BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(9)

PKscan: a program to identify H-type RNA pseudoknots in any RNA sequence with unlimited length

Xiaolan Huang[1], Zhihua Du[2], Jie Cheng[3] & Qiang Cheng[1]*

[1]Department of Computer Science, Southern Illinois University at Carbondale, IL 62901, USA; [2]Department of Chemistry and Biochemistry, Southern Illinois University at Carbondale, IL 62901, USA; [3]Department of Computer Science and Engineering, University of Hawaii at Hilo, HI 96720, USA; Qiang Cheng – Email: qcheng@cs.siu.edu; phone: (01) 618-453-6056; Fax: (01) 618-453-6044; *Corresponding author

Received March 04, 2013; Accepted March 08, 2013; Published May 25, 2013

Abstract:
A computer program written in C++ has been developed which can detect all potential H-type RNA pseudoknots within any given RNA sequence. There is no limit on the length of the input sequence. A validation run of the program using the full-length (8173 nt) genomic mRNA of simian retrovirus type-1 (SRV-1) identifies the established -1 frameshift stimulating pseudokont at the gag-pro junction as the most stable pseudoknot within the genomic mRNA.

Keywords: RNA, pseudoknot, -1 ribosomal frameshifting, computational biology.

Background:
RNA pseudoknot is a structural motif of RNA formed when a stretch of nucleotides within a loop region in a secondary structure basepair with residues outside that loop **[1, 2].** Many distinct folding topologies exist for pseudoknotted structures. The most frequently occurring pseudoknots are the so-called H-type (or hairpin type) pseudoknots. In a typical H-type pseudoknot, sequence within the loop of a hairpin (stem-loop) structure of RNA form intramolecular basepairing interaction with a complementary sequence outside the hairpin **(Figure 1).** All H-type pseudoknots contain two helical stems, S1 and S2, and two non-equivalent loops, L1 and L2. Some H-type pseudoknots also contain a third loop, L3. If L3 is absent, S1 and S2 can form a quasi-continuous double helix, with loops L1 and L2 crossing the major groove and minor groove of stem S2 and stem S1 respectively **(Figure 1C).** H-type pseudoknots assume diverse biological functions. The best-known function is to stimulate -1 ribosomal frameshifting in the translation of viral proteins **[3].** In a -1 ribosomal frameshifting event, a certain percentage of the translating ribosomes shift back by one nucleotide on the mRNA therefore change to the -1 reading frame. Although frameshifting occurs at a so-called slippery sequence with a typical composition of X XXY YYZ (XXX and YYY: a stretch of three identical nucleotides; the triplets XXY and YYZ indicate two codons in the 0 reading frame), an RNA structure several nucleotides downstream is required for efficient framshifting. Most often, the downstream RNA structure is an H-type pseudoknot. To facilitate the identification of H-type pseudoknots in RNAs, e.g. in the full-length genomic mRNAs of retroviruses, we have developed a computer program (written in C++) that is capable of identifying all potential H-type pseudoknots in any RNA sequence with unlimited length.

Methodology:
Figure 1A shows a linear presentation of the sequence elements of a typical H-type pseudoknot, which requires that both helical stems (S1 and S2) form simultaneously. If a given RNA sequence contains two pairs of complementary stretches (S1-5' complementary to S1-3', and S2-5' complementary to S2-3'. G-U

is considered a legitimate base pair) separated by two or three connecting unpaired regions (L1, L2 and optionally L3) with a sequential arrangement as shown in (**Figure 1A)**, then this sequence has the potential to form an H-type pseudoknot. The computer program tests all possible combinations of stem and loop lengths within certain ranges (which can be set by the user) to check whether the pseudoknot-forming criteria can be met.

