Figure 4A and Table 2). This poor fitness of the LDI conformers explains why they never make it to the final structure status in the unbiased runs.

### Sequential folding

We employed the sequential folding capability of the MPGAfold to model structure maturation of an elongating RNA sequence, thus approximately emulating folding during transcription. All the 618 nt sequences discussed above were folded in the sequential mode with populations varying from 4K to 32K.

The striking difference in the results of the elongating sequence folds, in comparison with their full domain length conformations, is the exclusively branched topology of the predicted structures. In the sequential folds the evolving structures do not transition through the LDI type conformations. This is the case even in the strains for which the LDI structures are predicted to be less fit than their BMH conformations in full domain length folds (Table 2). Thus it appears that in sequential folding the elongating sequences prefer local interactions in the poly(A) SL and SL1 (DIS) motifs and resist long-distance rearrangements of them. On the other hand, long-distance interactions disrupting the U5–AUG are predicted. The B(LD) conformations are present in a minority of runs for subtype B sequences, with the B(U5–AUG) conformation as the strongly dominant type. Strain CM240 folds are almost evenly split between the B(U5–AUG) and B(LD) conformers, U455 folds into B(LD) with a minority of alternative branched conformations and no B(U5–AUG)s, and the 93IN101 strain results transition from the B(U5–AUG) to alternative branched conformations as the folding population increases. All remaining sequences (ELI, NDK, MAL, IBNG, CPZGAB, 99ZACM9 and ANT70) were predicted to fold exclusively into the B(U5–AUG) type conformations.

### Consensus probability matrix algorithm results

The consensus probability matrix algorithm was applied to 155 HIV-1 non-redundant sequences, as explained in Materials and Methods. The system assigns scores that are higher than the threshold used for base pair prediction to stems SL1 and poly(A) SL, which support the BMH conformations, as well as to both stems supporting the LDI conformations, (90 260 5) and the alternative (102 264 8). These two LDI structure stems are subsets of the LAI-type and MAL-type key long-distance interactions predicted by the genetic algorithm. Full results of this algorithm are illustrated in Figures 2 and 3, where the stems supported by this algorithm are denoted by 'boxes' over the MPGAfold predicted secondary structures. As can be seen, motifs defining the BMH and LDI conformations are predicted by both methods. Specific scores assigned to different base pairs among the 'boxed' stems range from 0.1 to 1.0. A score of 0.1 is >4 SDs above the mean score ($Z$-score > 4) with respect to the elements of the consensus probability matrix. The scores represent, in a loose fashion, a lower bound on the probability of a base pair occurring. Not surprisingly the highest scores are assigned to the motifs shared by both the BMH and the LDI conformations (TAR, PBS/PAS, SL3). The motifs most impacted by the rearrangements in the two conformations 'split' the scores between them. For example, stems associated with the DIS sequence 'split' the score, in three ways, being involved in SL1 and two types of LDI interactions.

## DISCUSSION

In general, our computational experiments utilizing MPGAfold to find the folding pathways of multiple HIV-1 strains are consistent with and expand on the *in vitro* experimental results reported for the 5′-UTR structures of HIV-1 LAI and HXB2 and the pseudoknot interactions in HIV-1 MAL (1–3,23,31,44). They show very good agreement with the phylogenetic data on the BMH conformation (31). Our results show that functionally alternative conformational states; one branched (BMH), exposing the DIS sequence in a hairpin loop, the other partly linearized (LDI) through interactions of the DIS and poly(A) subdomains, can be found in all the tested HIV-1 and SIV subtypes, except for the O-type HIV-1 ANT70 strain. HIV-1 sequences of subtypes B and D readily form the BMH and LDI structures, while the subtypes A and C sequences favor BMH over LDI conformations. We found two alternative LDI pairing schemes retaining the same basic functionality, i.e. occluding the DIS sequence. One follows the *in vitro* long-distance pairing data for HIV-1 LAI, and the other, novel one, is exemplified by the interactions predicted for HIV-1 MAL. It also seems theoretically possible that occlusion of the DIS motif in alternative branched and hybrid structures provides yet another mechanism of dimerization control in strains IBNG, U455, MAL, and even the ANT70.

