BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 8(13)

shRNAPred (version 1.0): An open source and standalone software for short hairpin RNA (shRNA) prediction

Nishtha Singh[1], Tanmaya Kumar Sahu[1], Atmakuri Ramakrishna Rao[1]* & Trilochan Mohapatra[2]

[1]Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi, India, 110 012; [2]Central Rice Research Institute, Cuttack, Odisha, India, 753 006; AR Rao – E mail: arrao@iasri.res.in; Phone: +91-9999422935; Fax: +91-11-25841564; *Corresponding author

Received May 13, 2012; Accepted June 08, 2012; Published July 06, 2012

Abstract:
The small hairpin RNAs (shRNA) are useful in many ways like identification of trait specific molecular markers, gene silencing and characterization of a species. In public domain, hardly there exists any standalone software for shRNA prediction. Hence, a software *shRNAPred (1.0)* is proposed here to offer a user-friendly Command-line User Interface (CUI) to predict 'shRNA-like' regions from a large set of nucleotide sequences. The software is developed using PERL Version 5.12.5 taking into account the parameters such as stem and loop length combinations, specific loop sequence, GC content, melting temperature, position specific nucleotides, low complexity filter, etc. Each of the parameters is assigned with a specific score and based on which the software ranks the predicted shRNAs. The high scored shRNAs obtained from the software are depicted as potential shRNAs and provided to the user in the form of a text file. The proposed software also allows the user to customize certain parameters while predicting specific shRNAs of his interest. The shRNAPred (1.0) is open access software available for academic users. It can be downloaded freely along with user manual, example dataset and output for easy understanding and implementation.

Availability: http://bioinformatics.iasri.res.in/EDA/downloads/shRNAPred_v1.0.exe

Keywords: shRNA, shRNA prediction, RNAi, Gene silencing

Background:
A shRNA is a tight hairpin turn, with a loop of 4–23 nucleotides and a stem (two anti parallel strands) of 19–29 nucleotide base pairs **[1]**. shRNAs are pivotal in the field of gene silencing as these are cheaper than siRNAs for large-scale studies **[2]**. However, currently available web-based tools fail to predict shRNAs from a large set of nucleotide sequences. Also, hardly any stand alone software exists in the public domain for this purpose. Hence, the aim of this paper is to develop standalone software, to facilitate prediction of 'shRNA-like' regions, by considering an exhaustive list of hairpin parameters, from voluminous genomic sequence data. This software will cater the needs of the researchers and scientists working in the field of RNA interference in designing shRNA.

Methodology:
Initially, different properties of shRNA, like, stem and loop lengths, perfect stem complementarity, GC content, melting temperature (Tm), position specific nucleotides and low complexity regions are taken into consideration while developing the script for shRNAPred (version 1.0). The script is developed using Active PERL 5.12.3. Further, the executable file was generated by using Perl Packager (pp) module, provided by Perl Archive Toolkit (PAR) version 0.85_01 of

Comprehensive Perl Archive Network (CPAN) in windows environment. In addition, the script is configured with modules for each property and the software uses these modules based on the software options. The parameters GC content and Tm considered in the software are calculated as follows;

Calculation of GC content: [(C_count+G_count)*100]/[(2*(Stem_Length) +Loop length)]; Calculation of Tm **a)** $Tm[^oC]=64.9+[(41*(n_G+n_C-16.4))/(n_A+n_T+n_G+n_C)]$(if length>15) **[3]**; **b)** $Tm[^oC]=2*(n_A+n_T) + 4*(n_G+n_C)$(if length<=15) Where n= number of nucleotides