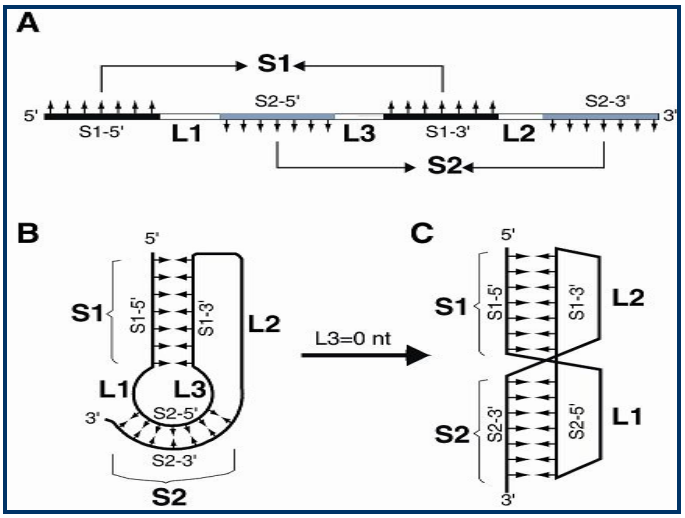

Figure 1: Schematic diagrams of the sequence elements for forming an H-type RNA pseudoknot. Abbreviations used are: S1, stem1; S2, stem2; S1-5' and S1-3', the 5' and 3' strands of stem1; S2-5' and S2-3', the 5' and 3' strands of stem2; L1, loop1; L2, loop2; L3, loop3. A: linear sequential arrangement of the pseudoknot-forming sequence elements. Complementary strands forming S1 and S2 are painted black and gray respectively. Small arrows indicated the complementarities of the strands. A pseudoknot requires that both S1 and S2 can form simultaneously. B: Schematic representation of folded pseudoknots with a non-zero L3. C: with the absence of L3, S1 and S2 can stack coaxially to form a quasi-continuous double helix. L1 and L2 are located on the same side of the molecule, with L1 crossing the major groove of S2 and L2 crossing the minor groove of S1.

In order to compare the relative thermodynamic stability of the identified pseudoknots within a given mRNA sequence, we have also implemented free energy (ΔG_{37}) calculation for the two helical stems S1 and S2. In calculating the free energy, the Turner's nearest-neighbor parameters are used **[4].** If L3=0, the two stems are taken as a continuous helical stem for the calculation, but only half of the value is given to the S1-S2 stack to account for the quasi-continuous nature of the stacked stems. Although this simplified free energy calculation should only be viewed as semi-quantitative, it provides a reasonable estimation of the relative stability of the detected pseudoknots, which are ranked according to the calculated free energies. Input and output. The input is an mRNA (or cDNA) sequence. The program recognizes G, C, A, U, and T (for cDNA sequences) in both upper and lower cases, this is to ensure that input sequence file can be generated easily and reliably from sequences downloaded from nucleotide sequence databases. Before the search, the user is prompted to set the ranges for S1, S2, L1, L2 and L3, or accept the default values (default ranges are 5 to 20 base pairs, 5 to 20 base pairs, 1 to 10 nucleotides, 3 to 50 nucleotides, and 0 to 10 nucleotides respectively for S1, S2, L1, L2 and L3). The threshold for calculated free energy (default value -18 kcal/mol) can also be set. Pseudoknots with a calculated free energy higher than the threshold will not be reported in the output. The output file of the program contains information about whether pseudoknots are found and how many are found; the detected pseudoknots are then listed in the order of calculated free energy of the stems. For each of the detected pseudoknots, the following information is given: lengths of S1, S2, L1, L2, and L3; free energy value of the stems; size and location of the pseudoknot. A schematic diagram is then drawn, showing the actual pseudoknot forming sequence and base-pairing schemes of the two stems; a sequence of 20 nucleotides immediately 5'- to the pseudoknot is also shown to facilitate the identification of potential slippery sequence in case of -1 frameshift stimulating pseudoknots.

```
Sequence #: M11841 (8173) nucleotides

Found 50 pseudoknots with stem free energy lower than -18.0 kcal/mol.

Pseudoknot 1: -34.710000 (kcal/mol): Start=2337 S1=6 S2=6 L1=1 L2=12 L3=0 End=2373
                          L2          L1
                    GAAACAAGCTTA       A

                          CCCCGG      ACCCCG
CCATCAGGGAAACGGACTGAGGGGCC            TGGGGCGG
  slippery sequence       S1          S2
Pseudoknot 2: -24.750000 (kcal/mol): Start=2351 S1=6 S2=6 L1=4 L2=39 L3=1 End=2417
                    GGTCAGCTTTGTTCCAGCCAACAAAAACAACCCATTTCA       AAAC

                    CGGGGT  A  TTCGAA
                            L3
GACTGAGGGGCCAGCCCCAGGCCCCG      AAGCTTAC
```

