*S*fold

Table of Contents

FUNCTION

***S*fold** predicts probable RNA secondary structures, assesses target accessibility, provides tools for the rational design of RNA-targeting nucleic acids, and predicts miRNA binding sites.

DESCRIPTION

***S*fold** is based on paradigm-shifting algorithms developed for RNA folding and prediction of target accessibility. Applications of the algorithms include rational design of RNA-targeting nucleic acids, and prediction of miRNA binding sites. The RNA folding algorithm generates a *statistical* sample of secondary structures from the Boltzmann ensemble of RNA secondary structures. From a statistical mechanics perspective, an RNA molecule can have a population of structures distributed according to a Boltzmann distribution, which gives the probability of a secondary structure I at equilibrium as $(1/U)exp[-E(I)/RT]$, where $E(I)$ is the free energy of the structure, R is the gas constant, T is the absolute temperature, and U is the partition function for all admissible secondary structures of the RNA sequence. The algorithm samples secondary structures *exactly* and *rigorously* according to the Boltzmann distribution, using Turner free energy rules.

General RNA folding features and output are available from the ***S*rna** module. Three design modules, ***S*irna**, ***S*oligo** and ***S*ribo**, provide tools for target-structure based design of short-

interfering RNAs (siRNAs), antisense oligonucleotides (oligos), and *trans*-cleaving hammerhead ribozymes, respectively. The last module ***S*TarMir** predicts miRNA binding sites on target RNA.

***S*fold** is described in its Wikipedia page at https://en.wikipedia.org/wiki/Sfold. ***S*fold** GitHub repository is located at https://github.com/Ding-RNA-Lab/Sfold.

*S*FOLD INSTALLATION AND RUNNING

*S*fold installation

***S*fold** only runs on Unix systems. Linux systems work well. Installation is straightforward. The user will need *R*, a statistics package and *Perl* installed on the system. Most Linux systems have both installed.

- Go to the Sfold GitHub page and click on the green 'Code' button. Choose the zip archive option and download the zip file.
- Move the zip file to the directory you wish to be the root of the Sfold installation and extract the zip archive.
- A directory called 'Sfold-main will be created, use '`cd Sfold-main`' to enter that directory.
- When the user lists (ls) the directory, a 'bin' directory, a small set of other directories and several text files will be seen. The user should read the '`README`' and the '`RUNNING_SFOLD`' file carefully.
- There is also a file '`configure`'. This is a utility that will configure your Sfold installation to run on your system. The user runs it using the command '`./configure`. The './ is not optional without it most systems will not run the script. A list of various tests will scroll by, and if the script is able to find everything it needs Sfold is installed. If it fails, the most typical reason is that *R* is not installed.
- The user must always run Sfold located in `Sfold-main/bin` for example `/researchtools/Sfold-bin/sfold`. It uses its own location to find the tools it needs to run.

Running of *S*fold

- As stated above, Sfold should always be invoked from the installation directory, and it would be safest to use the absolute path (starts with '/') to the executable.
- The minimum Sfold command, which corresponds to the Srna module on the website is `Sfold-main/bin/sfold -o myoutputdir/myseq.fa`. This would fold the sequence `myseq.fa`, perform clustering, and write the output file to `Sfold-main/bin/myoutdir/`. Note: this is why absolute pathnames should be used, relative pathnames can cause output to appear in surprising locations.
- There are other options that can be set. The full set of options is:
- -a <0 or 1> Run clustering on the sampled ensemble [default=1]
- -f <string> Name of file containing folding constraints

(Syntax follows UNAfold (Markham and Zuker 2008)
[default=no constraint]

- -h Display this information
- -l <+ve integer> Maximum distance between paired bases [default=no limit]
- -m <string> Name of file containing the MFE structure in GCG connect format. If provided, Sfold clustering module will determine the cluster to which this structure belongs.
- -o <string> Name of directory to which output files are written. Directory will be created if it does not already exist. Existing files will be overwritten. [default=$basedir/output]
- -w <+ve integer> Length of antisense oligos [default=20]
- -e Do not obliterate sample.out to save space [default=do]
- -i <0,1,2,3> 1=do Sirna, 2=do Soligo, 3=both, 0=neither [default=0]

Module-specific flags and constraints are described in the section below for duplicating the input options of the Sfold web service.

