AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

DATABASES AND SOFTWARE

StemLoop-Finder: a Tool for the Detection of DNA Hairpins with Conserved Motifs

Alyssa A. Pratt,[a,b,c] Ellis L. Torrance,[a,d] George W. Kasun,[a] Kenneth M. Stedman,[a] Ignacio de la Higuera[a]

[a]Department of Biology, Center for Life in Extreme Environments, Portland State University, Portland, Oregon, USA
[b]Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon, USA
[c]Department of Computer Science, Oregon State University, Corvallis, Oregon, USA
[d]Department of Biology, University of North Carolina Greensboro, Greensboro, North Carolina, USA

ABSTRACT Nucleic acid secondary structures play important roles in regulating biological processes. StemLoop-Finder is a computational tool to recognize and annotate conserved structural motifs in large data sets. The program is optimized for the detection of stem-loop structures that may serve as origins of replication in circular replication-associated protein (Rep)-encoding single-stranded (CRESS) DNA viruses.

Circular replication-associated protein (Rep)-encoding single-stranded (CRESS) DNA viruses are a highly diverse group of viruses that includes several virus families, such as the *Circoviridae*, *Nanoviridae*, and *Geminiviridae* (1, 2). CRESS DNA viruses replicate through a rolling circle mechanism (3, 4). To initiate replication, the viral Rep nicks a conserved nonanucleotide sequence within a stem-loop DNA structure (5–9). Locating this feature is important for understanding the characteristics of a particular CRESS genome (10–13). Detection of potential stem-loop structures with nonanucleotide motifs was previously performed manually (10, 14). This process is time-consuming, especially for large metagenomic data sets. By automating identification of the nonanucleotide motifs and secondary structures, StemLoop-Finder increases efficiency and produces an annotated file with scored potential stem-loops for each viral genome analyzed. The biological significance of the predicted stem-loop structures should be assessed rationally or experimentally by the user.

StemLoop-Finder is written in Python within the PyCharm integrated development environment and can be run through the command-line interface on Mac OS, Windows (virtual machine), or Linux operating systems. It uses the ViennaRNA 2.0 library (15) to predict secondary structures in a DNA sequence using user-supplied prediction parameters and the library's minimum free energy algorithms. It reads FASTA (with tinyfasta 0.1.0; https://pypi.org/project/tinyfasta/) and general feature format (GFF) sequence files and outputs stem-loop annotations as a GFF file and a more detailed comma-separated value (CSV) file (Fig. 1). Users input a desired CRESS DNA virus family or a 9-nucleotide sequence following the International Union of Pure and Applied Chemistry (IUPAC) degenerate base symbol standard (16). Another argument is used to determine the number of bases on either side of the nonanucleotide processed by the software for secondary structure prediction. These and other arguments are interpreted in Python with the argparse library.

ViennaRNA is used to predict the secondary structure of the defined region according to the parameters given, generating a dot-bracket model of the predicted structure (15). The user may use multiple parameter files and frame sizes to increase the number of stem-loop detections. In order to be scored, a stem-loop must have a stem length of at least 5 nucleotides and a loop length of at least 7 nucleotides. Each putative stem-loop is scored +1 point for each deviation of 1 nucleotide from the ideal stem or loop length and −5 points for high similarity to a specific nonanucleotide sequence, determined by

Citation Pratt AA, Torrance EL, Kasun GW, Stedman KM, de la Higuera I. 2021. StemLoop-Finder: a tool for the detection of DNA hairpins with conserved motifs. Microbiol Resour Announc 10:e00424-21. https://doi.org/10.1128/MRA.00424-21.

Editor Simon Roux, DOE Joint Genome Institute

Address correspondence to Ignacio de la Higuera, ide@pdx.edu.