Scoring System

Different scores are assigned for each parameter by considering various favorable and unfavorable properties of shRNA. The favorable properties are assigned with positive scores whereas unfavorable properties are assigned with penalty. For property TM, a score +1 is given when it lies in the range 20^0C-60^0C **[4].** In a similar way, for GC content a score of +1 is given in the range 35% - 60% **[5].** A penalty of -1 is added for presence of Poly A or Poly C and +1, otherwise **[6].** A penalty -0.2 is added for each complimentary base pair in the loop sequence since more the number of complementary bases the lesser is the chance of the sequence being a loop. Presence or absence of certain nucleotides at specific positions in stem often increases the efficacy of the shRNA **[7, 8].** Hence, suitable scores, are given for the properties like A at 3rd position of 5′sense strand (1,0), T at 10^{th} position of 5′ sense strand(1,0), G/C is present at 1st position of 5′ sense strand (1,-1), A/T at 19^{th} position of 5′ sense strand(1,-1), G at 13^{th} position of 5′ sense strand(-1,0), T is at 13^{th} position of 5′ sense strand(1,0), T at 1^{st} position of 5′ antisense strand(2,0) and A/U nucleotide in any of the first five position at 5′ antisense strand(1,0); where the first value in every parenthesis ( ) is meant for presence and the second value meant for absence of a nucleotide at a given position on 5′ sense / antisense strand. Based on the above set parametric score a total score is computed for each shRNA-like region. Ranks are then assigned to these regions based on magnitude of the score, with rank 1 being given to highest scored region.

Software input / output:

Input

The software accepts the input sequence file in FASTA format. The header line for every sequence should contain the information *viz.*, Gene ID, accession number and definition, separated by pipelines and multiple sequences must be separated by a new line.

Options

The software provides three different options and the user can choose one option at a time for predicting shRNA from the input.

User defined Stem and loop lengths

In this choice the user needs to enter stem length and loop length of his choice. Besides, range of GC-content and number of loop-end complementary residues is to be provided by the user under this option.

Predefined Stem and loop length combinations

In this case, the users don't need to provide the stem and loop lengths, as the stem and loop length combinations *viz*. 29-4, 19-9, 19-10, 21-9, 25-10, 19-4, 29-9 and 27-4 **Table 1 (see supplementary material)** are predefined based on literature. Here also the user is required to enter the range of GC content and number of loop-end complementary residues as given in option - User defined stem and loop lengths.