MODULE SPECIFIC INSTRUCTIONS AND OUTPUT

*S*rna

This module provides tools and statistics to statistically characterize the Boltzmann ensemble through the sampled structures. By default, the clustering of structures is on. This provides much useful information, so there is no reason to turn it off.

Flags

All of the Sfold flags are available. To use this module, the following are the most common:

- -o The location of the output directory. All files will be written to this Directory.
- -i 0 Use neither the SiRNA functionality nor the Soligo functionality.
- -l <num> Maximum distance between paired bases.

The input is a file containing a single mRNA in FASTA format. If the '-f' flag is used, a file of folding constraints can be included.

Output files

File `sclass.out` provides complete information for all structural clusters.

File `2dhist.out` contains base pair frequencies for constructing *2Dhist*. The first and second column are positions of a base pair; the third column is the number of occurrences in the sample; the last column is the size of the sample (i.e., number of structures generated).

File `fe.out` gives free energies (in kcal/mol, column 2) for all sampled structures.

File `cdf.out` is used for constructing the free energy CDF plot.

File `pdf.out` is a density version of `cdf.out`. It gives the probability with which structures in the sample will fall into an interval of width 5% with respect to the SMFE. The intervals are, (0%, 5%], ..., (90%, 95%], and (95%, 100%]. The probability of structures with SMFE is computed and is listed in column 2 in line 1 of `pdf.out`. Starting from line 2, the first column is the upper bound percentage of each interval, and the second column is the associated probability.

_S_irna

Design methodology. The *S*fold siRNA design method is based on factors supported by scientific evidence. More specifically, for siRNA screening, *S*fold combines target accessibility prediction, siRNA duplex thermodynamics rules, typical design rules and the empirical rules reported by Reynolds *et al*. (2004). Target accessibility evaluation is a unique feature of *S*fold and is expected to improve the chance of success. By integrating target accessibility evaluation, thermodynamic properties and sequence features for siRNA duplexes, *S*irna provides a unique combination of tools for siRNA design.

Scoring of siRNAs. *S*fold computes a total score of predicted siRNA potency. The total score is the sum of target accessibility score, duplex sequence feature score and duplex thermodynamics score.

Using the Sirna functionality will cause Sfold to generate information for designing small interfering RNAs.

All Sfold options and flags are available. The default for Sirna is to accept the defaults for all except that (1) '-o' should be set to the output directory, and (2) '-i' should be set to '1'.

There must be a file containing a single RNA sequence in FASTA format. It is possible to include a second file containing folding constraints, if the '-f' parameter is used.

In addition to the Srna output, the files `sirna_f.out`, `sirna_s.out`, `sirna.out`, `stability.out` and `Dharmacon_therm.out` will be created by Sfold running with -i 1 for Sirna. To obtain the disruption energy, an extra column in the output, it will be necessary to run a helper application

Obtaining disruption energy

- Confirm that `sirna.disrupten.pl` and `filter.sirna.disrupten.pl` are both present in the `Sfold-man/bin` directory of the distribution that was downloaded from the Sfold GitHub repository.
- Confirm that both of the scripts above have been set to executable.
- Locate the directory where the Sfold results were output. The user needs to run the script in that directory to confirm that `filter.sirna.disrupten.pl` can find the Sfold energy and base pairing files it needs to run.
- The script to run is `filter.sirna.disrupten.pl`. It will locate and run `sirna.disrupten.pl` during its execution.
- The script should be run with its full pathname so that the helper scripts it needs can be located. Below is an example run:
    - Assume the data from the Sfold run was output to `/data/sfold/run1/`
    - Run the command `cd /data/sfold/run1/`
    - Assume Sfold has been installed to `/programs/Sfold/Sfold-main/`
    - Run the command `/programs/Sfold/Sfold-main/filter.sirna.disrupten.pl`
      The user needs to provide at a minimum the following options
        - -c -p -- deal with overlapping sites and un-useful sites
        - -i <sirna.out file>
        - -s <sstrand.out> file containing single stranded probabilities
        - -l <length of sirna>
        - -o <file to output results to>

An example command line is

```
/home/williamrennie/Sfold-main/bin/filter.sirna.disrupten.pl -c
-p -i ./sirna.out -s ./sstrand.out -l 19 -o sirna_de.out
```

Using the *-h* option will print a help message which lists the other available options.