The MPGAfold exploration of the HIV-1 and SIV domains of varying lengths shows that the stability of the two observed conformations, the branched and the linear, depends on the

choice of the folding context, which is dramatically illustrated by the difference in the solution spaces of the 355 and 368 nt long domains of HIV-1 LAI.

A motif of interest reported for the LDI conformation of the HIV-1 LAI is the extended Ψ structure. Our folding results indicate that HIV-1 LAI appears to be the only examined sequence with any significant propensity to form it. What is more, interactions supported by our consensus probability matrix algorithm do not include the extended Ψ motif stems, further indicating its uniqueness to one, possibly a few strains.

The structural motifs shared by all results form a kind of structural scaffold. These motifs include the TAR and the extended PBS substructure, maintained in all folding domains (with variations in IBNG), the SL2 (SD) and SL3 (Ψ) stem–loop, present in the 368 and 618 nt domain, and the 3′ subdomain (past the position 368 in HXB2R) in the 618 nt and longer sequences (Figures 2A and 3). Thus the structural polymorphism depends exclusively on the localized rearrangements of the poly(A) and the SL1 (DIS) stem–loop motifs in the BMH and LDI conformations. The above has to be qualified in light of the proposed long-range pseudoknot (LRP) interaction between the 5′ poly(A) hairpin nucleotides and a sequence from the matrix coding region, shown for the HIV-1 MAL (44) and HXB2 strains (31). A similar intra or, potentially, inter-monomeric interaction was reported for MLV and is also possible in HIV-1 (50). Such interactions, compatible only with the BMH conformation (Figure 2A), could control translation. The currently available free energy rules do not allow for the calculation of the energetic contribution of this kind of pseudoknot. It is worthwhile mentioning, however, that the folds of the HIV-1 MAL in the 632 nt domain show that in the two dominant conformations, the matrix coding 3′ side nucleotides of the potential LRP are more open to form the proposed motif, without the need for any major structural rearrangements, than the secondary structure presented in Ref. (44). Similarly open to form the 3′ side of the LRP interactions are the nucleotides in the secondary structures of other folded strains, except for HIV-1 CM240 and SIV CPZGAB, in which they are mostly occluded.

Other MPGAfold predicted interactions in the 5′ fragment of the *gag* coding region remain in agreement with the majority of the published phylogenetic and biochemical probing data and include stems (363 539 6), (375 397 6), (399 484 9), (400 483 8), (447 457 3) and (501 526 6). Other stems presented in Ref. (31) fall downstream of our 618 nt folding domain, and may be part of long-distance interactions with the 3′ poly(A) sequence, based on our full HIV-1 genome folds (35).

Given that our consensus probability matrix algorithm, as well as the phylogenetic and biochemical probing data (18,23,31,32) support the SL1 (DIS) and SL2 (SD) stems, it is reasonable to assume that SL1s and SL2sh variations predicted by MPGAfold, better fit by only a 0.1 kcal/mol, represent the same structure within the margin of error of the free energy rules or indicate 'breathing' structures. A recent *ex vivo* study reported that SL1's 3′ side nucleotides C267-A271 exhibited weak reactivity suggesting structural breathing (32). These nucleotides include the 3′ side bulge loop in SL1s. It is also worthwhile noting that the lower stem (229 279 6) of the phylogenetically supported SL1 motif [DIS1 in (31)], corresponds to a stem predicted by MPGAfold in low fitness (immature) solutions, while the biochemically probed stem is present in the mature structures (Figure 2C). An NMR study of the SL1 structure of an *in vitro* transcript of HIV-1 HXB2c2 revealed the presence of the so-called loop A (Figure 2A) with a 3D structure very much like that of the Rev-binding loop in the Rev response element (51). This loop, which is not part of the phylogenetic and structure probing predictions, was present in the vast majority of the BMH conformers of HXB2R, MN and ELI folded with the genetic algorithm. Two other sequences from the B and D subtypes have mutations weakening the stem supporting the loop (236 282 3) or (237 281 2). Also worth noting is the fact that phylogenetic data support a large H-loop in the SL2 structure [SD in (31)], which accommodates both the SL2 and SL2sh H-loops.