Figure 2: A typical output of running the PKscan program, using the full-length genomic mRNA of the simain retroviruses type-1 (SRV-1, accession number M11841) as the input sequence. A total of 50 pseudoknots with a calculated stem free energy lower than-18 kcal/mol were identified (only the 1st & 2nd ranked pseudoknots are shown for clarity). The texts and boxes in red are not parts of the original output. They are added to indicate the sequence elements of the identified pseudoknots**.** The software can be downloaded from: http://www2.cs.siu.edu/~xhuang.

Validation:

To show the usefulness of PKscan, we have performed a search on the full-length genomic RNA of simain retroviruses type-1 (SRV-1, accession number M11841) that has 8173 nucleotides. Using the default values for stems and loops, it took ~78 minutes to finish the search on a Linux workstation (Red-hat Enterprise Linux 5.6 on Dell precision T5500). As shown in (**Figure 2)**, a total of 50 potential pseudoknots were identified within the SIV-1 genomic RNA. Significantly, the established -1 frameshift stimulating pseudoknot at the gag-pro junction **[5, 6]** is identified as the most stable pseudoknot (lowest calculated free energy, -34.71 kcal/mol). The calculated free energy of this frameshift stimulating pseudoknot is lower than the second ranked pseudoknot by ~10 kcal/mol. Scanning the SRV-1 genomic RNA for pseudoknot detection showcases a unique utility of the PKscan program which allows assessment on the strategic importance of the frameshift-stimulating pseudoknot within the viral genome. The results also show that there is no ultra-stable pseudoknot in the mRNA sequence which would become a roadblock to terminate translation **[7].** Of course, PKscan can also be useful in many other scenarios.

Conclusion:

PKscan can be used to efficiently and reliably identify all potential H-type pseudoknots in any given RNA sequence. There is no limit on the length of the RNA sequence therefore very long RNA sequences such as the full-length viral RNAs (several thousands to tens of thousands of nucleotides) can be scanned. As long as the sequence elements of a potential pseudoknot fall within the ranges defined by the program, the pseudoknot will not elude being detected.

Acknowledgement:

The work was supported by the start-up fund and a seed grant from Southern Illinois University Carbondale to Z.D., and a grant from University of Hawaii to J.C. This work is also supported by the National Science Foundation under Grant No. 1218712. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References:

[1] Pleij CW *et al. Nucleic Acids Res*. 1985 **13:** 1717 [PMID: 4000943]
[2] Pleij CW, *Trends Biochem Sci*. 1990 **15**: 143 [PMID: 1692647]
[3] Gesteland RF & Atkins JF, *Annu Rev Biochem*. 1996 **65:** 741 [PMID: 8811194]
[4] Serra MJ & Turner DH, *Methods Enzymol*. 1995 **259**: 242 [PMID: 8538457]
[5] Dam EBt *et al*. *RNA*. 1995 **1**: 146 [PMID: 7585244]
[6] Michiels PJ *et al. J Mol Biol.* 2001 **310**: 1109 [PMID: 11501999]
[7] Tholstrup J *et al. Nucleic Acids Res*. 2012 **40**: 303 [PMID: 21908395]

Edited by P Kangueane

Citation: **Huang *et al.*** Bioinformation 9(9): 440-442 (2013)