Bayegan *et al. BMC Bioinformatics* (2016) 17:530
DOI 10.1186/s12859-016-1389-7

BMC Bioinformatics

RESEARCH ARTICLE Open Access

New tools to analyze overlapping coding regions

Amir H. Bayegan, Juan Antonio Garcia-Martin and Peter Clote*

Abstract

Background: Retroviruses transcribe messenger RNA for the overlapping Gag and Gag-Pol polyproteins, by using a programmed -1 ribosomal frameshift which requires a slippery sequence and an immediate downstream stem-loop secondary structure, together called frameshift stimulating signal (FSS). It follows that the molecular evolution of this genomic region of HIV-1 is highly constrained, since the retroviral genome must contain a slippery sequence (sequence constraint), code appropriate peptides in reading frames 0 and 1 (coding requirements), and form a thermodynamically stable stem-loop secondary structure (structure requirement).

Results: We describe a unique computational tool, `RNAsampleCDS`, designed to compute the number of RNA sequences that code two (or more) peptides p, q in overlapping reading frames, that are identical (or have BLOSUM/PAM similarity that exceeds a user-specified value) to the input peptides p, q. `RNAsampleCDS` then samples a user-specified number of messenger RNAs that code such peptides; alternatively, `RNAsampleCDS` can exactly compute the position-specific scoring matrix and codon usage bias for all such RNA sequences. Our software allows the user to stipulate overlapping coding requirements for all 6 possible reading frames simultaneously, even allowing IUPAC constraints on RNA sequences and fixing GC-content.
We generalize the notion of *codon preference index* (CPI) to overlapping reading frames, and use `RNAsampleCDS` to generate control sequences required in the computation of CPI. Moreover, by applying `RNAsampleCDS`, we are able to quantify the extent to which the overlapping coding requirement in HIV-1 [resp. HCV] contribute to the formation of the stem-loop [resp. double stem-loop] secondary structure known as the frameshift stimulating signal. Using our software, we confirm that certain experimentally determined deleterious HCV mutations occur in positions for which our software `RNAsampleCDS` and `RNAiFold` both indicate a single possible nucleotide. We generalize the notion of codon preference index (CPI) to overlapping coding regions, and use `RNAsampleCDS` to generate control sequences required in the computation of CPI for the Gag-Pol overlapping coding region of HIV-1. These applications show that `RNAsampleCDS` constitutes a unique tool in the software arsenal now available to evolutionary biologists.

Conclusion: Source code for the programs and additional data are available at http://bioinformatics.bc.edu/clotelab/RNAsampleCDS/.

Keywords: Overlapping coding region, Ribosomal frameshift, Frameshift stimulating signal, HIV-1, HCV

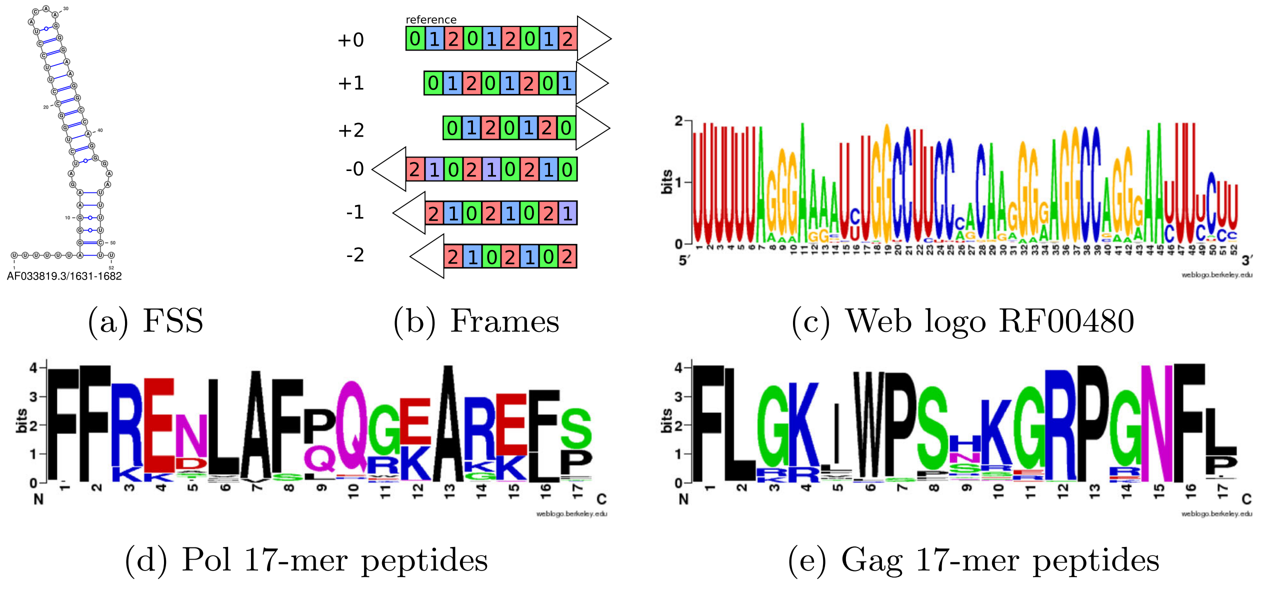
Background

Programmed ribosomal frameshift (PRF) is a curious phenonenon, exploited especially by certain viruses, in order to translate two different protein products from the same messenger RNA. The frameshift is caused by particular sequence and structural elements of the mRNA which sometimes cause the ribosome to slip and readjust the reading frame, thus allowing viruses to pack more information into their genomes. Since the ratio of the protein products coded in overlapping reading frames depends on the PRF efficiency, which has been finely tuned by evolution, any chemical that can modify this efficiency could prove to be a useful anti-viral agent. Though partcularly important for the life cycle of certain viruses, such as HIV-1 and HCV, programmed ribosomal frameshift can be found in all kingdoms of life [1].

*Correspondence: clote@bc.edu
Biology Department, Boston College, 140 Commonwealth Avenue, 02467 Chestnut Hill MA, USA

In HIV-1, Pol is obtained from a fused Gag-Pol polyprotein via a programmed -1 ribosomal frameshift, which naturally occurs with a frequency of 5–10%; moreover, an increase of ribosomal frameshift frequency is associated with a decrease in viral infectivity [2]. The -1 ribosomal frameshift is caused by two *cis*-acting RNA elements, together known as *frameshift stimulating signal* (FSS): (1) a heptameric *slippery sequence* (U UUU UUA), where the Gag reading frame is indicated, and (2) a downstream stem-loop secondary structure, often with either internal loop or right bulge. The FSS from HIV-1 genome (AF033819.3/1631-1682) is shown in Fig. 1a, where the minimum free energy (MFE) secondary structure was determined by `RNAfold` from *Vienna RNA Package* 2.1.9 [3]. The Pol reading frame is -1 with respect to the Gag reading frame, or equivalently, the Gag reading frame is +1 with respect to the Pol reading frame (convention adopted throughout this paper) – Fig. 1b depicts the six reading frames considered in this paper. While the entire Gag-Pol overlap region in HIV-1 AF033819.3 is from position 1631 to 1838 (Pr55 Gag polyprotein is coded at AF033819.3/336-1838), the 17-mer Pol [resp. Gag] peptide coded in the 52 nt FSS region 1631-1682 is FFREDLAFLQGKAREFS [resp. FLGKIWPSYKGRPGNFL]. Moreover, we found the secondary structure from Fig. 1a to be the most common MFE structure for 52 nt segments of the Pol coding region, which begin by UUUUUUA, taken from the HIV Sequence Database in Los Alamos National Laboratory (LANL) available at www.hiv.lanl.gov. Due to its importance, a collection of 145 HIV-1 ribosomal frameshift elements is given in the family RF00480 in Rfam 12.0 [4]. Figure 1c displays the sequence logo obtained from the 145 sequences in the seed alignment of RF00480, while Fig. 1d and e respectively display the sequence logos for the 17-mer Pol and Gag peptides coded in RF00480.

For decades, research in evolutionary biology has focused mostly on protein-coding regions, leading to the development of sophisticated computational tools, such as `PAML` [5] and `HYPHY` [6], to compute the ratio dN/dS of non-synonomous mutation rate dN to the synonomous mutation rate dS [7–9]. Pedersen and Jenson [10] extended the codon substitution model of Goldman and Yang [8] to overlapping genes in a site-specific manner, where evolutionary constraints of both genes are taken into account. However, estimation of



(a) FSS (b) Frames (c) Web logo RF00480

(d) Pol 17-mer peptides (e) Gag 17-mer peptides

Fig. 1 a Minimum free energy (MFE) structure of the initial 52-nt Gag-Pol overlapping reading frame in positions 1631-1682 of the HIV-1 complete genome (GenBank AF033819.3). This frameshift stimulating signal (FSS) contains the initial slippery sequence heptamer, given by `U UUU UUA` in the Gag reading frame, as well as the displayed stem-loop secondary structure, which together promote a programmed -1 frameshift `UUU UUU A` in the Pol reading frame. **b** Depiction of all 6 possible reading frames – `RNAsampleCDS` samples RNA sequences that code in all possible reading frames, allowing IUPAC sequence constraints **c** Sequence logo for 145 RNA HIV-1 frameshift signal sequences from the RF00480 seed alignment from Rfam 12.0 [4]. **d** Sequence logo for the Pol peptide coded by 138 RNA HIV-1 frameshift signal sequences from the RF00480 seed alignment from Rfam 12.0; Pol peptide translated from nucleotide positions 1-51. **e** Sequence logo for the Gag peptide coded by 138 RNA HIV-1 frameshift signal sequences from the RF00480 seed alignment from Rfam 12.0; Gag peptide translated from nucleotide positions 2-52. Since some sequences from RF00480 contained IUPAC codes for uncertain data, the data were disambiguated–for instance, the code B (not A) was disambiguated by randomly assigning either C,G or U with probability 1/3. Seven sequences were removed from the seed alignment of 145 RNAs due to gaps in the alignment, and another five sequences were removed since either the Pol or Gag peptide contained a stop codon–resulting in 133 sequences for nucleotide analysis. Peptide sequence logos for the 138 Pol and Gag peptides were created using `WebLogo` [26] evolutionary parameters in this model required computationally expensive Markov chain Monte Carlo simulations. By dropping the condition of site specificity, Sabath et al. [11] were able to apply a maximum likelihood method to estimate parameters in a more efficient manner. The resulting tool has been used to predict functionality of overlapping reading frames [12]. An evolutionary model has been developed for coding regions with conserved RNA secondary structures [13] as well. This approach was used to determine the effects of structural elements on nucleotide substitution in hepatitis C virus.