The impact of sequence variations on the key LDI structure features is illustrated in Figure 4A and B. Interacting nucleotides are highlighted in the relevant fragments of aligned sequences. These sequence alignments indicate that, at least theoretically, the LDI interactions of the LAI type (yellow) should be possible in the MAL-type LDI group (purple) and vice versa. Indeed, a few LAI and HXB2R structures were predicted with MAL-type LDIs in the unbiased runs (2.2% less fit). Biased MPGAfold runs with these interactions designated as 'sticky stems,' highlighted in gray, showed them to be generally less fit than both (BMH and LDI) of the naturally folding states (5.3% less fit than the LDI for HIV-1 MN). The fact that two major stable alternatives exist, both of which occlude the DIS motif, suggests a functional importance of the LDI conformation, even though lack of compensatory base pair mutations does not give it absolute support.

The R–Gag long-distance duplex is another biochemically probed motif associated with linear conformations in the 368 nt domain (23). It was also prominent in our earlier study of HIV-1 LAI folds (without coaxial stacking energy calculations, EFN2) in this and longer domains (37). Figure 4B illustrates the ability to form it as a function of primary sequence. In addition, we highlight a 3 bp-long stem, which is also prominent in our predictions for the majority of the tested strains.

The comparisons illustrated in Figures 4A and 4B show that the cumulative impact of mutations in HIV-1 ANT70 on the two key areas associated with linear conformations is too disruptive for this topology to form. Computational experiments with the key linear motifs designated as 'sticky stems' show that the SIV CPZGAB can fold into LAI-type and MAL-type linear conformations with a poor fitness (Tables 1 and 2). Some external stabilizing factor would probably be required to maintain this topology.

The sequential folding results indicate strong propensity among all the tested sequences to fold into a branched topology, mostly of the B(U5–AUG) type, and show no LDI conformers. Thus they appear to be in better agreement with the recent *ex vivo* study of the genomic RNA conformation of HIV-1 HXB2NEO, which reported no evidence of LDI structures (32). The sequential folding results also appear to be consistent with the suggestion expressed in that paper that sequences may be kinetically trapped in the BMH conformation during transcription. However, contrary to the earlier reports suggesting that the BMH conformation must be thermodynamically less stable than the LDI (1–3,32), we have

shown in the non-sequential full domain folds that the branched conformations do not always correspond to energetically less fit states (Table 2). In fact, the optimal structure predicted by Mfold for the entire 9229 nt genomic RNA of HIV-1 LAI contains the BMH conformation in its 5′-UTR, while the top suboptimals contain the LDI conformations.

The key difference between the *ex vivo* results and our BMH conformation predictions is the retention of the U5–AUG interaction in the majority of sequences, in agreement with the *in vitro* data. On the other hand, the alternative branched conformations, which disrupt the U5–AUG stem, and which are prominent in non-sequential folds and present in the sequential folding results of subtype B sequences, indicate that the long U5–AUG structure may be relatively easily opened or rearranged. This is not surprising, since the *gag* start codon is present in the 3′ side of this interaction.