Output files

Implementation of target accessibility rule. The probability profile displays predicted accessible sites on the target RNA. Because an accessible site can be targeted by a number of siRNAs, selection of the "optimal" one can be based on binding energy of the antisense strand, together with other design rules. Stronger binding is indicated by smaller binding energy (stacking energies are *negatively valued*). For example, an antisense siRNA with a binding energy of -15 kcal/mol is predicted to be more effective than an antisense siRNA with a binding energy of -10 kcal/mol. The antisense siRNA binding energy is a weighted sum of the RNA/RNA stacking energies (Xia *et al.* 1998) for the hybrid formed by the antisense siRNA and the targeted sequence. For a base-pair stack, the weight for the sum is calculated by the probability of the unpaired dinucleotide in the target sequence that is involved in the stack. In addition, A-U

terminal penalty is included and is weighted by the probability of the unpaired terminal base. This weighting scheme accounts for the structural variation at the target site. The target accessibility rule is implemented by requiring the siRNA binding energy to be below a threshold value. The current default of the threshold is -10 kcal/mol.

siRNA duplex thermodynamics. ***S*irna** computes a number of thermodynamics indexes for the implementation of rules on siRNA duplex stabilities, based on recent RNA thermodynamics parameters (Xia *et al.* 1998; Mathews *et al*. 1999). 5'-antisense stability (AntiS, in kcal/mol) is computed by a sum of free energies for four base pair stacks and the 3' dangling T and a penalty for terminal A-U for the 5' end of the antisense siRNA strand; 5'-sense stability (SS, in kcal/mol) is the sum for the 5' end of the sense siRNA strand. Differential stability of siRNA duplex ends (DSSE, in kcal/mol) is the difference between the 5'-antisense stability and the 5'-sense stability, i.e., DSSE=AntiS-SS. For each of positions 2-18 of the antisense strand, the internal stability is the sum of 4 base pair stacks, starting at this position in the 5' → 3' orientation. For position 1, the internal stability is the 5'-antisense stability. For position 19, the internal stability is the 5'-sense stability. The internal stabilities are used for constructing an internal stability profile for each siRNA duplex. For positions 16-19 on the 3' end of the antisense strand, Khvorova and colleagues (Khvorova *et al.* 2003) extended the target sequence with the target RNA for the purpose of calculation. This treatment can lead to inaccurate information. For example, the profile is not guaranteed to be symmetric when the bases for the ends of the duplex are symmetric. For a correct comparison of the stabilities for the two siRNA duplex ends, we simply reverse the orientation to 3' → 5' in the calculation for positions 16-19. Average internal stability at the cleavage site (AIS, in kcal/mol) is the average of internal stability values for positions 9-14 of the antisense strand (Khvorova *et al.* 2003).

Filtering and scoring. Filters and a number of siRNA scores are used by ***S*irna** for siRNA output files. The filters are described in file headers. Target accessibility score is 0 if antisense siRNA binding energy > -2 (kcal/mol); the accessibility score is k (k=2,…, 7), if $-2k \leq$ binding energy $< -2(k+1)$; the accessibility score is 8 if the binding energy < -16. The siRNA duplex feature score is computed with the algorithm by Reynolds *et al.* (2004) and has a minimum of -2 points and a maximum of 10 points. The duplex thermodynamics score can have value 0, 1 or 2, with 1 point contributed by DSSE > 0 (kcal/mol), and another point by AIS > -8.6 (kca/mol). The total siRNA score is the sum of accessibility score, duplex feature score and duplex thermodynamics score. The maximum total score is 20 points.

Total duplex stability, sum of probabilities of unpaired bases of the target sequence, and the dinucleotide leader preceding the target sequence are also included in the output files. The user has the option to consider dinucleotide leader motifs such as AA or NA, although there is a lack of evidence to support the significance of the leader for siRNA function.

File `sirna_de.out` presents binding site disruption energies (Shao *et al* Roninson and Ding 2007)

File `sirna_f.out` gives output for siRNAs that meet all filter criteria:

Line 1:
Column 1: target position (starting - ending)
Column 2: sense siRNA (5' → 3')
Column 3: antisense siRNA (5' → 3')
Column 4: dinucleotide leader preceding the target sequence

Line 2:
Column 1: total score for siRNA duplex
Column 2: target accessibility score
Column 3: duplex feature score
Column 4: duplex thermodynamics score
Column 5: siRNA GC content
Column 6: antisense siRNA binding energy (kcal/mol)
Column 7: differential stability of siRNA duplex ends (DSSE, in kcal/mol)
Column 8: average internal stability at the cleavage site (AIS, in kcal/mol)
Column 9: total stability of siRNA duplex (kcal/mol)
Column 10: sum of probabilities of unpaired target bases
(column 4 of output file `sstrand.out`)