Several methods have been developed to sample sequences using an evolutionary model derived from a given phylogeny [14–16]. To the best of our knowledge, however, there is no previously published method for sampling sequences in overlapping coding regions. The program SISSI [16] incorporates a user-defined system of dependencies between the nucleotides; however, it is not possible using SISSI to sample sequences that code in overlapping reading frames, since SISSI requires that any position in an RNA sequence must belong to a single codon. Moreover, SISSI does not allow sequence and structural dependencies to be specified simultaneously. Our work in this paper is orthogonal to the foregoing computational models and tools of mathematical evolution theory and does not rely on phylogeny information. In full generality, the new software RNAsampleCDS supports the following. For each reading frame $r \in \{+0, +1, +2, -0, -1, -2\}$ illustrated in Fig. 1b, let p_r be a length n sequence in the 22-letter alphabet consisting of IUPAC codes for each amino acid, together with symbol X (any residue) and O (any residue or STOP). RNAsampleCDS computes the number of RNA sequences $a_0, \ldots, a_{3n+2}$ which simultaneously code protein p'_r in reading frame r, such that either p'_r is identical to p_r, or (optionally) whose BLOSUM/PAM similarity to p_r exceeds a user-specified value. (Throughout the article, we say that the peptide p is *BLOSUM[PAM]* θ *similar* to another peptide p', if each amino acid of p has BLOSUM[PAM resp.] similarity of *at least* θ with the corresponding amino acid of p'.) RNAsampleCDS can then compute the PSSM and codon usage frequency for such proteins, as well as sample a user-specified number of such sequences. RNAsampleCDS runs in linear time and space, although if GC-content is optionally controlled, then time and space requirements are quadratic. For expository reasons, we describe the algorithms for only two proteinsp, q respectively in reading frame 0 and 1; however, our code is general as just described – see the Additional file 1 for details on the general algorithm. Using RNAsampleCDS, we undertake a preliminary analysis of the Gag-Pol overlapping reading frame in human immunodeficiency virus (HIV-1) and of the triple overlapping reading frame of hepatitis C virus (HCV).

Methods

RNAsampleCDS

Let $p = p_1, \ldots, p_n$ and $q = q_1, \ldots, q_n$ be two peptides of equal length. In this section, we are interested in the following questions.

1. Which sequences $a_0, \ldots, a_{3n}$ of messenger RNA translate the peptide p in reading frame 0 and also translate the peptide q in reading frame $+1$?
2. Which sequences $a_0, \ldots, a_{3n}$ of messenger RNA translate peptides $p' = p'_1, \ldots, p'_n$ in reading frame 0 and peptide $q' = q'_1, \ldots, q'_n$ in reading frame $+1$, where the BLOSUM/PAM similarity of p with p' and q with q' is greater than or equal to a user-specified threshold θ?
3. What is the profile, or PSSM, for the collection of mRNAs from (1) and (2)?
4. What is the total number of sequences satisfying (1) and (2), and how can we sample sequences $a_0, \ldots, a_{3n}$ of messenger RNA in an unbiased manner, in order to satisfy either (1) or (2)?

By developing software to sample mRNA sequences that code user-specified proteins in different reading frames, we can then analyze the samples with other tools to provide an estimate of the probability of satisfying a given property of interest, hence give approximate answers for questions like the following: What is the expected stem size in the minimum free energy (MFE) structure of RNAs that translate peptides p', q' in reading frames 0,1, where the BLOSUM/PAM similarity of p, p' and of q, q' is at least a user-specified threshold value of θ? As we show, it is not difficult to see that questions (1,2) are easily answered using breadth first search (BFS); however, for large values of n, it can happen that BFS in not practical, since the number of messenger RNAs can be of size exponential in n. For that reason, we describe a novel dynamic programming (DP) algorithm to answer questions (3) and (4).

We first need a few definitions. If xyz is a trinucleotide, then let $tr(xyz)$ denote the amino acid whose codon is xyz in the genetic code; i.e. $tr(xyz)$ is the amino acid translated from codon xyz, unless xyz is a stop codon. If $xyzu$ is a tetranucleotide, then let $tr_0(xyzu)$ [resp. $tr_1(xyzu)$] denote the amino acid whose codon is xyz [resp. yzu]; i.e. $tr_0(xyzu) = tr(xyz)$ and $tr_1(xyzu) = tr(yzu)$. For each $k = 1, \ldots, n$, define the collection L_k of 4-tuples $s = s_0, s_1, s_2, s_3$ such that $tr_0(s) = tr(s_0, s_1, s_2) = p_k$ and $tr_1(s) = tr(s_1, s_2, s_3) = q_k$. Define two 4-tuples $s = s_0s_1s_2s_3$ and $t = t_0t_1t_2t_3$ to be *compatible* if $s_3 = t_0$ – i.e. the tail of s equals the head of t. Note that if 4-tuples s, t are compatible, then the *merge* $s_0, s_1, s_2, t_0, t_1, t_2, t_3$ of s, t has the property that amino acids are translated by each of the four codons $s_0s_1s_2$, $s_1s_2s_3$, $t_0t_1t_2$, and $t_1t_2t_3$.

Algorithm 1 (BFS computation of sequences that code in reading frames 0 and 1)

Define the tree T by induction on depth as follows.

- **Base case:** The root of T is $\emptyset$; the children of the root are those 4-tuples s, such that $tr_0(s) = p_1$, $tr_1(s) = q_1$. The depth of the root is 0, and the depth of each child of the root is 1.
- **Inductive case:** If s is a 4-tuple in T of depth k, then the children of s are those 4-tuples t, such that $s_3 = t_0$ (compatibility requirement) and $tr_0(t) = p_{k+1}$, $tr_1(t) = q_{k+1}$ (coding requirement). The depth of each child of s is $k + 1$.

Suppose that $\sigma_1, \sigma_2, \ldots, \sigma_k$ is a *path* from root to level k; i.e. $\sigma_1, \sigma_2, \ldots, \sigma_k$ is a sequence of 4-tuples belonging to T, where for each $i = 1, \ldots, k$, the level of σ_i is equal to i, and for each $i = 1, \ldots, k-1$, σ_{i+1} is a child of σ_i. Define the *merge* of $\sigma_1, \sigma_2, \ldots, \sigma_k$ to be the RNA sequence $a_0, a_1, \ldots, a_{3k}$, where $\sigma_1 = a_0a_1a_2a_3$, $\sigma_2 = a_3a_4a_5a_6$, $\sigma_3 = a_6a_7a_8a_9$, $\ldots$, $\sigma_k = a_{3(k-1)}a_{3k-2}a_{3k-1}a_{3k}$. By induction, it is easy to establish that in this case $tr_0(\sigma_i) = p_i$, $tr_1(\sigma_i) = q_i$ for each $i = 1, \ldots, k$. An easy application of breadth first search then allows one to generate the collection of level n nodes of T. It follows that the answer to question (1) is the set of RNAs obtained by merging the paths from root to level n nodes of T.

Using our implementation of the BFS approach in Algorithm 1, we can easily determine that there are exactly 32 52-nt RNAs that translate the 17-residue Pol peptide FFREDLAFLQGKAREFS in reading frame 0, and the 17-residue Gag peptide FLGKIWPSYKGRPGNFL in reading frame +1. These 17-mer peptides are those which constitute the beginning of the Gag-Pol overlap in the HIV-1 genome (nucleotides 1631-1682 in GenBank AF033819.3). The entire Gag-Pol overlap region is from 1631-1835, whereby the 68-mer Pol [resp. Gag] peptide is coded in the region 1631-1834 [resp. 1632-1835 with a Gag STOP codon at 1836-1838]. Our implementation of the BFS method returns exactly 256 205-nt RNAs that code the Pol [resp. Gag] 68-mers from HIV-1 (GenBank AF033819.3).

Figure 2 displays the centroid secondary structure, `RNAalifold` [17] consensus structure, and the corresponding mountain plot for the alignment of all 256 205-nt RNA sequences that code the Pol and Gag 68-mer peptides from HIV-1 (Pol 1631-1835, Gag 1632-1836 in GenBank AF033819.3), *not* necessarily containing the slippery sequence UUUUUUA.

Further analysis (data not shown) indicates that there is considerable variation in the low energy structures of RNAs that exactly code the same 68-mer Pol and Gag peptides as those coded by AF033819.3/1631-1836.

Question (2) is an obvious generalization of (1), and is easy to answer by generalizing the collection L_k of 4-tuples $s = s_0, s_1, s_2, s_3$ such that $tr_0(s) = tr(s_0, s_1, s_2) = p'_k$ and $tr_1(s) = tr(s_1, s_2, s_3) = q'_k$, where the BLOSUM/PAM similarity of p_k, p'_k and of q_k, q'_k is at least a user-specified threshold θ.

It is more interesting to turn to question (3), which requires a different strategy, since the number of RNAs returned by BFS may be exponentially large. Indeed, if RNA sequences are required to code peptides p [resp. q] whose amino acids have BLOSUM62 similarity of at least θ to those of the Pol [resp. Gag] 17-mer peptide coded in reading frame 0 [resp. 1] in AF033819.3/1631-1682, then the number of solution sequences is 256 ($\theta = 4$), 34,560 ($\theta = 3$), 90,596,966,400 ($\theta = 2$), 2.14285987145e+32 ($\theta = 1$), 3.61150917928e+56 ($\theta = 0$), 1.20555937201e+81 ($\theta = -1$), 1.17643153215e+106 ($\theta = -2$)! To address question (3), define the forward and backwards partition function ZF, ZB as follows.

- **Forward partition function:** For integer $k = 1, \ldots, n$ and nucleotide $ch \in \{A, C, G, U\}$, define $ZF(k, ch)$ to be the number of RNAs $\mathbf{a} = a_0, \ldots, a_{3k}$ such that a_{3k} is the nucleotide ch, and $\mathbf{a}$ translates the peptide $p_1, \ldots, p_k$ resp. $q_1, \ldots, q_k$ in reading frame 0 resp. 1; i.e. $tr_0(\mathbf{a}) = p_1, \ldots, p_k$ and $tr_1(\mathbf{a}) = q_1, \ldots, q_k$.
- **Backward partition function:** For integer $k = 1, \ldots, n$ and nucleotide $ch \in \{A, C, G, U\}$, define $ZB(k, ch)$ to be the number of RNAs $\mathbf{a} = a_{3k}, a_{3k+1}, \ldots, a_{3n}$ such that a_{3k} is the nucleotide ch, and $\mathbf{a}$ translates the peptide $p_k, \ldots, p_n$ resp. $q_k, \ldots, q_n$ in reading frame 0 resp. 1; i.e. $tr_0(\mathbf{a}) = p_k, \ldots, p_n$ and $tr_1(\mathbf{a}) = q_k, \ldots, q_n$.

By dynamic programming, it is straightforward to compute the forward and backward partition functions in linear time and space, as done in Algorithm 2.

Recall that the *indicator function* $I[$ boolean condition$]$ returns the value 1 if the boolean condition within its scope is true, and otherwise the value returned is 0.