Received 29 April 2021
Accepted 8 June 2021
Published 1 July 2021

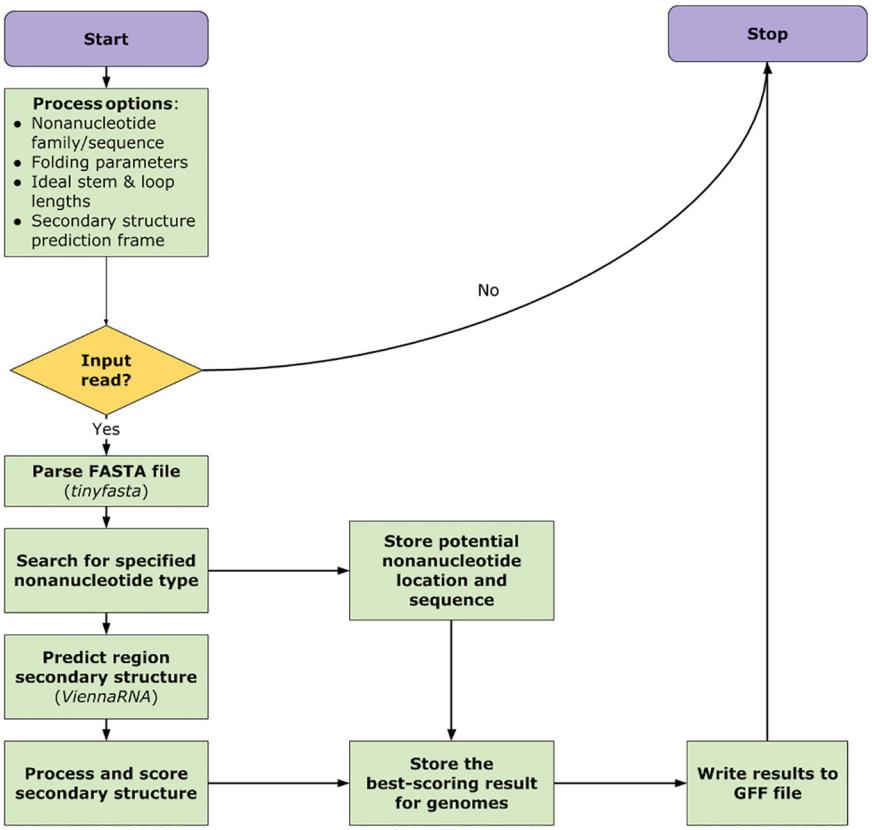

FIG 1 Flow chart depicting the StemLoop-Finder pipeline, with third-party tools indicated by italicized text.

the user as an argument or by the input viral family name. In order for a stem-loop to be annotated within the GFF file, it must have a score of less than 15 (or another user-defined value) and cannot have a nonanucleotide within 4 bases of the start or end of the potential stem-loop structure.

StemLoop-Finder was tested with a diverse set of publicly available CRESS DNA viral sequences from terrestrial arthropods for which stem-loops had been manually annotated (10). StepLoop-Finder detected stem-loops in 33 of the 44 sequences using the nonanucleotide motif NANTATTAC, which was used for the manual search (10). In six that were not detected, the nonanucleotide found manually did not fit NANTATTAC, and in the remaining five, the sequence surrounding the putative nonanucleotides was not predicted to form a stem-loop structure. Thus, StemLoop-Finder can be reliably used to automatically predict stem-loop structures in genomic and metagenomic data sets (12).

Data availability. The software source code is available on a public Bitbucket repository (https://bitbucket.org/crucicrew/sl-finder/src/master/) to be compiled from source or as a Docker container. It will remain freely available for the next 10 years alongside instructions for use and any applicable updates.

ACKNOWLEDGMENTS

We thank Arvind Varsani for discussions.

This work was supported by grant 80NSSC17K0301 from NASA, grant MCB-2025305 from the NSF, the Apprenticeships in Science and Engineering Program, Sigma Xi of the Columbia Willamette, John Howieson, and Alison Stenger.

REFERENCES

1. Zhao L, Rosario K, Breitbart M, Duffy S. 2019. Eukaryotic circular Rep-encoding single-stranded DNA (CRESS DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. Adv Virus Res 103:71–133. https://doi.org/10.1016/bs.aivir.2018.10.001.