It is also worth noting that the reactivity probing results in the *ex vivo* study appear to be in conflict with the proposed LAI-type LDI interaction, (90 260 8) (yellow in Figure 4A), to a greater extent than with the novel MAL-type LDI stem, (102 264 9), supported by our results (purple in Figure 4A), or the subtype C LDI stem (88 264 6), (green in Figure 4A), a minor variation on the LAI-type motif. Excluding the end of stem mismatches between the *ex vivo* probing data for highly reactive and unreactive nucleotides in the BMH (without the U5–AUG) and MAL-type LDI secondary structure models, we arrive at 7.1% (9/126) error rate for the BMH and 11.1% (14/126) error rate for the MAL-type LDI. Whereas the LAI-type stem (90 260 8) includes 3 nts indicated as open in *ex vivo* probing ($Me_2SO_4$-modified), only one of them, A263, is involved in the MAL-type stem (102 264 9). The other two, A255 and A256, are in a loop and at the end of a short stem. Even though the MAL-type stem is rare in the predicted subtype B LDI conformers, strains LAI and HXB2R (closest to the strain used in *ex vivo* experiments) included it in some LDI conformers, which were only 1% less fit than the BMH structures. In the case of the subtype C LDI stem (88 264 6), it is harder to make direct comparisons with the experimental data, as the A263 is mutated to U and no other conflicts appear (Figure 4A).

As we noted above, a fairly prominent alternative to the dominant B(U5–AUG) conformation in the 618 nt domain is the B(LD). The 5′ and 3′ side nucleotides of its (107 446 7) stem are strongly conserved, which excludes verification via compensatory base mutations. We have to note that no biochemical probing data have been reported for such a conformation, and that the 3′ side of the (107 446 7) stem overlaps the 3′ side of the proposed LRP interaction (44). We therefore report our findings only as a hypothesis worth examining experimentally.

## CONCLUSIONS

In conclusion, our folding results extend to the major subtypes of HIV-1 the experimental results reported for only a few strains of HIV-1. Most importantly, MPGAfold predicts two types of LDI conformers. One is exemplified by the previously reported LAI-type interaction, and the other, novel alternative, is exemplified by the MAL long-distance interactions. These findings are bolstered by the data from our consensus probability matrix algorithm, which shows support for them in a larger set of HIV-1 strains (155 sequences). Our computational experiments also stress the impact of the selected folding domains and sequence variations on the ability and the propensity of different strains to fold into the LDI conformation. Our folding data explicitly and strongly support the existence of the structural polymorphism in the tested strains of HIV-1 subtypes B and D, shown previously *in vitro* only for the subtype B strains HIV-1 LAI and HXB2. It is worthwhile noting here that the sequences of these two subtypes, which show the strongest propensity to fold into the BMH and the LDI conformations in the genetic algorithm runs constitute nearly 76% of all the HIV-1 sequences collected around the world (http://www.hiv.lanl.gov/). Strains from the subtypes A1 (U455), 01_AE (CM240) and 02_AG (IBNG) exhibit not only the branched and linear topologies, but also the energetically best alternative branched states, in which the DIS sequence is occluded, while the poly(A) SL is preserved. Since the alternative branched structures and branched-linear hybrids are only marginally better fit than the better of the main two conformation types (which varies with strain and domain), they may not, in the presence of some external factors play any role *in vitro* or *in vivo*. Alternatively, they may indicate that subtype A strains have two ways of occluding the DIS sequence and disabling its function. Again, the presence of external stabilizing factors may strengthen the naturally weak topologies they share with the other strains. The SIV CPZGAB readily folds into the branched conformation, but its linear topology's fitness is poor and the energy difference between the two states is much greater than for the above HIV-1 strains. Finally, the type O strain ANT70 seems incapable of folding into the LDI conformation having too many mutations disrupting the key long-distance interactions in this topology, but it can form low fitness alternative branched conformations occluding the DIS.