By appropriately redefining L_k, the recursions of Algorithm 2 can easily be modified to instead count the number of sequences coding $p'_1, \ldots, p'_n$ in reading frame 0 and $q'_1, \ldots, q'_n$ in reading frame +1, such that for each i, the BLOSUM/PAM similarity of p_i, p'_i and of q_i, q'_i exceeds a user-specified threshold θ, or for which the Kyte-Doolittle hydrobicity of p_i, p'_i and q_i, q'_i differ by at most a user-specified upper bound, etc. The same remark applies to *all* algorithms of this section, although for reasons of space, we do not explicitly mention such extensions. Nevertheless, such extensions are supported by the software `RNAsampleCDS`.

By refining the definition of forward and backward partition function, Algorithms 1 and 2 can be modified

Algorithm 2 (DP partition function for sequences that code in reading frames 0 and 1)

Given n-mer peptides p_0, q_0, for $k = 1, \ldots, n$ and $ch \in \{A, C, G, U\}$ define the *forward partition function* $ZF(k, ch)$ inductively as follows:

- Case 1: $k = 1$
  $ZF(k, ch) = \sum_{s_0 s_1 s_2 s_3 \in L_k} I[s_3 = ch]$
- Case 2: $k = 2, \ldots, n$
  $ZF(k, ch) = \sum_{s_0 s_1 s_2 s_3 \in L_k} I[s_3 = ch] \cdot ZF(k-1, s_0)$

For $k = n, \ldots, 1$ and $ch \in \{A, C, G, U\}$, define the *backward partition function* ZB inductively as follows:

- Case 1: $k = n$
  $ZB(k, ch) = \sum_{s_0 s_1 s_2 s_3 \in L_k} I[s_0 = ch]$
- Case 2: $k = n-1, \ldots, 1$
  $ZB(k, ch) = \sum_{s_0 s_1 s_2 s_3 \in L_k} I[s_0 = ch] \cdot ZB(k+1, s_3)$

Note the use of the boolean valued indicator function $I[\ldots]$, which has the value 1 if the expression within the brackets is true, and otherwise has the value 0. It follows that

$$Z = \sum_{ch \in \{A,C,G,U\}} ZF(n, ch) = \sum_{ch \in \{A,C,G,U\}} ZB(1, ch)$$

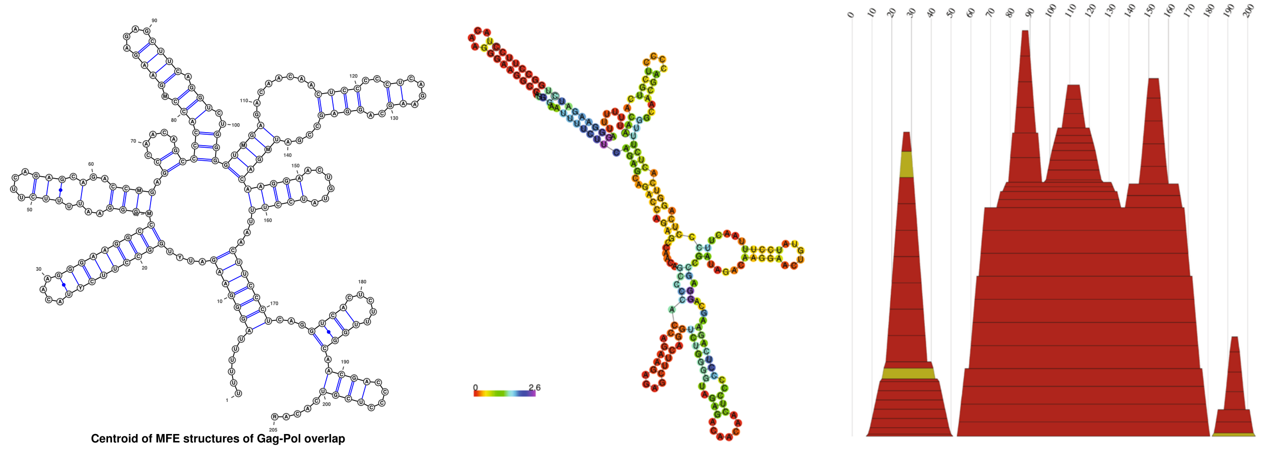
is the total number of RNA sequences that translate p in reading frame 0 and q in reading frame +1.

to keep track of the GC-content, albeit at an overhead for the space required. For an arbitrary RNA sequence $\mathbf{a}$, let $gccount(\mathbf{a})$ denote the number of Gs or Cs occurring in $\mathbf{a}$.

- **Forward partition function accounting for GC-content:** For integer $k = 1, \ldots, n$ and nucleotide $ch \in \{A, C, G, U\}$, define $ZF_{GC}(k, x, ch)$ to be the number of RNAs $\mathbf{a} = a_0, \ldots, a_{3k}$ such that a_{3k} is the nucleotide ch, $gccount(\mathbf{a}) = x$, and $\mathbf{a}$ translates the peptide $p_1, \ldots, p_k$ resp. $q_1, \ldots, q_k$ in reading frame 0 resp. 1; i.e. $tr_0(\mathbf{a}) = p_1, \ldots, p_k$ and $tr_1(\mathbf{a}) = q_1, \ldots, q_k$.
- **Backward partition function accounting for GC-content:** For integer $k = 1, \ldots, n$ and nucleotide $ch \in \{A, C, G, U\}$, define $ZB_{GC}(k, x, ch)$ to be the number of RNAs $\mathbf{a} = a_{3k}, a_{3k+1}, \ldots, a_{3n}$ such that a_{3k} is the nucleotide ch, $gccount(\mathbf{a}) = x$, and $\mathbf{a}$ translates the peptide $p_k, \ldots, p_n$ resp. $q_k, \ldots, q_n$ in reading frame 0 resp. 1; i.e. $tr_0(\mathbf{a}) = p_k, \ldots, p_n$ and $tr_1(\mathbf{a}) = q_k, \ldots, q_n$.

Though not explicitly described, *all* the following algorithms (PSSM computation and sampling) can be modified to account for GC-content. Our program, `RNAsampleCDS`, implements all the algorithms described in this section, including versions that account for GC-content. Moreover, our program supports any *two or more* overlapping coding regions in any of the 6 reading frames – i.e. reading frame 0,1,2 on the plus-strand and 0,1,2 on the minus-strand, as shown in Fig. 1b.

Note that an easy modification of the above algorithm allows one to compute the total number of RNAs of



(a) Centroid (b) Consensus (c) Mountain plot

Fig. 2 **a** The centroid secondary structure, **b** `RNAalifold` consensus structure, and **c** the corresponding mountain plot for the alignment of all 256 205-nt RNA sequences that code the Pol and Gag 68-mer peptides from HIV-1 (Pol 1631-1835, Gag 1632-1836 in GenBank AF033819.3) length $3n + 1$, which code n-mer peptides p [resp. q] in reading frames 0 [resp. 1], i.e. for which neither reading frame contains a stop codon. This modification is later used to compute the probability that a random RNA of length $3n + 1$ will code in both reading frames 0 and 1. Algorithm 3 applies Algorithm 2 in order to compute the exact value of the position specific scoring matrix (PSSM).

Algorithm 3 (PSSM computation of sequences that code in reading frames 0 and 1)

Given n-mer peptides p_0, q_0, for $i = 0, \ldots, 3n$ and $ch \in \{A, C, G, U\}$, define the profile or PSSM of nucleotides at positions $0, \ldots, 3n$ as follows:

- CASE 1: $i = 0$. Then $PSSM(i, ch)$ equals $\sum_{s \in L_1} I[s_0 = ch] \cdot ZB(1, ch)/Z$
- CASE 2: $i \equiv 0 \bmod 3$. Then $PSSM(i, ch)$ equals $ZF(i/3, ch) \cdot ZB(i/3, ch)/Z$
- CASE 3: $i \equiv 1 \bmod 3$. Then $PSSM(i, ch)$ equals $\sum\limits_{s \in L_{\lfloor i/3 \rfloor}} I[s_1 = ch] \cdot ZF(\lfloor i/3 \rfloor, s_0) \cdot ZB(\lceil i/3 \rceil, s_3)/Z$
- CASE 4: $i \equiv 2 \bmod 3$. Then $PSSM(i, ch)$ equals $\sum\limits_{s \in L_{\lfloor i/3 \rfloor}} I[s_2 = ch] \cdot ZF(\lfloor i/3 \rfloor, s_0) \cdot ZB(\lceil i/3 \rceil, s_3)/Z$

The recursions can be easily modified, if the RNA sequence is instead required to code $p'_1, \ldots, p'_n$ in reading frame 0 and $q'_1, \ldots, q'_n$ in reading frame +1, such that for each i, the BLOSUM/PAM similarity of p_i, p'_i and of q_i, q'_i exceeds a user-specified threshold θ. This answers question (3). The resulting DP program is very fast, since the run time is linear in n, while the BFS program has run time that is exponential in n.

Given a gapless alignment S of mRNA sequences of length $3n + 1$, each of which codes a protein in reading frame 0 and 1, define the *positional codon frequency* $PCF(w, k, r)$ to be the number of occurrences of w in the kth codon position in reading frame $r \in \{0, 1\}$ of a sequence in S. If S is the collection of all mRNAs that code proteins p, q respectively in reading frame 0,1, which are identical to (or alternatively have BLOSUM/PAM similarity that exceeds threshold θ), then the positional codon frequency can be defined from the partition functions ZF, ZB as done in Algorithm 4.

Next, in order to sample RNA sequences that code peptides $p = p_1, \ldots, p_n$ resp. $q = q_1, \ldots, q_n$ in reading frames 0 resp. 1, we construct the sampled sequence from last to first character, each time ensuring that $ZF(k, ch) > 0$ where ch is the leading character of the current sample $a_{3k-1}, a_{3k}, \ldots, a_{3n}$. This is described in done in Algorithm 5, where we recall that L_k denotes the collection of 4-tuples $s = s_0, s_1, s_2, s_3$ such that $tr_0(s) = tr(s_0, s_1, s_2) = p'_k$ and $tr_1(s) = tr(s_1, s_2, s_3) = q'_k$, and the BLOSUM/PAM similarity of p_k, p'_k and of q_k, q'_k is at least a user-specified threshold θ.

Algorithm 4 (Positional codon frequency)

Given n-mer peptides p_0, q_0, integer $k = 1, \ldots, n$, codon $w = w_0 w_1 w_2 \in (\{A, C, G, U\})^3$, and reading frame $r \in \{0, 1\}$, the positional codon frequency $PCF(w, k, r)$ for the set of all mRNAs that code p_0, q_0 respectively in reading frame 0, 1 can be computed as follows.