Filter criteria:
A) Antisense siRNA binding energy ≤ -10 kcal/mol (target accessibility rule);
B) Duplex feature score of 6 or higher;
C) DSSE > 0 kcal/mol (asymmetry rule);
D) AIS > -8.6 kcal/mol (cleavage site instability rule);
E) 30% ≤ GC % ≤ 60%;
F) Exclusion of target sequence with at least one of AAAA, CCCC, GGGG, or UUUU.

Notes:
1) The starting (ending) position of the target sequence corresponds to position 19 (1) of the antisense siRNA (i.e., dinucleotide leader and nt 22 and nt 23 in Tuschl patterns are not considered by us to be part of the target sequence);
2) Sense siRNA=target sequence + 3' dTdT overhang; dTdT for both sense and antisense siRNAs can be replaced by UU;
3) GC % = GC count in siRNA (excluding overhangs)/19 x 100%;
4) DSSE = stability of 5'-antisense end of 4 base pairs - stability of 5'-sense end of 4 base pairs; the asymmetry rule is enforced by DSSE > 0 (see Schwartz *et al. Cell*, **115**, 199-208, 2003).
5) AIS = average of internal stability values for positions 9-14 of the antisense strand; starting at a position, the internal stability is for 4 BP stacks; the rule of relative instability at the cleavage site is enforced by AIS > -8.6 kcal/mol, the midpoint between the minimum of -3.6 and the maximum of -13.6 (see Khvorova *et al*. *Cell*, **115**, 209-216, 2003).
6) Total siRNA duplex score is the sum of target accessibility score, duplex feature score and duplex thermodynamics score, with a maximum of 20 points; the accessibility score is based on antisense siRNA binding energy and has a range of [0, 8]; the duplex feature score is computed with the algorithm by Reynolds *et al*. (*Nature Biotech*., **22**, 326-330, 2004), and has a range of

[-2, 10]; the duplex thermodynamics score has a range of [0, 2], with contribution of 1 point for DSSE > 0, and 1 point for AIS > -8.6 kcal/mol.

File `sirna_s.out` provides output information for siRNAs with total score greater or equal to a preset threshold. The current threshold is 12 points.

File `sirna.out` contains output information for all siRNAs.

File `stability.out` gives output for siRNA ends and internal stabilities:

Line 1: target position antisense siRNA (5' → 3')
5'-antisense stability (AntiS, in kcal/mol)
5'-sense stability (SS, in kcal/mol)
differential stability of siRNA duplex ends (DSSE, in kcal/mol)
average internal stability at the cleavage site (AIS, in kcal/mol)
Line 2: internal stability for antisense positions 1-10
Line 3: internal stability for antisense positions 11-19

Notes:
1) AntiS is computed by a sum of energies for 4 base pair stacks and the 3' dangling T for the 5' end of the antisense siRNA strand; SS is the sum for the 5' end of the sense strand;
2) DSSE = AntiS-SS; the symmetry rule is enforced by DSSE > 0 (see Schwartz *et al.* 2003);
3) AIS = average of internal stability values for positions 9-14 of the antisense strand; starting at a position, the internal stability is for 4 BP stacks; the rule of relative instability at the cleavage site is enforced by AIS > -8.6 kcal/mol, the midpoint between the minimum of -3.6 and the maximum of -13.6 (see Khvorova *et al.* 2003).

File `Dharmacon_thermo.out` presents siRNA duplex features proposed by Dharmacon (Reynolds *et al*. 2004) and duplex thermodynamics.

File `sstrand.out` contains information for probability profiling and for probability-weighted calculations for antisense siRNA binding energy and antisense oligo binding energy:

Column 1: nucleotide position *i*
Column 2: nucleotide
Column 3: complementary nucleotide
Column 4: the probability that nucleotide *i* is unpaired (i.e., *W*=1)
Column 5: probability that dinucleotide *i* and *i*+1 are both unpaired (i.e., *W*=2)
Column 6: the probability that nucleotide *i*, *i*+1, *i*+2, and *i*+3 are *all* unpaired (i.e., *W*=4)

Note:
Column 5 is used for probability weighted calculations of antisense siRNA binding energy and antisense oligo binding energy.
Column 6 is used for probability profiling for single-stranded fragments of 4 bases.