Results of sequential folds indicate that the dynamics of folding as the sequence is elongated (as in transcription) strongly favors BMH conformations, which may reflect the reported *ex vivo* results. At the same time they indicate that some structural rearrangements are possible as the sequences elongate. They do not exclude, therefore, the possibility that the LDI conformers, predicted in the non-sequential, full-length domain folds and identified *in vitro*, exist transitionally in cells, or are missed by the specific *in situ* probing method owing to interference caused by cellular factors. The MAL-type form of the LDI interactions may also be compatible with the *ex vivo* probing data. We hope that the multiple predicted structures of the two LDI and the alternative conformations types, all of which support the existence of structures occluding the DIS site, will prompt further experimental studies aimed at clarifying the existence of and the role that this theoretically very likely structure may play *in vivo*.

## ACKNOWLEDGEMENTS

We wish to thank Dr Alan Rein of NCI-Frederick for helpful comments and the staff of the Advanced Biomedical Computing Center (ABCC) at NCI-Frederick for their assistance. This research has been funded in part with federal funds from the National Cancer Institute, National Institutes

of Health, under contract N01-C0-12400 and was also supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. Funding to pay the Open Access publication charges for this article was provided by NCI-Frederick.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Huthoff,H. and Berkhout,B. (2001) Two alternating structures of the HIV-1 leader RNA. *RNA*, **7**, 153–157.
2. Huthoff,H. and Berkhout,B. (2002) Multiple secondary structure rearrangements during HIV-1 RNA dimerization. *Biochemistry*, **41**, 10439–10445.
3. Berkhout,B., Ooms,M., Beerens,N., Huthoff,H., Southern,E. and Verhoef,K. (2002) *In Vitro* evidence that the untranslated leader of the HIV-1 genome is an RNA checkpoint that regulates multiple functions through conformational changes. *J. Biol. Chem.*, **277**, 19967–19975.
4. Berkhout,B., Silverman,R.H. and Jeang,K.T. (1989) Tat *trans*-activates the human immunodeficiency virus through a nascent RNA target. *Cell*, **59**, 273–282.
5. Klasens,B.I., Das,A.T. and Berkhout,B. (1998) Inhibition of polyadenylation by stable RNA secondary structure. *Nucleic Acids Res.*, **26**, 1870–1876.
6. Klasens,B.I., Thiesen,M., Virtanen,A. and Berkhout,B. (1999) The ability of the HIV-1 AAUAAA signal to bind polyadenylation factors is controlled by local RNA structure. *Nucleic Acids Res.*, **27**, 446–454.
7. Berkhout,B., Vastenhouw,N., Klaens,B. and Huthoff,H. (2001) Structural features in the HIV-1 repeat region facilitate strand transfer during reverse transcription. *RNA*, **7**, 1097–1114.
8. Beerens,N., Groot,F. and Berkhout,B. (2001) Initiation of HIV-1 reverse transcription is regulated by a primer activation signal. *J. Biol. Chem.*, **276**, 31247–31256.
9. Beerens,N. and Berkhout,B. (2002) The tRNA primer activation signal in the HIV-1 genome is important for initiation and processive elongation of reverse transcription. *J. Virol.*, **76**, 2329–2339.
10. Skripkin,E., Paillart,J.C., Marquet,R., Ehresmann,B. and Ehresmann,C. (1994) Identification of the primary site of the human immunodeficiency virus type 1 RNA dimerization *in vitro*. *Proc. Natl Acad. Sci. USA*, **91**, 4945–4949.
11. Laughrea,M. and Jette,L. (1994) A 19-nucleotide sequence upstream of the 5′ major splice donor is part of the dimerization domain of human immunodeficiency virus 1 genomic RNA. *Biochemistry*, **33**, 13464–13474.
12. Jossinet,F., Paillart,J.-C., Westhof,E., Hermann,T., Skripkin,E., Lodmell,J.S., Ehresmann,C., Ehresmann,B. and Marquet,R. (1999) Dimerization of HIV-1 genomic RNA of subtypes A and B: RNA loop structure and magnesium binding. *RNA*, **5**, 1222–1234.
13. Takahashi,K.-I., Baba,S., Chattopadhyay,P., Koyanagi,Y., Yamamoto,N., Takaku,H. and Kawai,G. (2000) Structural requirement for the two-step dimerization of human immunodeficiency virus type 1 genome. *RNA*, **6**, 96–102.
14. Lodmell,J.S., Ehresmann,C., Ehresmann,B. and Marquet,R. (2000) Convergence of natural and artificial evolution on an RNA loop–loop interaction: the HIV-1 dimerization initiation site. *RNA*, **6**, 1267–1276.