- CASE 1: $r = 0$. Then $PCF(w, k, 0)$ equals $ZF(k-1, w_0) \cdot \sum_{ch \in \{A,C,G,U\}} ZB(k, ch)$.
- CASE 2: $r = 1$. Then $PCF(w, k, 1)$ equals $\sum_{ch \in \{A,C,G,U\}} ZF(k-1, ch) \cdot ZB(k, w_2)$

Algorithm 5 (Uniform sampling of RNAs that code in reading frames 0 and 1)

```
k = n //initialize to the common length of peptides p,q
rna = "" //initialize to empty sequence
ch = random nucleotide in { A,C,G,U } satisfying $ZF(k,ch) > 0$
while k>0
    choose random 4-tuple $s = s_0, s_1, s_2, s_3$ such that $s_3 = ch$
    $rna = s_1, s_2, s_3 + rna$
    ch = $s_0$
    k = k-1
$rna = ch + rna$ //prepend the remaining initial nucleotide
```

It is straightforward to modify the previous algorithm to sample in a *weighted* fashion as done in Algorithm 6. First, recall that L_k denotes the collection of 4-tuples $s = s_0, s_1, s_2, s_3$ such that $tr_0(s) = tr(s_0, s_1, s_2) = p'_k$ and $tr_1(s) = tr(s_1, s_2, s_3) = q'_k$, and the BLOSUM/PAM similarity of p_k, p'_k and of q_k, q'_k is at least a user-specified threshold θ. Additionally, if $ch \in \{A, C, G, U\}$ then let $L_{k,ch}$ denote the set of tuples t in L_k, whose last element t_3 is ch.

Our implementation of the algorithms described in this section allows the user to stipulate *sequence constraints* using any IUPAC nucleotide codes, for instance, designating the first 7 nucleotides to be the slippery sequence UUUUUUA, or to consist of an alternation of purines and pyrimidines RYRYRYR, etc.

Finally, we note that all the previous algorithms in this section can be extended to handle *multiple* overlapping reading frames in all six reading frames, i.e. reading frames +0,+1,+2 on the plus strand and reading frames -0,-1,-2

Algorithm 6 (Weighted sampling of RNAs that code in reading frames 0 and 1)

```
k = n //initialize to the common length of peptides p,q
rna = "" //initialize to empty sequence
a = ZF(k,A); c = ZF(k,C); g = ZF(k,G); u = ZF(k,U);
z = a+c+g+u
a = a/z; c = c/z; g = g/z; u = u/z
select ch from A,C,G,U with prob a,c,g,u using roulette wheel
while k>0
    sum = 0; r = random(0,1) · ZF(k-1,ch)
    for t in $L_{k-1,ch}$ //note that $t = t_0 t_1 t_2 t_3$ and $t_3$ = ch
        sum = sum + $ZF(k-1, t_0)$
        if r < sum
            rna = t + rna; ch = $t_0$; k = k-1; break
return rna
```

on the minus strand, as illustrated in Fig. 1b. For instance, in order to compute the forward partition function for reading frames 0,1,2 we define $ZF(k, ch1, ch2)$ to be the number of RNA sequences $\mathbf{a}$ of length $3k+2$ whose last two nucleotides are $ch1, ch2$, such that $tr_0(\mathbf{a}) = p_1, \ldots, p_k$, $tr_1(\mathbf{a}) = q_1, \ldots, q_k$, $tr_2(\mathbf{a}) = r_1, \ldots, r_k$, for user-specified peptides $\mathbf{p} = p_1, \ldots, p_n$, $\mathbf{q} = q_1, \ldots, q_n$, $\mathbf{r} = r_1, \ldots, r_n$. Now we define L_k to be the set of 5-tuples $s = s_0, \ldots, s_4$ such that $s_0 s_1 s_2$ codes residue p_k, $s_1 s_2 s_3$ codes residue q_k, and $s_2 s_3 s_4$ codes residue r_k. The definition of the generalization of the forward partition function $ZF(k, ch1, ch2)$, analogous to that defined in Algorithm 2, is as follows:

- Case 1: $k = 1$. Then $ZF(k, ch1, ch2)$ equals
$$\sum_{s_0 s_1 s_2 s_3 s_4 \in L_k} I[\, s_3 = ch1, s_4 = ch2]$$
- Case 2: $k = 2, \ldots, n2, \ldots, n$. Then $ZF(k, ch1, ch2)$ equals
$$\sum_{s_0 s_1 s_2 s_3 s_4 \in L_k} I[\, s_3 = ch1, s_4 = ch2] \cdot ZF(k-1, s_0, s_1)$$

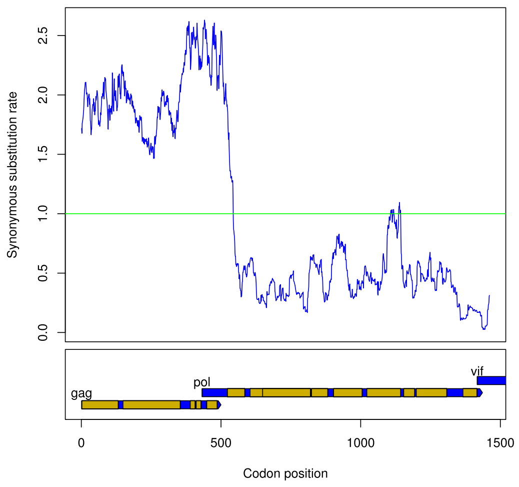
Our publicly available code RNAsampleCDS supports all the above described variants of Algorithms 1-6 with possible IUPAC sequence constraints, stipulation of GC-content, and where the user may stipulate that particular peptides are coded in any or all of the six reading frames displayed in Fig. 1b. See Additional file 1 for details of how we determine the run time estimate of $\approx 0.58831373 \cdot L + 0.00550239 \cdot N$ to generate compute the partition function and generate N samples of RNA sequences of length L that code any peptide in each of the six possible reading frames.

Results and Discussion

In this section, we use RNAsampleCDS to study novel aspects of human immunideficiency virus HIV-1 and hepatitis C virus HCV, that cannot be determined using methods other than those described in this paper.

HIV-1 programmed -1 frameshift

Analysis of HIV-1 overlap: Since HIV-1 and other retroviruses have a -1 ribosomal frameshift in the initial portion of the Gag-Pol overlap, this can be detected by the software FRESCo [18], which predicts regions of excess synonymous constraint in short, deep alignments. Figure 3a displays the dN/dS ratio we obtained for HIV-1 AF033819.3 with respect to the Gag reading frame, when aligned with other HIV-1 genomes from the Los Alamos HIV Database – see also Additional file 1: Figure S1. This figure indicates that there is *positive selection* in the Gag region before the Gag-Pol overlap. In contrast, starting with the beginning of the Gag-Pol overlap (nucleotide 1631), there is *purifying selection*; i.e. Fig. 3a suggests the presence of an important signal starting around position 1631. Figure 3b displays the dN/dS ratio of the 52 nt Gag-Pol overlap region, for both the Gag and Pol reading frames, using the method of [11] which computes a rate matrix for overlapping reading frames – an aspect ignored by PAML and other software. Since Sabath's program computes dN/dS from a pairwise alignment, which is wholly inappropriate for the short 52 nt sequences considered here, we modified the approach by first producing multiple alignments of 52 nt Gag-Pol overlap regions, and then computed the number of (observed) synonomous and nonsynonomous mutations within the Gag [resp. Pol] reading frame, taking account for all codon pairs in the same column. We then modified Sabath's Matlab program to compute dN/dS by maximum likelihood using counts obtained from the multiple alignments. The multiple alignments considered in Fig. 3b are from Rfam family RF00480 and from 52 nt RNA sequences generated by the programs RNAsampleCDS and RNAiFold 2.0. RNAsampleCDS generates 52 nt sequences, that translate peptides in the Gag [resp. Pol] reading frame, each of whose amino acids has BLOSUM62 similarity of either 0 or 1 to the corresponding amino acids in the Gag [resp. Pol] reading frame of the peptides translated by the 52 nt HIV-1 overlap region of AF033819.3/1631-1682. RNAiFold 2.0 generates 52 nt sequences, that not only satisfy the same coding requirements as RNAsampleCDS, but which also fold into the minimum free energy secondary structure shown in Fig. 1a. In each case, RNAiFold 2.0 generates *all* sequences that satisfy both the coding and structure



(a) FRESCo

	dN/dS Pol	*dN/dS Gag*	*branch len t*	*transition/transversion*	*num seq*
Lanl-B0-CDS	1.03	0.24	2.93	1.54	100,000
Lanl-B0-ifold	0.66	0.38	1.23	2.78	42,534
Lanl-B1-CDS	0.18	0.13	1.13	0.97	100,000
Lanl-B1-ifold	0.16	0.17	0.63	4.73	1,196
Ofori-B0-CDS	1.13	0.33	2.45	1.94	100,000
Ofori-B0-ifold	0.66	0.44	1.08	3.09	26,640
Ofori-B1-CDS	0.25	0.16	0.89	1.48	100,000
Ofori-B1-ifold	0.25	0.19	0.45	15.15	276

(b) dN/dS

Fig. 3 a Output from the program `FRESCo` [18], when run on the Gag reading frame of an alignment of 200 sequences from the LANL HIV-1 database using 50 nt windows. Note the precipitous drop in dN/dS value at the beginning of Gag-Pol overlap region. **b** Values of dN/dS, branch length, and transition/transversion rate (see [8] for definitions) for the 52 nt Gag-Pol overlap regions within a multiple alignment from Rfam family RF00480 as well as from 52 nt RNA sequences generated by the programs `RNAsampleCDS` and `RNAiFold`. These programs generate sequences that code peptides, each of whose amino acids has BLOSUM62 similarity of either 0 or 1 to the corresponding amino acids in the Gag [resp. Pol] reading frame of the peptide translated by the 52 nt HIV-1 overlap region of [2] or by GenBank accession code AF033819.3/1631-1681. The program `RNAsampleCDS` ensures only coding requirements, while `RNAiFold` ensures both coding requirements and that the 52 nt RNAs fold into the minimum free energy structure of the Gag-Pol overlap region of HIV-1 from [2] and GenBank accession code AF033819.3/1631-1682

requirements, their number being substantially less than the 100,000 sequences generated by `RNAsampleCDS`. Note the presence of purifying selection for the Gag reading frame, as indicated by dN/dS values less than 1.

Codon preference index: In this section, we generalize the notion of *codon preference index* (CPI) [19] to the context of overlapping coding regions. For RNA sequence $\mathbf{a} = a_0, \ldots, a_{3n}$ which codes n-mer peptides in reading frames 0, 1, for codon $w \in (\{A, C, G, U\})^3$ and reading frame $r \in \{0, 1\}$, define $f_{(w,\mathbf{a},r)}$ to be the number of occurrences of codon w in reading frame r of $\mathbf{a}$, and for amino acid AA, define $f_{(AA,\mathbf{a},r)}$ to be the number of occurrences of codons coding AA in reading frame r of $\mathbf{a}$. Define the *observed codon preference* in $\mathbf{a}$ by $p_{obs}(w, \mathbf{a}) = \sum_{r=0}^{1} f_{(w,\mathbf{a},r)} / \sum_{r=0}^{1} f_{(AA,\mathbf{a},r)}$. If $S$is a set of mRNAs of length $3n+1$, each of which codes n-mer peptides in both reading frames 0,1, then define the *observed codon preference* in S by $p_{obs}(w, S) = \sum_{r=0}^{1} \sum_{\mathbf{a}\in S} f_{(w,\mathbf{a},r)} / \sum_{r=0}^{1} \sum_{\mathbf{a}\in S} f_{(AA,\mathbf{a},r)}$. Note that $p_{obs}(w, S)$ is the *probability* that codon w will be used for amino acid AA in the collection S of overlapping coding sequences. Finally, define the *codon preference index* $I(w)$ of codon w in S by $I(w) = p_{obs}(w, S)/p_{obs}(w, S')$, where S' is a *control* set of mRNAs of length $3n + 1$.