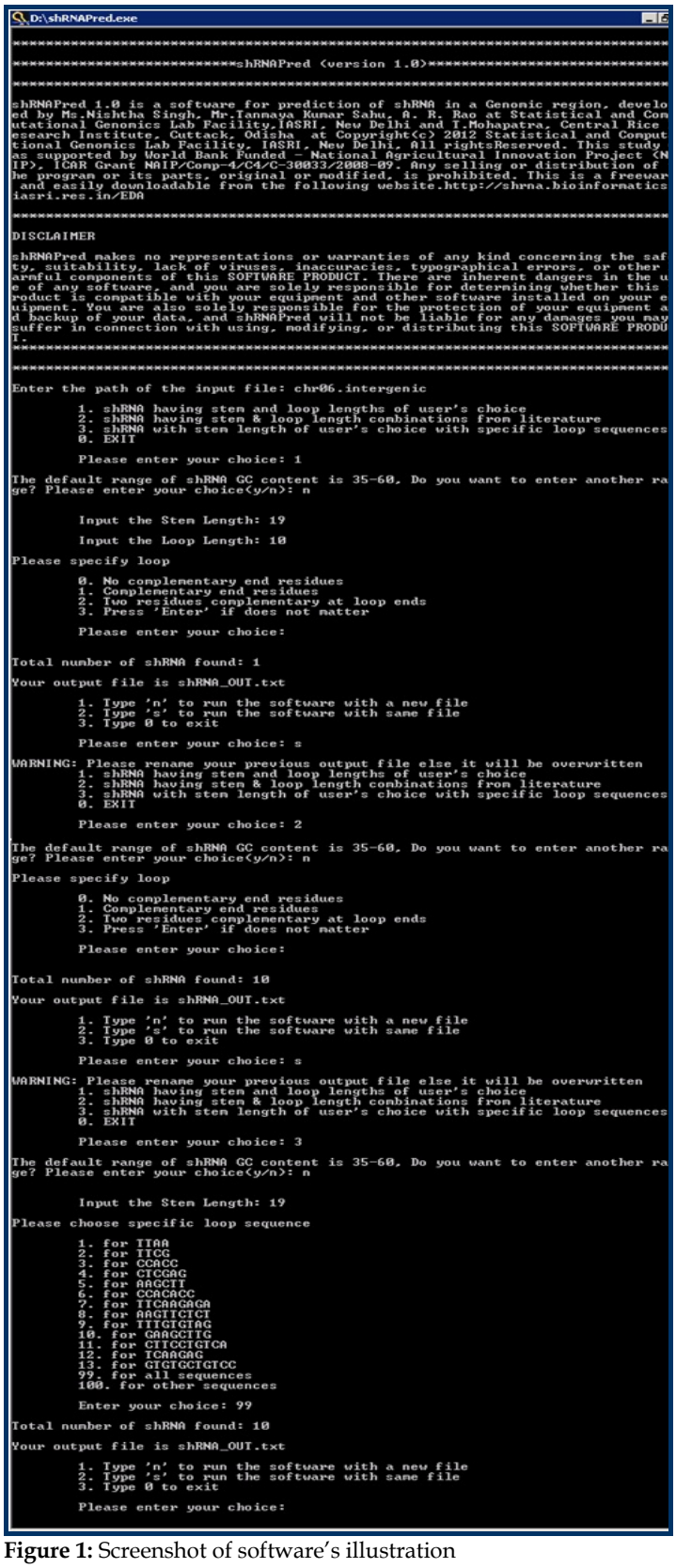

Figure 1: Screenshot of software's illustration

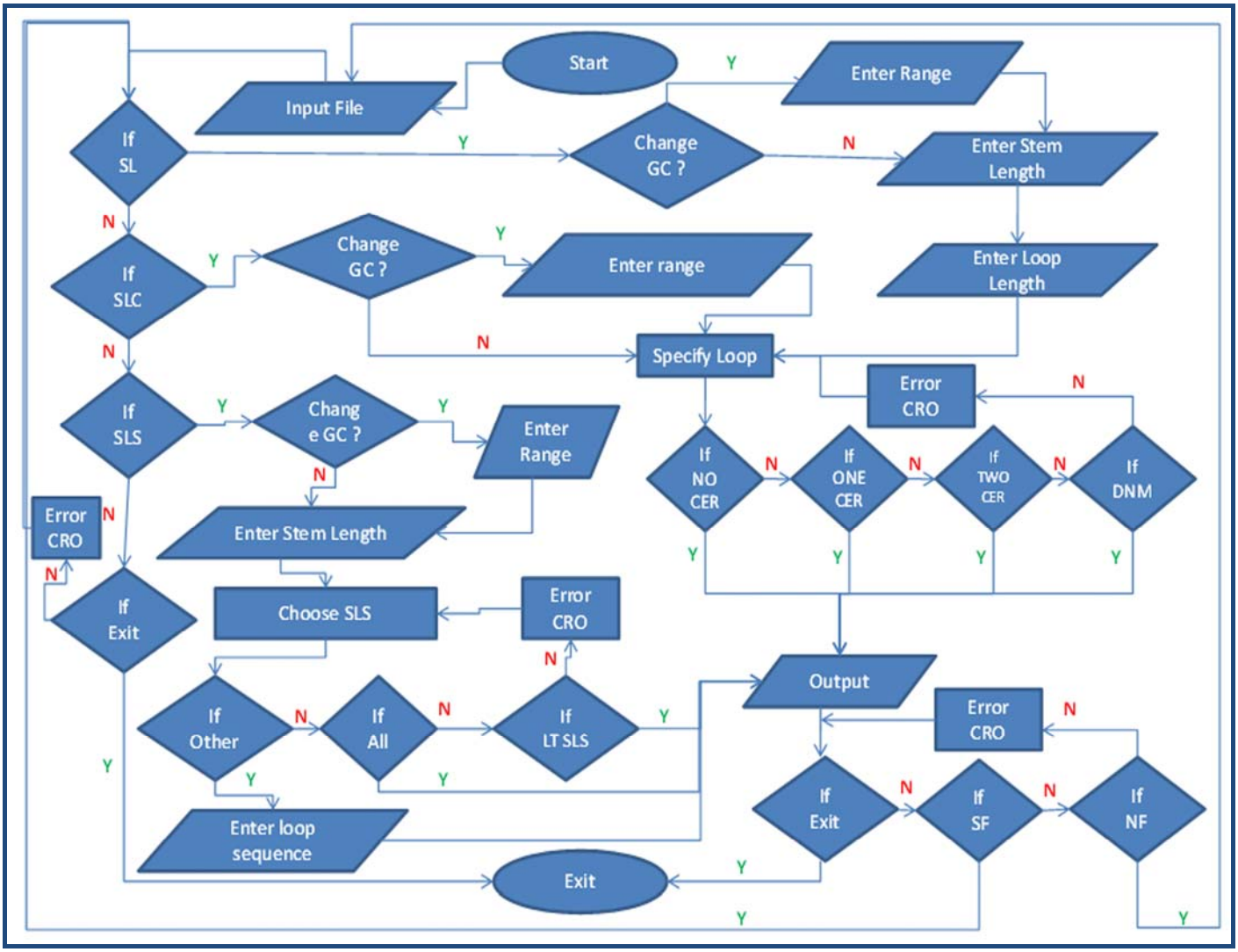

Figure 2: Flowchart for shRNAPred (version 1.0); SL-shRNA having Stem and Loop lengths of user's choice (option "1"), SLC-shRNA having Stem and Loop length Combinations from literature (option "2"), SLS- shRNA with Stem Length of user's choice with Specific loop sequences (option "3"), GC-range of Guanine-Cytosine content, LTSLS- Literature based Specific Loop Sequences, CER- Complementary End Residues, CRO- Choose Right Option, DNM- Does Not Matter, SF- re-execution with Same File, NF- re-execution with New File , Y- Yes, N- No.

Specific Loop Sequences (SLS)

This option, initially, prompts the user to enter the stem length. Further, it asks the user to either choose a loop sequence from a set of literature based SLS or define a new SLS. The literature based SLS are TTAA, TTCG, CCACC, CTCGAG, AAGCUU, CCACACC, TTCAAGAGA, AAGTTCTCT, AAGTTCTCT, TTTGTGTAG, GAAGCTTG, CTTCCTGTCA, TCAAGAG, GTGTGCTGTCC, TTCAAGAAC, TTGTGAGA **Table 2 (see supplementary material)**.

In addition, this option also prompts the user to enter GC content range as an additional parameter. After a successful execution, the software displays the total number of shRNAs predicted and the name of the generated output file. **Figures 1 & 2** show the illustration and flowchart of the software.

Output

The output file is created in the directory where the input file exists. It is a tab delimited text file containing the shRNA information, *viz.* stem length, loop length, shRNA sequence, position of shRNA in the input sequence, GC content, Tm, shRNA score, presence of polyAs and PolyCs in both positive and negative strands, number of complimentary nucleotides in loop sequence, accession number of the mRNA, gene ID and definition of the mRNA sequence. All the columns of the output can easily be imported in excel, access and MySQL databases. The program also generates a log file containing the sequence accessions having no shRNA complements.

Caveat and future development:

In the next version, web-based software will be developed with an additional module on prediction of shRNAs from all possible stem and loop length combinations.

Acknowledgement:

This study was supported by World Bank Funded - National Agricultural Innovation Project (NAIP), ICAR Grant NAIP/Comp-4/C4/C-30033/2008-09.