15. Shen,N., Jette,L., Wainberg,M.A. and Laughrea,M. (2001) Role of stem B, loop B, and nucleotides next to the primer binding site and the kissing-loop domain in human immunodeficiency virus type 1 replication and genomic-RNA Dimerization. *J. Virol.*, **75**, 10543–10549.
16. Ennifar,E., Walter,P., Ehresmann,B., Ehresmann,C. and Dumas,P. (2001) Crystal structures of coaxially stacked kissing complexes of the HIV-1 RNA dimerization initiation site. *Nature Struct. Biol.*, **8**, 1064–1068.
17. Pattabiraman,N., Martinez,H.M. and Shapiro,B.A. (2002) Molecular modeling and dynamics studiues of HIV-1 kissing loop structures. *J. Biomol. Struct. Dyn.*, **20**, 397–412.
18. Paillart,J.C., Shehu-Xhilaga,M., Marquet,R. and Mak,J. (2004) Dimerization of retroviral RNA genomes: an inseparable pair. *Nature Rev. Microbiol.*, **2**, 461–472.
19. Clever,J.L., Sassetti,C. and Parslow,T.G. (1995) RNA secondary structure and binding sites for *gag* gene products in the 5′ packaging signal of human immunodeficiency virus type 1. *J. Virol.*, **69**, 2010–2109.
20. De Guzman,R.N., Rong Wu,Z., Stalling,C.C., Pappalardo,L., Boer,P.N. and Summers,M.F. (1998) Structure of the HIV-1 nucleocapsid protein bound to the SL3 Ψ-RNA recognition element. *Science*, **279**, 384–388.
21. Zeffman,A., Hassard,S., Varani,G. and Lever,A. (2000) The major HIV-1 packaging signal is an extended bulged stem loop whose structure is altered on interaction with the *gag* polyprotein. *J. Mol. Biol.*, **297**, 877–893.
22. Amarasinghe,G.K., De Guzman,R.N., Turner,R.B. and Summers,M.F. (2000) NMR structure of stem-loop SL2 of the HIV-1 Ψ RNA packaging signal reveals a novel A-U-A base-triple platform. *J. Mol. Biol.*, **299**, 145–156.
23. Abbink,T.E.M. and Berkhout,B. (2003) A novel long distance base-pairing interactions in human immunodeficiency virus type 1 RNA occludes the gag start codon. *J. Biol. Chem.*, **278**, 11601–11611.
24. Shapiro,B.A. and Navetta,J. (1994) A massively parallel genetic algorithm for RNA secondary structure prediction. *J. Supercomput.*, **8**, 195–207.
25. Shapiro,B.A. and Wu,J.-C. (1996) An annealing mutation operator in the genetic algorithm for RNA folding. *Comput. Appl. Biosci.*, **12**, 171–180.
26. Shapiro,B.A. and Wu,J.-C. (1997) Predicting RNA H-type pseudoknots with the massively parallel genetic algorithm. *Comput. Appl. Biosci.*, **13**, 459–471.
27. Wu,J-C. and Shapiro,B.A. (1999) A Boltzmann filter improves the prediction of RNA folding pathways in a massively parallel genetic algorithm. *J. Biomol. Struct. Dyn.*, **1763**, 581–595.
28. Shapiro,B.A., Wu,J-C., Bengali,D. and Potts,M. (2001) The massively parallel genetic algorithm for RNA folding: MIMD implementation and population variation. *Bioinformatics*, **17**, 137–148.
29. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
30. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
31. Damgaard,C.K., Andersen,E.S., Knudsen,B., Gorodkin,J. and Kjems,J. (2004) RNA Interactions in the 5′ Region of the HIV-1 Genome. *J. Mol. Biol.*, **336**, 369–379.
32. Paillart,J-C., Detenhoffer,M., Yu,X-F., Ehresmann,C., Ehresmann,B. and Marquet,R. (2004) First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J. Biol. Chem.*, **279**, 48397–48403.
33. Holland,J.H. (1975) *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.
34. Shapiro,B.A., Bengali,D., Kasprzak,W. and Wu,J.-C. (2001) RNA folding pathway functional intermediates: their prediction and analysis. *J. Mol. Biol.*, **312**, 27–44.
35. Gee,A., Kasprzak,W. and Shapiro,B.A. (2005) Structural differentiation of the HIV-1 Poly(A) Signals. *J. Biomol. Struct. Dynamics*.In press.
36. Shapiro,B.A., Bengali,D., Kasprzak,W. and Wu,J-C. (2001) Computational insights into RNA folding pathways: getting from here to there. In *Proceedings of the Atlantic Symposium on Computational Biology and Genome Systems and Technology*. pp. 10–13.
37. Kasprzak,W. and Shapiro,B.A. (2002) Structural dependencies of the HIV-1 dimer initiation site as determined by the massively parallel genetic algorithm. In *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences*, CSREA Press, Vol. I, 48–54.
38. Shapiro,B.A. and Kasprzak,W. (1996) STRUCTURELAB: a heterogeneous bioinformatics system for RNA structure analysis. *J. Mol. Graph.*, **14**, 194–205.
39. Kasprzak,W. and Shapiro,B.A. (1999) 'Stem Trace': an interactive visual tool for comparative RNA structure analysis. *Bioinformatics*, **15**, 16–31.
40. Bindewald,E. and Shapiro,B.A. (2005) Secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*. In Press.

41. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
42. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
43. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
44. Paillart,J-C., Skripkin,E., Ehresmann,B., Ehresmann,C. and Marquet,R. (2002) *In vitro* evidence for a long range pseudoknot in the 5′-untranslated and matrix coding regions of HIV-1 genomic RNA. *J. Biol. Chem.*, **277**, 5995–6004.
45. Clever,J.L., Wong,M.L. and Parslow,T.G. (1996) Requirements for kissing-loop-mediated dimerization of human immunodeficiency virus RNA. *J. Virol.*, **70**, 5902–5908.
46. Huthoff,H., Bugala,K., Barciszewski,J. and Berkhout,B. (2003) On the importance of the primer activation signal for initaiation of $tRNA^{lys3}$-primed reverse transcription of the HIV-1 RNA genome. *Nucleic Acids Res.*, **31**, 5186–5194.
47. Beerens,N. and Berkhout,B. (2002) Switching the *in vitro* tRNA Usage of HIV-1 by simultaneous adaptation of the PBS and PAS. *RNA*, **8**, 357–369.
48. Isel,C., Westhof,E., Massire,C., Le Grice,S., Ehresmann,B., Ehresmann,C. and Marquet,R. (1999) Structural basis for the specificity of the initiation of HIV-1 reverse transcription. *EMBO J.*, **18**, 1038–1048.
49. Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.
50. Hibbert,C. and Rein,A. (2005) Preliminary physical mapping of RNA–RNA linkages in the genomic RNA of moloney murine leukemia virus. *J. Virol.*, **79**, 8142–8148.
51. Gallego,J., Greatorex,J., Zhang,H., Yang,B., Arunachalam,S., Fang,J., Seamons,J., Lea,S., Pomerantz,R.J. and Lever,A.M.L. (2003) Rev binds specifically to a purine loop in the SL1 region of the HIV-1 leader RNA. *J. Biol. Chem.*, **278**, 40385–40391.