With these notations, Fig. 4 depicts a heat map for the codon preference index $I(w)$, computed over 5,125 entire Gag-Pol overlap regions of average length 205 ± 10 (Gag and Pol peptide size $\approx$ 68) extracted from LANL HIV-1 database, each starting with the slippery sequence UUUUUUA and terminating with the last Gag codon; additionally the heat map includes Gag-only and Pol-only values for the same overlap region. For this figure, the control set S' is defined differently for each column $1 - 5$, although in all cases, each sequence in S' contains the initial slippery sequence UUUUUUA. For column 1 [resp. 2] S' is the set of all mRNAs that code proteins in the Gag [resp. Pol] reading frame that are coded by some sequence of S. For column 3, S' is the set of all mRNAs that code proteins p and q that are identical to proteins coded in the Gag and Pol reading frames of some sequence $\mathbf{a}$ of S. For column 4, S' is defined as in the case for column 3, except that 'identical to' is replaced by 'BLOSUM62 +1 similar to'. For column 5, S' is the set of all mRNAs that code proteins p and q that are BLOSUM62 +1 similar to proteins coded in the Gag and Pol reading frames of a sequence $\mathbf{a}$ of S, and whose GC-content lies in the range of GC-content of $\mathbf{a} \pm 5$. The heat map of Fig. 4 shows that for serine, $I(AGU, Gag) < I(AGU, Pol) < I(AGU, Gag/Pol) \approx 1$; for valine, $I(GUG, Gag) < 1 < I(GUU, Gag)$ but $I(GUG, Gag/Pol) > 1 > I(GUU, Gag/Pol)$; for proline, $I(CAU, Gag) < I(CAU, Pol) < I(CAU, Gag/Pol) \approx 1$, but when the control set is taken to be BLOSUM62 +1 similar peptides to Gag and Pol, then $I(CAU, Gag/Pol + 1) \gg 1$. See Additional file 1: Figures S2 and S3 and the

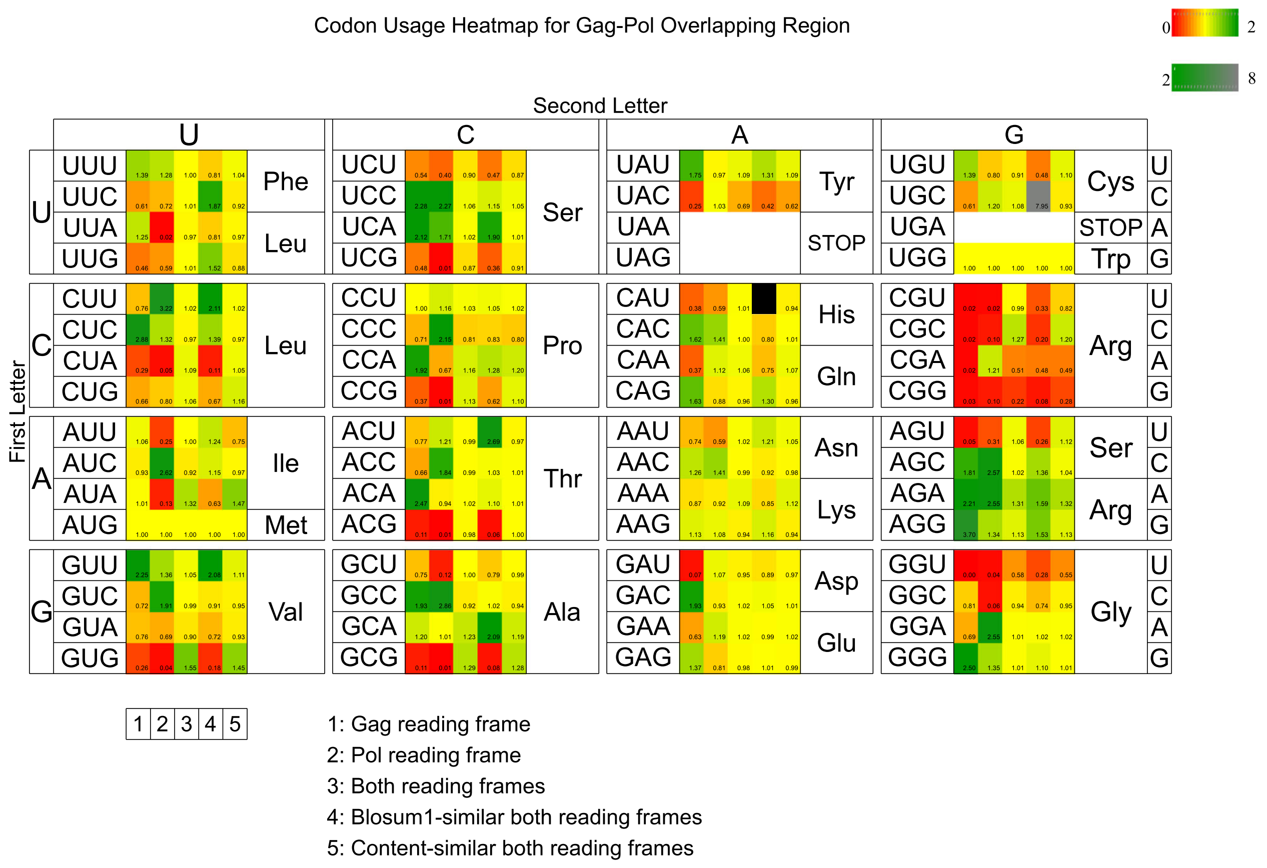


Fig. 4 Heat map of the *codon preference index* (CPI) for a collection of 5125 entire Gag-Pol overlap regions of average length 205 ± 10 extracted from LANL HIV-1 database. CPI values shown at bottom right of each square. See text for additional explanation

text from Additional file 1 for more detailed explanation. These figures show that the codon usage bias observed at the Gag-Pol junction is not due to natural selection [20] or to the underlying mutational bias, but rather imposed by the overlapping coding constraints.

Overlapping coding and stem-loop formation: Here we describe how to quantify the extent to which coding HIV-1 17-mer peptides in overlapping reading frames induces a stem-loop structure. In particular, we consider the following questions.

1. What is the probability that random RNA forms a stem-loop structure?
2. What is the probability that RNA forms a stem-loop structure, if it is required to code (any arbitrary) peptides in reading frames 0 and 1?
3. What is the probability that RNA forms a stem-loop structure, if it is required to code peptides in reading frames 0 and 1, which are *similar* to peptides coded in the HIV-1 frameshift stimulating signal (FSS)?
4. To what extent do HIV-1 coding requirements in the Pol-Gag overlap region alone induce stem-loop formation?
5. What is the (conditional) probability of coding peptides in reading frames 0 and 1 if the RNA forms a secondary structure similar to the FSS stem-loop structure of HIV-1?

To answer question 1, we generated 200,000 52-nt RNAs, where the first seven nucleotides constituted the slippery sequence UUUUUUA, and each nucleotide in position 8 through 52 was randomly selected with probability 0.25 for each of A,C,G,U. Using RNAshapes, cf. [21], we determined the Boltzmann probability that each RNA sequence has shape [ ] [22], i.e. $P($ [ ] $) = \sum_s \exp(-E(s)/RT)$, where the sum is taken over all *stem-loop* secondary structures, which may contain internal loops and bulges, but no multiloops or multiple stem-loops. Throughout the sequel of the paper, the probability that a given RNA sequence will form a *stem-loop* structure is identified with $P($ [ ] $)$. A finer analysis could consider type 1 shapes of the form _[_[ ]_] or _[[ ]_], corresponding to a stem loop with internal loop or right bulge, with left flanking unpaired region, but in this paper we consider only the type 5 stem loop shape [ ]. By

MFE stem-loop structure, we mean the stem-loop secondary structure which has the minimum free energy, taken over all stem-loop structures. Similarly, *stem-loop MFE* means the minimum free energy of all stem-loop structures. Note that the stem-loop MFE is not necessarily equal to the MFE, since it is possible that a structure having two or more external loops, or containing a multiloop, could have lower energy than that of any stem-loop structure. By uniformly sampling 200,000 52 nt RNAs with no coding requirements, we estimate an average probability of stem-loop formation of 60.7% with standard deviation of 36.2%, and average stem-loop MFE was −7.65 kcal/mol with standard deviation 3.42 kcal/mol – again, this is for 52 nt RNA with no constraints.

Before answering question 2, we first note that the conditional probability is 45.32% that a 52-nt RNA codes in both reading frames 0,1 assuming that it begins by the slippery heptamer UUUUUUA is 23.14%, and that the conditional probability that a 52-nt RNA codes in reading frame 1, given that it begins by the slippery heptamer UUUUUUA *and* that it already codes in reading frame 0 45.32% – i.e. $P(A|B,C) = 0.4532$, where event A is that a 52-nt RNA codes in reading frame 0, event B is that the 52-nt RNA contains slippery heptamer UUUUUUA, and event C is that reading frame 0 of the 52-nt RNA contains no stop codon. In contrast, the conditional probability that a 52-nt RNA codes in reading frame 0 assuming that it begins by the slippery heptamer UUUUUUA is 51.06%.

Indeed, using RNAsampleCDS, we determine that the number x_1 of 52-nt RNAs beginning by UUUUUUA and which code in both reading frames 0,1 is $2.86451 \cdot 10^{26}$. In contrast, the number x_2 of 52-nt RNAs beginning by UUUUUUA and which code in reading frame 0 is $x_2 = 16 \cdot 61^{14} \cdot 4 = 6.32117 \cdot 10^{26}$, since there are 16 codons that begin by A, a choice of 61 coding codons for the remaining 14 residues (since the first two residues must be FF and the third residue have a codon beginning by A), times 4 for the last nucleotide to ensure the RNA length is 52. The number x_3 of all 52-nt RNAs that begin by UUUUUUA is clearly $4^{45} = 1.23794 \cdot 10^{27}$. These computations justify the previous probabilities, and suggest the potential utility of RNAsampleCDS when speculating about molecular evolution.