References:

[1] Dykxhoorn DM *et al. Nat Rev Mol Cell Biol.* 2003 **4**: 457 [PMID: 12778125]

[2] Sarnova L *et al. J Negat Results Biomed.* 2010 **9**: 8 [PMID: 20939886]

[3] Wallace RB *et al. Nucleic Acids Res.* 1979 **6:** 3543 [PMID: 158748]

[4] Gong W *et al. BMC Bioinformatics.* 2006 **7:** 516 [PMID: 17129386]

[5] Wang L & Mu FY, *Bioinformatics.* 2004 **20:** 1818 [PMID: 15001477]

[6] Hardin CC *et al. Biochemistry.* 1992 **31:** 833 [PMID: 1731941]

[7] Zhou H *et al. Informatica.* 2006 **30:** 357

[8] Reynolds A *et al. Nat Biotechnol.* 2004 **22:** 326 [PMID: 14758366]

Edited by P Kangueane

Citation: Singh *et al.* Bioinformation 8(13): 629-633 (2012)

BIOINFORMATION

open access

Supplementary material:

Table 1: Stem and loop length combinations with the corresponding references

Stem length	Loop length	Reference
29	4	Li *et al.* **[1]**
19	9	Kim *et al.* **[2]**; Li *et al.* [1]; Terasawa *et al.* **[3]**
19	10	Terasawa *et al.* **[3]**; Ge *et al.* **[4]**
21	9	Miyagishi *et al.* **[5]**
25	10	Ge et al. **[4]**
19	4	Li *et al.* **[1]**
29	9	Li *et al.* **[1]**
27	4	Siolas *et al.* **[6]**

Table 2: Specific loop sequences with the corresponding references

Specific Loop Sequence	Reference
TTAA	Khomyakova *et al.* **[7]**
TTCG	Lee *et al.* **[8]**
CCACC	Paul *et al.* **[9]**
CTCGAG	Editorial, Nature Cell Biology **[10]**
AAGCUU	Editorial, Nature Cell Biology **[10]**
CCACACC	Paul *et al.* **[9]**
TTCAAGAGA	Yu *et al.* **[11]**
AAGTTCTCT	Brummelkamp *et al.* **[12]**
TTTGTGTAG	Galy *et al.* **[13]**
GAAGCTTG	http://www.genelink.com/sirna/shrnai.asp **[14]**
CTTCCTGTCA	Galy *et al.* **[13]**
TCAAGAG	http://www.genelink.com/sirna/shrnai.asp **[14]**
GTGTGCTGTCC	Tanaka *et al.* **[15]**
TTCAAGAAC	Schopman *et al.* **[16]**
TTGTGAGA	Schopman *et al.* **[16]**

References for supplementary material:

[1] Li L *et al. RNA.* 2007 **13:** 1765 [PMCID: 1986814]
[2] Kim J *et al. J Biochem Mol Biol.* 2006 **39:** 329 [PMID: 16756763]
[3] Terasawa K *et al. J Nucleic Acids.* 2011 **2011**: 131579 [PMID: 21776374]
[4] Ge Q *et al. RNA.* 2010 **16:** 106 [PMCID: 2802021]
[5] Miyagishi M *et al. J Gene Med.* 2004 **6:** 715 [PMID: 15241778]
[6] Siolas D *et al. Nat Biotechnol.* 2005 **23:** 227 [PMID:15619616]
[7] Khomyakova EB *et al. FEBS Lett.* 1998 **422:** 265 [PMID: 9490021]
[8] Lee NS *et al. Nat Biotechnol.* 2002 **20:** 500 [PMID:11981565]
[9] Paul CP *et al. Nat Biotechnol.* 2002 **20:** 505 [PMID:11981566]
[10] Editorial *Nat Cell Biol.* 2003 **5**: 489 [PMID: 12776118]
[11] Yu JY *et al. Proc Natl Acad Sci U S A.* 2002 **99:** 6047 [PMID: 11972060]
[12] Brummelkamp TR *et al. Science.* 2002 **296:** 550 [PMID: 11910072]
[13] http://patentscope.wipo.int/search/en/WO2009083188.
[14] http://www.genelink.com/sirna/shrnai.asp.
[15] Tanaka H *et al. Mol Biotechnol.* 2009 **41**: 173 [PMID: 18821064]
[16] Schopman NC *et al. Antiviral Res.* 2010 **86:** 204 [PMID: 20188764]