2. Krupovic M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, Rosario K, Yutin N, Wolf YI, Harrach B, Zerbini FM, Dolja VV, Kuhn JH, Koonin EV. 2020. Cressdna-viricota: a virus phylum unifying seven families of Rep-encoding viruses with single-stranded, circular DNA genomes. J Virol 94:71–133. https://doi.org/10.1128/JVI.00582-20.
3. Saunders K, Lucy A, Stanley J. 1991. DNA forms of the geminivirus African cassava mosaic virus consistent with a rolling circle mechanism of replication. Nucleic Acids Res 19:2325–2330. https://doi.org/10.1093/nar/19.9.2325.
4. Jeske H, Lütgemeier M, Preiß W. 2001. DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. EMBO J 20:6158–6167. https://doi.org/10.1093/emboj/20.21.6158.
5. Heyraud-Nitschke F, Schumacher S, Laufs J, Schaefer S, Schell J, Gronenborn B. 1995. Determination of the origin cleavage and joining domain of geminivirus rep proteins. Nucleic Acids Res 23:910–916. https://doi.org/10.1093/nar/23.6.910.
6. Steinfeldt T, Finsterbusch T, Mankertz A. 2006. Demonstration of nicking/joining activity at the origin of DNA replication associated with the Rep and Rep′ proteins of porcine circovirus type 1. J Virol 80:6225–6234. https://doi.org/10.1128/JVI.02506-05.
7. Hafner GJ, Stafford MR, Wolter LC, Harding RM, Dale JL. 1997. Nicking and joining activity of banana bunchy top virus replication protein in vitro. J Gen Virol 78:1795–1799. https://doi.org/10.1099/0022-1317-78-7-1795.
8. Laufs J, Jupin I, David C, Schumacher S, Heyraud-Nitschke F, Gronenborn B. 1995. Geminivirus replication: genetic and biochemical characterization of rep protein function, a review. Biochimie 77:765–773. https://doi.org/10.1016/0300-9084(96)88194-6.
9. Stanley J. 1995. Analysis of African cassava mosaic virus recombinants suggests strand nicking occurs within the conserved nonanucleotide motif during the initiation of rolling circle DNA replication. Virology 206:707–712. https://doi.org/10.1016/S0042-6822(95)80093-X.
10. Rosario K, Mettel KA, Benner BE, Johnson R, Scott C, Yusseff-Vanegas SZ, Baker CCM, Cassill DL, Storer C, Varsani A, Breitbart M. 2018. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. PeerJ 6:e5761. https://doi.org/10.7717/peerj.5761.
11. Dayaram A, Galatowitsch ML, Argüello-Astorga GR, van Bysterveldt K, Kraberger S, Stainton D, Harding JS, Roumagnac P, Martin DP, Lefeuvre P, Varsani A. 2016. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. Infect Genet Evol 39:304–316. https://doi.org/10.1016/j.meegid.2016.02.011.
12. de la Higuera I, Kasun GW, Torrance EL, Pratt AA, Maluenda A, Colombet J, Bisseux M, Ravet V, Dayaram A, Stainton D, Kraberger S, Zawar-Reza P, Goldstien S, Briskie JV, White R, Taylor H, Gomez C, Ainley DG, Harding JS, Fontenele RS, Schreck J, Ribeiro SG, Oswald SA, Arnold JM, Enault F, Varsani A, Stedman KM. 2020. Unveiling crucivirus diversity by mining metagenomic data. mBio 11:e01410-20. https://doi.org/10.1128/mBio.01410-20.
13. de la Higuera I, Torrance EL, Pratt AA, Kasun GW, Maluenda A, Stedman KM. 2019. Genome sequences of three cruciviruses found in the Willamette Valley (Oregon). Microbiol Resour Announc 8:e00447-19. https://doi.org/10.1128/MRA.00447-19.
14. Cheung AK, Ng TFF, Lager KM, Alt DP, Delwart E, Pogranichniy RM. 2015. Identification of several clades of novel single-stranded circular DNA viruses with conserved stem-loop structures in pig feces. Arch Virol 160:353–358. https://doi.org/10.1007/s00705-014-2234-9.
15. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA package 2.0. Algorithms Mol Biol 6:26. https://doi.org/10.1186/1748-7188-6-26.
16. Cornish-Bowden A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res 13:3021–3030. https://doi.org/10.1093/nar/13.9.3021.