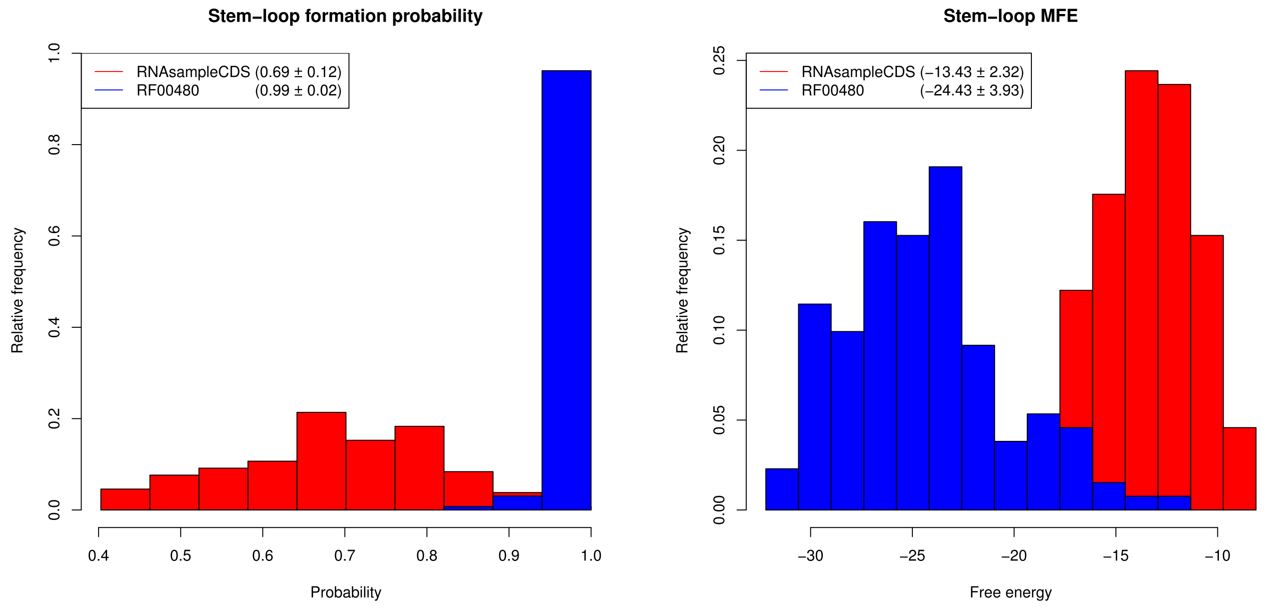
To answer question 2, we used RNAsampleCDS to generate 200,000 52-nt RNA sequences, each of which contains the slippery sequence UUUUUUA and codes 17-mer peptides in both reading frames 0 and 1. Executing RNAshapes as previously described yielded an average probability of stem-loop formation of 59.8% with standard deviation of 36.7%, and average stem-loop MFE of −8.06 kcal/mol with standard deviation 3.58 kcal/mol.

To answer question 3, we extracted 145 52-nt Pol-Gag overlapping FSS sequences in family RF00480 from the Rfam 12.0, of which 133 sequences remained after disambiguation and removal of sequences containing gaps or stop codons. For each of the 133 sequences, we generated 100,000 sequences using RNAsampleCDS, each of which begins by the same initial 7 nucleotides of the Rfam sequence constituting a slippery sequence (since most but not all RF00480 sequences begin with UUUUUUA), and which code peptides p [resp. q] having BLOSUM62 similarity of at least +1 with the corresponding amino acids of the 17-mer peptide coded by the Rfam sequence in frame 0 [resp. 1].

After removing two outliers (discussed shortly), we have the following statistics for the remaining 131 sequences from RF00480. Average probability of stem-loop formation for RF00480 is 99.3 ± 2.2%, and average stem-loop MFE is −24.43 ± 3.91 kcal/mol. For the collection of 100,000 sequences generated by RNAsampleCDS for each sequence from Rfam family RF00480, coding BLOSUM62 +1 similar peptides to those coded by the Rfam sequence, the average stem-loop formation probability is is 69 ± 12%, and average stem-loop MFE is −13.43 ± 2.32 kcal/mol. Figure 5a and b depict respectively the stem-loop formation probabilities and stem-loop minimum free energies. In contrast, a similar computational experiment using RNAsampleCDS shows that the average probability of stem-loop formation is 98.1% ± 8.1 if each sampled sequence is required to code *exactly* the same peptides as those from HIV-1 in RF00480. This answers question 4.

The previous analysis was performed for 131 Rfam sequences, obtained after removal of the sequences AF442567.1/1455-1506 and L11798.1/1290-1341, from the set of 133 Rfam sequences obtained from 145 sequences in RF00480, after disambiguation and removal of sequences containing gaps or stop codons. These two sequence were removed as outliers, since their stem-loop formation probabilities were respectively 53.1% and 55.5% – far removed from the average of 99.3 ± 2.2% of the remaining sequences. GenBank annotations indicate that AF442567.1 is highly G to A hypermutated with very many, mostly in-frame, stop codons throughout the genome, and that the Gag gene of L11798.1 has a premature termination at position residue 46.

Together, these results show that stem-loop formation is a consequence of the *precise* HIV-1 Gag and Pol 17-mer peptides, but not of BLOSUM62 +1 similar peptides. As well, stem-loop formation probability is not statistically different (T-test) between random sequences, sequences that have no stop codon in reading frame 0 or 1, and sequences that code peptides having BLOSUM62 similarity of at least +1 to HIV-1 peptides. To determine particular nucleotide positions in the 52-nt FSS that appear to be critical in stem-loop formation, we computed the position-dependent nucleotide frequency (PSSM), denoted by π_1, for 200,000 sequences generated by RNAsampleCDS that begin



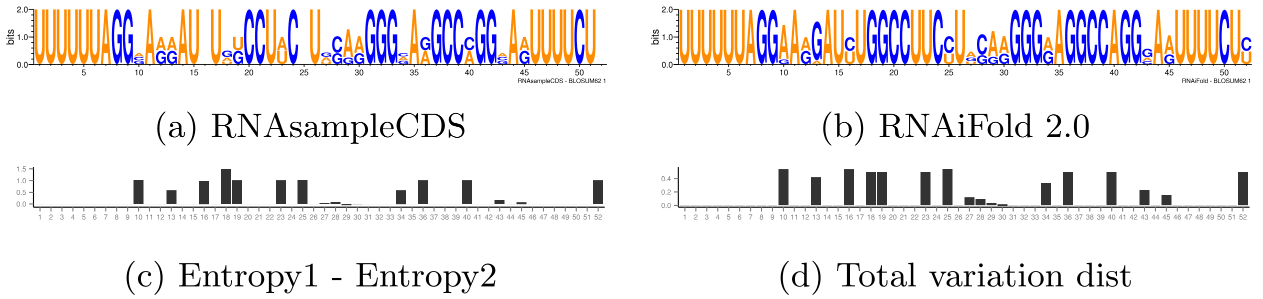
(a) Stem-loop formation probability (b) Stem-loop MFE

Fig. 5 For each of 131 52 nt frameshift stimulating signals (FSS) from family RF00480 from the Rfam 12.0, RNAsampleCDS generated 100,000 RNAs that have the same slippery sequence as the Rfam sequence, and code 17-mer peptides *p* [resp. *q*] in reading frame 0 [resp. 1] each of whose amino acids has BLOSUM62 similarity of at least +1 with the corresponding amino acid in the Pol [resp. Gag] peptide coded by the Rfam sequence. Stem-loop formation probability, *P*([ ]), and stem-loop minimum free energy (MFE) were computed by RNAshapes [21] with the command RNAshapes -q -m '[]'. **a** Average stem-loop formation probability for 100,000 sequences sampled from RNAsampleCDS for each RF00480 sequence (red); stem-loop formation probability of HIV-1 frameshift stimulating signals from RF00480 (*blue*). Overall mean RNAsampleCDS samples is 69% ± 12 (*red*), while that for the RF00480 sequences is 99.3 ± 2.2 (*blue*). **b** Average stem-loop MFE for 100,000 sequences sampled by RNAsampleCDS for each RF00480 sequence (red); stem-loop minimum free energy for HIV-1 frameshift stimulating signals from RF00480 (*blue*). Overall mean for RNAsampleCDS samples is 13.43 ± 2.32 kcal/mol (*red*), while that for RF00480 sequences is −24.43 ± 3.91 kcal/mol (*blue*)

by the slippery sequence UUUUUUA, and code peptides p [resp. q], each of whose amino acids has BLOSUM62 similarity greater than or equal to 1 with the corresponding amino acids of the Pol [resp. Gag] 17-mer peptides FFREDLAFPQGKAREFS [resp. FLGKIWPSHKGRPGNFL] coded in AF033819.3/1631-1682. Using RNAiFold 2.0, we also computed the PSSM, denoted by π_2, for all possible sequences that begin by slippery heptamer UUUUUUA, and fold into the MFE structure of AF033819.3/1629-1682 shown in Fig. 1a, and which code peptides that are BLOSUM62 +1 similar to the peptides coded by AF033819.3/1631-1682. We then computed the position-dependent total variation distance between π_1 and π_2, defined by $\delta(\pi_{1,i}, \pi_{2,i}) = 1/2 \cdot \sum_{x\in\{A,C,G,U\}} |\pi_{1,i}(x) - \pi_{2,i}(x)|$, where $\pi_{1,i}$ resp. $\pi_{2,i}$ denotes the mononucleotide frequency at position i of the PSSM for sequences generated by RNAsampleCDS resp. RNAiFold 2.0. With the exception of specific regions, the total variation distance is close to zero, thus pinpointing critical nucleotides necessary for stem-loop formation of the FSS. Figure 6a, b display the sequence logo for the PSSM π_1 and π_2, and Fig. 6c and d respectively depict the position-dependent entropy and total variation distance.

To answer question 5, we used RNAiFold 2.0 with target structure as depicted in Fig. 1a, in order to generate 200,000 52-nt RNA sequences, each containing the slippery sequence UUUUUUA and each folding into the target structure. We determined that 61.91% of these sequences have no stop codon in reading frames 0 or 1. The percentage of sequences that have no stop codon in reading frame 0 [resp. 1] alone is somewhat higher, with value 78.7% [resp. 79.59%]. We additionally determined that the average base pair distance between the MFE structure of the sampled sequences and the target FSS secondary structure is 2.04 and average ensemble defect is 3.58.