File `loopr.out` contains information for probability profiling of loops:

Column 1: nucleotide position
Column 2: nucleotide
Column 3: the probability that this nucleotide is in a hairpin loop
Column 4: the probability that this nucleotide is in a bulge loop
Column 5: the probability that this nucleotide is in an interior loop
Column 6: the probability that this nucleotide is in a multi-branched loop
Column 7: the probability that this nucleotide is in the exterior loop
Column 8: sum of columns 3 through 7 (this is the same as column 4 of file *sstrand.out)*

*S*oligo

Soligo provides, in addition to structural and energy information, information valuable for designing antisense oligos.

All Sfold flags are available. Soligo recommends that if the sequence is from a procaryote, the max distance between bases "-l" is set to 50, to address local target folding. A file of other folding constraints can be passed to Sfold with the '-f' flag.

An example command line is:

```
<path to sfold>/sfold -l 50 -i 2 myseq.fa -o <sfold output
directory>
```

If the user wishes to obtain the disruption energy output oligo, a helper script is needed and is described below.

- Confirm that `soligo.disrupten.pl` and `filter.soligo.disrupten.pl` are both present in the `Sfold-man/bin` directory of the distribution that was downloaded from the Sfold GitHub repository.
- Confirm that both of the scripts above have been set to executable.
- Locate the directory where the Sfold results were output to. The user needs to run the script in that directory to confirm that `filter.soligo.disrupten.pl` can find the Sfold energy and base pairing files it needs to run.
- The script to run is `filter.sirna.disrupten.pl`. It will locate and run `sirna.disrupten.pl` during its execution.
- The script should be run with its full pathname so that the helper scripts it needs can be located. Below is an example run.
    - Assume the data from the Sfold run was output to `/data/sfold/run1/`
    - Run the command `cd /data/sfold/run1/`
    - Assume Sfold has been installed to `/programs/Sfold/Sfold-main/`

- Run the command `/programs/Sfold/Sfold-main/filter.soligo.disrupten.pl`
  The user needs to provide at a minimum the following options
  - -c -p -- deal with overlapping sites and unuseful sites
  - -s <sstrand.out> file containing single stranded probabilities
  - -l <length of desired oligo>
  - -o <file the results should be sent to>

An example command line is

```
/home/williamrennie/Sfold-main/bin/filter.soligo.disrupten.pl -c
-p -s ./sstrand.out -l 20 -o oligo_de.out
```

Output files

The probability profile displays predicted accessible sites on the target RNA. Because an accessible site can be targeted by a number of antisense oligos, selection of the "optimal" one can be based on binding energy, together with other empirical rules such as GC content, avoidance of GGGG (or more stringent GGG) motifs, etc. Stronger binding is indicated by smaller binding energy (stacking energies are *negatively valued*). For example, an antisense oligo with a binding energy of -10 kcal/mol is more effective than an oligo with a binding energy of -5 kcal/mol. The antisense oligo binding energy is a weighted sum of the DNA/RNA stacking energies (Sugimoto *et al*. 1995) for the hybrid formed by the antisense oligo and the targeted sequence. For a base-pair stack, the weight for the sum is calculated by the probability of the unpaired dinucleotide in the target sequence that is involved in the stack. This weighting scheme accounts for the structural variation at the target site among the structures in the sample.

File `oligo_f.out` gives filtered output for design of antisense oligos:

Column 1: target position (starting - ending)
Column 2: target sequence (5' → 3')
Column 3: antisense oligo (5' → 3')
Column 4: GC content
Column 5: oligo binding energy (kcal/mol)

Filter criteria:
A) 40% ≤ GC % ≤ 60%;
B) Antisense oligo binding energy ≤ -8 kcal/mol;
C) No GGGG in the target sequence.

File `oligo.out` gives complete output for design of antisense oligos:

Column 1: target position (starting - ending)
Column 2: target sequence (5' → 3')

Column 3: antisense oligo (5' → 3')
Column 4: GC content
Column 5: oligo binding energy (kcal/mol)
Column 6: GGGG indicator

Note:
GGGG indicator=1 for at least one GGGG in the target sequence; indicator=0 otherwise.