The probability of stem-loop formation for frameshift stimulating signal (FSS) regions of HIV-1 is close to 1, with average value of 99% ± 2 for RF00480 as shown in Fig. 5a. This value is much larger than that of random 52-nt RNAs (≈ 61%), or 52-nt RNA having no stop codons in reading frames 0 or 1 (≈ 60%), or even 52-nt RNA coding peptides in reading frames 0,1 with BLOSUM62 similarity



(a) RNAsampleCDS

(b) RNAiFold 2.0

(c) Entropy1 - Entropy2

(d) Total variation dist

Fig. 6 a Sequence logo from `RNAsampleCDS` for *all* 8,819,712 sequences that code peptides *p* [resp. *q*], each of whose amino acids has BLOSUM62 similarity ≥ +1 with the corresponding amino acids of the Pol [resp. Gag] 17-mer peptides FFREDLAFPQGKAREFS [resp. FLGKIWPSHKGRPGNFL] in AF033819.3/1631-1682. The PSSM is (exactly) computed by `RNAsampleCDS` with flag `-pssm`, and the logo plot was produced using `WebLogo` [26]. The average pairwise Hamming distance is 10.92 ± 4.32 (length-normalized value of 0.21 ± 0.083), when computed with a random sample of 1000, 5000, and 10,000. **b** Sequence logo for all 1196 sequences determined by `RNAiFold 2.0` to fold into the frameshift stimulating signal (FSS) given by the MFE structure from AF033819.3/1629-1682 and code peptides P,Q, each of whose BLOSUM62 similarity with the Gag,Pol peptides in the overlap region is greater than or equal to +1. The average pairwise Hamming distance is 5.80 ± 1.84 (length-normalized value of 0.11 ± 0.035). **c** The position-dependent entropy is defined by $H_i = -p_A \ln p_A - p_C \ln p_C - p_G \ln p_G - p_U \ln p_U$ for each nucleotide position $i = 1, \ldots, 52$. Subfigure (**c**) shows the position-dependent *difference* $H_i^a - H_i^b$ in entropies of (**a**) minus (**b**). **d** Position-dependent total variation distance $\delta(\pi_{1,i}, \pi_{2,i}) = 1/2 \cdot \sum_{x \in \{A,C,G,U\}} |\pi_{1,i}(x) - \pi_{2,i}(x)|$ in the 52 nt region of the Gag-Pol overlap in the HIV-1 genome (GenBank AF033819.3/1631-1682) that contains the frameshift stimulating signal (FSS). Here $\pi_{1,i}$ resp. $\pi_{2,i}$ is the mononucleotide frequency at position *i* of the PSSM in the left resp. right panel. If total variation distance is zero, then it is suggestive that the coding constraint automatically may already entail the FSS secondary structure constraint

of at least +1 to HIV-1 peptides (≈ 69%). It follows that coding BLOSUM62 +1 similar peptides to those of HIV-1 at most slightly induces stem-loop formation. Yet the probability that stem-loop structures do not have a stop codon in either reading frame 0 or 1 is only about 62%, without requiring that the peptides be similar to those of HIV-1. It follows that BLOSUM62 +1 similarity to HIV-1 peptides cannot induce the required stem-loop FSS structure, nor can the target FSS structure from Fig. 1a induce BLOSUM62 +1 similarity to HIV-1 peptides. We speculate that starting from a genomic region that codes a polyprotein similar to that of Gag, a series of pointwise mutations could slowly induce a stem-loop FSS structure and at the same time slowly create a Pol-like reading frame. Although speculative, it is possible to create an adaptive walk or Monte Carlo program to test the likelihood of this hypothesis, using intermediate sequences generated by `RNAsampleCDS` and `RNAiFold2.0`.

As shown in Fig. 6a, the average pairwise Hamming distance of sequences generated by `RNAsampleCDS` with the overlapping coding constraint and the slippery sequence constraint is 10.92 ± 4.32 (length-normalized value of 0.21±0.083), when computed with a random sample of 1000, 5000, and 10,000. As shown in Fig. 6b, the average pairwise Hamming distance of sequences generated by `RNAiFold` with the frameshift stimulating structure (FSS) constraint, overlapping coding constraint and the slippery sequence constraint is 5.80 ± 1.84 (length-normalized value of 0.11 ± 0.035). Essentially, this means that approximately 11% of the positions (pairwise) are different for `RNAiFold` sampled sequences, compared with approximately 21% of the positions (pairwise) for `RNAsampleCDS`, compared with 81% of the positions (pairwise) for random RNA in positions 8-52 (i.e. after the fixed 7 nt slippery sequence). The greatest reduction in pairwise Hamming distance appears to be due to overlapping coding constraints, with an additional small reduction due to the FSS structural constraint.

HCV programmed −1 and +1 frameshifts

There is both in vitro and in vivo experimental evidence for a -2/+1 (hereafter designated as +1) and -1/+2 (hereafter designated as +2) programmed ribosomal frameshift in the core protein of the hepatitis C virus (HCV) [23]. The +1 frameshift produces a 17 kDa protein called protein F (Frameshift), also designated as ARFP (Alternative Reading Frame Protein). In addition, the +2 frameshift produces a 1.5 kDa protein. As measured by in vitro assays, the +1 ribosomal frameshift efficiency is ∼ 12 − 15%, while the +2 ribosoma frameshift efficiency is ∼ 30 − 45% [23]. Figure 7 depicts the organization of the overlapping coding region for the HCV genome (GenBank M62321.1), including a double stem-loop RNA structure designated as *frameshift stimulating signal* (FSS) depicted in Fig. 8. According to [23], the frameshift is caused by a poly-A slippery sequence (A AAA AAA AAC) in the triple coding region, although a mutated slippery sequence (A AGA AAA ACC) has also been shown to cause a frameshift, but with a lower efficiency. Out of 6,589 sequence hits for the HCV1 frameshift signal for the LANL HCV database (www.hcv.lanl.gov), we found that 94% of the sequences started with (A AGA

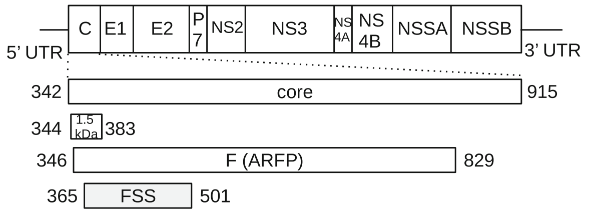
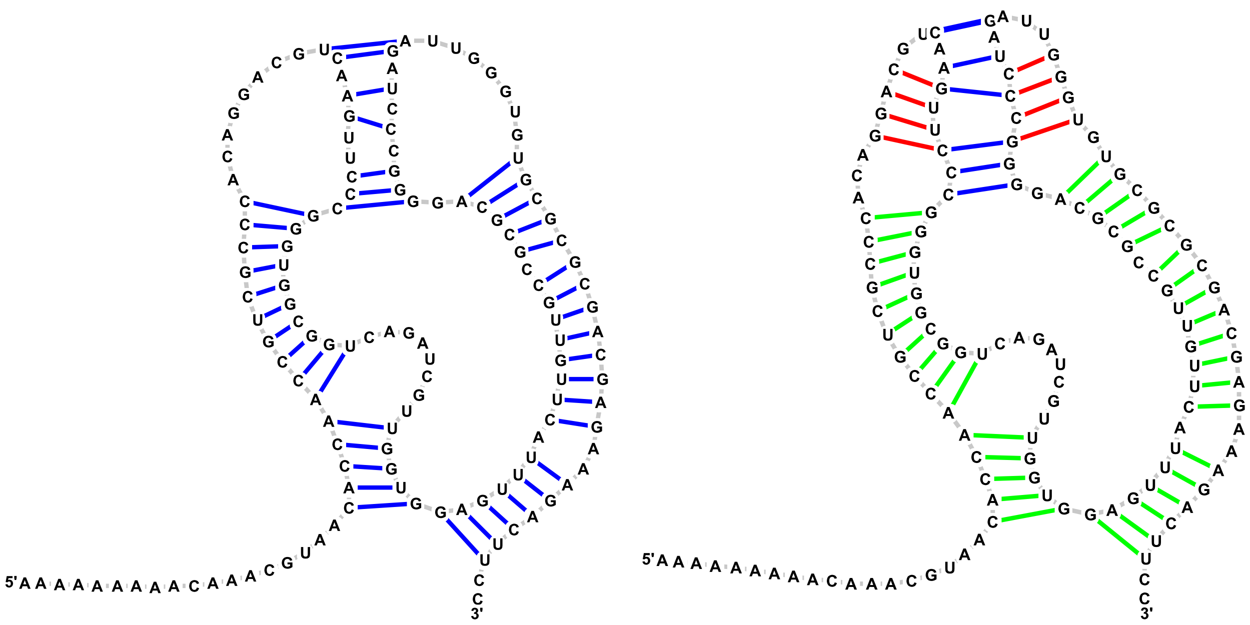


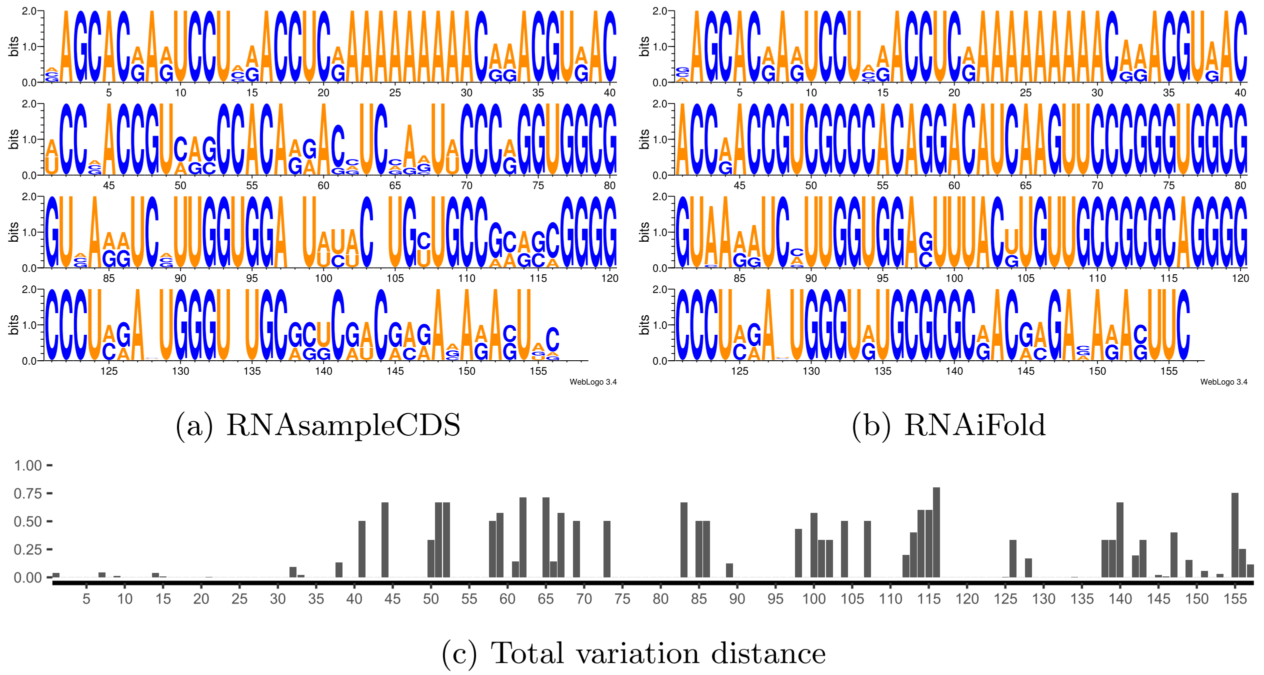
Fig. 7 Organization of the initially triple, then double overlapping reading frame region of hepatitis C virus (HCV) (GenBank M62321.1). The top gene organization map is adapted from Fig. 1a of [23]. All coding regions mentioned in the following include a terminal stop codon. The second line depicts the core in-frame protein, coded in nucleotides 342–915. Next, a 1.5 kDa protein is coded in nucleotides 344–383, while protein F is coded in nucleotides 346–829. The double stem-loop frameshift stimulating signal (FSS) is found at nucleotides 365-501; the FSS structure is depicted in Fig. 8

AAA ACC). Furthermore, downstream of the slippery sequence a double stem-loop structure facilitates translational frameshifting (Fig. 8). For this analysis, we took nucleotides 344-500 from the 9401 nt HCV subtype 1a genome (GenBank M62321.1) [23], corresponding to the region starting at the triple coding region and extending to the end of double-stem loop. Using `RNAsampleCDS` we computed the logo plot for all sequences that code BLOSUM62 +1 similar peptides to those coded by the reference genome (Fig. 9a). Using `RNAiFold 2.0` [24], we generated more than 11 million sequences that fold into the double-stem loop structure indicated in Fig. 8 and which have BLOSUM62 similarity of at least +1 to the reference genome peptides (Fig. 9b). Although `RNAiFold 2.0` does not support pseudoknot structures, by providing structural compatibility constraints, we ensured that every sequence returned by `RNAiFold 2.0` has the property that the nucleotides, which participate in the "kissing hairpin" model of Fig. 1a of [23], can indeed form a base pair together. Note that the set of all sequences returned by `RNAiFold 2.0`, which satisfy both the coding and structural requirements, forms a proper subset of the set of all sequences returned by `RNAsampleCDS`, which are required to satisfy only the coding requirements. Figure 9c depicts the total variation distance between these sequence two profiles. At positions where the total variation distance is zero, the secondary structure is likely to be *induced* by the overlapping coding constraints. Indeed, a mutation in such positions could lead to a disruption of the double stem-loop or to a modification of the amino acid in one of the overlapping reading frames. Our results from Fig. 9c agree with experimental evidence showing that modifications of nucleotides at positions 64, 91, 130 and 137 lead to *detrimental mutations* for the hepatitis C virus [25].



(a) Double stem-loop FSS from [23] (b) MFE structure from `RNAfold`

Fig. 8 HCV ribosomal frameshift stimulating signal (FSS). **a** Proposed pseudoknotted structure from [23]. **b** Minimum free eneergy (MFE) structure computed by `RNAfold 2.1.9` (*green*, *red*), with added pseudoknot (*blue*). Green arcs indicate common base pairs; *red arcs* indicate base pairs predicted by RNAfold but not present in the structure from [23]; *blue arcs* indicate pseudoknot base pairs from the model proposed by [23] that are absent from the `RNAfold` MFE structure. Figures produced using `jViz` [27]



(a) RNAsampleCDS

(b) RNAiFold

(c) Total variation distance

Fig. 9 a Exact sequence logo determined by `RNAsampleCDS` for all 2.55 × 10^{17} sequences, whose initial 39 nucleotides code amino acids having BLOSUM62 +1 similarity to the corresponding amino acids from each of the three reading frames in the triple overlapping coding region 344-383 of the reference HCV genome, and whose remaining nucleotides code amino acids having BLOSUM62 +1 similarity to the corresponding amino acids from each of the two reading frames in the double overlapping coding region 383-501 of the reference HCV genome. **b** Sequence logo determined by `RNAiFold 2.0` for the more than 11 million sequences that fold into the HCV FSS structure depicted in Fig. 8, whose initial 39 nucleotides code BLOSUM62 +1 amino acids having BLOSUM62 +1 similarity to the corresponding amino acids from each of the three reading frames in the triple overlapping coding region 344-383 of the reference HCV genome, and whose remaining nucleotides code amino acids having BLOSUM62 +1 similarity to the corresponding amino acids from each of the two reading frames in the double overlapping coding region 383-501 of the reference HCV genome. **c** Total variation distance shown for each nucleotide position, determined by computing the total variation distance between the position-specific profiles of (**a**) and (**b**)

Mutations at these positions resulted in an attenuated HCV infection in chimpanzee. According to our analysis, an introduction of mutations at positions whose variation distance is much greater than zero, should allow the disruption of the double-stem loop with minimal effects on the protein function. This hypothesis could be tested experimentally.

To further investigate whether the overlapping coding requirement of HCV possibly induces the FSS double stem-loop structure, we proceeded in a manner analogous to that for our HIV-1 analysis. We sampled 100,000 RNA sequences using `RNAsampleCDS` with BLOSUM62 similarity of +1 and 0 to the reference peptides in each reading frame. Using `RNAshapes`, we computed the average Boltzmann probability of formation of a double-stem loop with shape `[ ][ ]`, in the sampled RNA sequences as well as 6,589 sequences from LANL database (Additional file 1: Figure S5). Average Boltzmann probability of the double stem-loop shape `[ ][ ]` is 19% [resp. 9%] for BLOSUM62 similarity of +1 [resp. 0], compared with 98% probability for the sequences from LANL HCV database. In contrast, dinucleotide shuffles of sequences generated by `RNAsampleCDS` having BLOSUM62 +1 similarity to the reference peptides have average probability of 5% of double stem-loop formation, while the probability double stem-loop formation is 6% for random RNA sequences generated with probability of $\frac{1}{4}$ for each nucleotide. Additional file 1: Figure S5 displays average double stem-loop probability and free energy results for the HCV overlapping coding region, which are analogous to results for HIV-1 presented in Fig. 5.

Conclusion

In this paper, we have developed the novel program `RNAsampleCDS`, the only existent program which computes the number of RNA sequences that code user-specified peptides in one to six overlapping reading frames, as depicted in Fig. 1b. More importantly, `RNAsampleCDS` can compute (exact) PSSMs and sample, in an unweighted or weighted fashion, a user-specified number of RNA sequences that code the specified proteins (or code proteins having BLOSUM/PAM similarity that exceeds a user-specified threshold to the given proteins). With extensions to `RNAiFold2.0` made in this paper, `RNAsampleCDS` and `RNAiFold2.0` complement each other and together allow one to analyze the HIV-1 Gag-Pol overlapping reading frame and the HCV triple overlapping reading frame in a manner that cannot be supported by any other software, thus augmenting the software arsenal available to evolutionary biologists.

Additional file

Additional file 1: Supplementary information for "New tools to analyze overlapping coding regions". (PDF 416 kb)

Abbreviations

ARFP: Alternative reading frame protein; CPI: Codon preference index; DP: Dynamic programming; FSS: Frameshift stimulating signal; HCV: Hepatitis C virus; HIV: Human immunodeficiency virus; MFE: Minimum free energy; PSSM: Position-specific scoring matrix, also called profile

Acknowledgements

Research of the Clote Lab was supported by National Science Foundation grant DBI-1262439. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Funding

The research was funded by National Science Foundation grant DBI-1262439. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Availability of data and materials

Source code for the python software `RNAsampleCDS` is available at http://bioinformatics.bc.edu/clotelab/RNAsampleCDS.

Authors' contributions

Project design PC. Algorithm design PC. Implementation PC, AHB, JAGM. Software testing and data analysis PC, AHB, JACM. Manuscript preparation PC, AHB. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 21 August 2016 Accepted: 26 November 2016
Published online: 13 December 2016

References

1. Dinman JD. Programmed Ribosomal Frameshifting Goes Beyond Viruses: Organisms from all three kingdoms use frameshifting to regulate gene expression, perhaps signaling a paradigm shift. Microbe Wash DC. 2006;1(11):521–7. doi:17541450.
2. Ofori LO, Hilimire TA, Bennett RP, Brown Jr NW, Smith HC, Miller BL. High-affinity recognition of HIV-1 frameshift-stimulating RNA alters frameshifting in vitro and interferes with HIV-1 infectivity. J Med Chem. 2014;57(3):723–32.
3. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. Viennarna Package 2.0. Algorithms Mol Biol. 2011;6:26.
4. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43(Database issue): D130–D137.
5. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997;13(5):555–6. doi:9367129.
6. Pond SL, Frost SD, Muse SV. Hyphy: hypothesis testing using phylogenies. Bioinformatics. 2005;21(5):676–9. doi:15509596.
7. Gojobori T, Ishii K, Nei M. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J Mol Evol. 1982;18(6):414–23. doi:7175958.
8. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994;11(5):725–36. doi:7968486.
9. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000;155(1):431–49. doi:10790415.
10. Pedersen AM, Jensen JL. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol Biol Evol. 2001;18(5):763–6. doi:10.1093/oxfordjournals.molbev.a003859.
11. Sabath N, Landan G, Graur D. A method for the simultaneous estimation of selection intensities in overlapping genes. PLoS ONE. 2008;3(12):3996. doi:19098983.
12. Sabath N, Graur D. Detection of functional overlapping genes: simulation and case studies. J Mol Evol. 2010;71(4):308–16. doi:20820768.
13. Pedersen JS, Forsberg R, Meyer IM, Hein J. An evolutionary model for protein-coding regions with conserved RNA structure. Mol Biol Evol. 2004;21(10):1913–22. doi:10.1093/molbev/msh199.
14. Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. 1997. doi:10.1093/bioinformatics/13.3.235.
15. Hudelot C, Gowri-Shankar V, Jow H, Rattray M, Higgs PG. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences,. Mol Phylogenet Evol. 2003;28(2):241–52.
16. Gesell T, von Haeseler A. In silico sequence evolution with site-specific interactions along phylogenetic trees. Bioinformatics. 2006;22(6):716–22. doi:10.1093/bioinformatics/bti812.
17. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinforma. 2008;9:474.
18. Sealfon RS, Lin MF, Jungreis I, Wolf MY, Kellis M, Sabeti PC. FRESCo: finding regions of excess synonymous constraint in diverse viruses. Genome Biol. 2015;16:38. doi:25853568.
19. Gribskov M, Devereux J, Burgess RR. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res. 1984;12(1):539–49. doi:6694906.
20. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12(1):32–42. doi:21102527.
21. Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics. 2006;22(4):500–3.
22. Giegerich R, Voss B, Rehmsmeier M. Abstract shapes of RNA. Nucleic Acids Res. 2004;32(16):4843–851.
23. Choi J, Xu Z, Ou JH. Triple decoding of hepatitis C virus RNA by programmed translational frameshifting. Mol Cell Biol. 2003;23(5): 1489–97. doi:10.1128/MCB.23.5.1489.
24. Garcia-Martin JA, Dotu I, Clote P. RNAiFold 2.0: a web server and software to design custom and Rfam-based RNA molecules. Nucleic Acids Res. 2015;43(W1):513–21. doi:26019176.
25. McMullan LK, Grakoui A, Evans MJ, Mihalik K, Puig M, Branch AD, Feinstone SM, Rice CM. Evidence for a functional RNA element in the hepatitis C virus core gene,. Proc Natl Acad Sci U S A. 2007;104(8): 2879–84. doi:10.1073/pnas.0611267104.
26. Crooks GE, Hon G, Chandonia JM, Brenner SE. Weblogo: a sequence logo generator. Genome Res. 2004;14(6):1188–1190.
27. Wiese KC, Glen E, Vasudevan A. JViz,Rna–a Java tool for RNA secondary structure visualization. IEEE Trans Nanobioscience. 2005;4(3):212–8.