File `oligo_de.out` presents binding site disruption energies (Shao *et al* Roninson and Ding 2007).
Column 1: starting target position
Column 2: ending target position
Column 3: target sequence (5p --> 3p)
Column 4: antisense oligo (5p --> 3p)
Column 5: GC content
Column 6: average unpaired probability for target site nucleotides
Column 7: binding site disruption energy (kcal/mol)

Filter criteria ("<=": less than or equal to; ">=": greater than or equal to):

A) 40% <= GC % <= 60%;
B) No GGGG in the target sequence;
C) Average unpaired probability for target site nucleotides >= 0.5;
D) For each peak in the accessibility profile that is above the threshold probability of 0.5, all sites targeted to this same peak are ranked by their average unpaired probability (the higher the better) and at most n sites are selected for each peak, where n is determined by max ([width of peak/site length], 2);
E) Among sites satisfying criteria A-D, the top 20 unique ones with the highest average unpaired probability are listed.

Files `sstrand.out` and `loopr.out` are the same as described for ***S*irna**.

***S*ribo**

Sribo outputs information for potential cleavage sites for hammerhead ribozymes. It is quite computationally intensive and will take some time to run.

Input files should be specified using absolute path names (although the script was designed to run with relative pathnames, occasional failures have been observed).

The executable should also be specified using an absolute path name, this facilitates the script locating its helper scripts. Steps for running Sribo are the following:

- Confirm the four ribozyme scripts. `rz-dghybrid.pl`,`rz_dgswitch.pl`, `parse_lib.pl` and `rz_energies.pl` are in the `Sfold-main/bin` directory that was downloaded from the Sfold GitHub repository. Confirm they are marked as executable.
- Run Sfold on the target RNA sequence. The default options should be sufficient. an example command line is
    - `<path to Sfold>/sfold myseq.fa -o <Sfold output directory>`
    - Make note of the sfold output directory
- The script `rz_energies.pl` is the top-level script. It will call the other scripts and Sfold as it runs.
- Script options
    - -m <target sequence> in FASTA format
    - -s <sfold result directory>/bp.out
    - -f <sfold result directory>/fe.out
    - -a <length of Helix III>
    - -b <length of Helix I>
    - -t <triplet> one of the NUH triplets
    - -o <file the results should be written to>
- An example command line is

```
Sfold-main/bin/rz_energies.pl -m myseq.fa -s
SfoldResult/bp.out -f Sfold/result/fe.out -a 9 -b 11
-t GUC -o ribo_energies.out
```

Output files

File `rz_energies.out` provides computed energies for ribozyme cleavage site:

Column 1: site ID (only those sites with long enough arms will be shown)
Column 2: target starting position
Column 3: target ending position
Column 4: dG_total = dG_hybrid - dG_switch - dG_disrupt (kcal/mol)
Column 5: dG_hybrid (kcal/mol)
Column 6: dG_switch (kcal/mol)
Column 7: dG_disrupt (kcal/mol)
Column 8: target site sequence (5p --> 3p)
Column 9: ribozyme sequence (5p --> 3p)

Files `sstrand.out` and `loopr.out` are the same as described for ***S*irna**. The user can examine column 4 of `sstrand.out` for prediction of other cleavage triplets.

***S*TarMir**

STarmir runs it on the Ubuntu distribution, version 18.04 or newer. Unbuntu's tutorial for installation is available at https://ubuntu.com/tutorials/install-ubuntu-desktop#1-overview.

The zip archive for STarMir can be downloaded from https://github.com/Ding-RNA-Lab/Sfold, by selecting the green "Code" button. Experienced GitHub users can also "fork" the archive. The zip file should be saved in a directory where the user plans to install the archive. The zip file can be extracted using an unzip utility tool. After unzipping, a directory called "Sfold-main" is created containing the components of the Sfold package including the STarMir program.

Installation of other required software and utilities

RNAhybrid

RNAhybrid (Rehmsmeier *et al* 2004) is used by STarMir to create a set of candidate binding sites. It can be downloaded from https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid. Download and installation instructions are available on this site.

R

R is a statistical package (https://www.r-project.org) used by STarMir for executing the prediction models. R is usually available on most Linux systems and can be downloaded and installed using the Linux distributions package manager. The installed executable must be globally accessible on the host computer.

Perl

The bulk of the STarMir code was written in Perl. Version 5 or newer is required. Perl is usually pre-installed on any Linux system. Two Perl modules used by STarMir, Bio::Seq and Bio::SeqIO, must be installed, typically using CPAN. The main Bioperl installation page provides recommendations and links to the main CPAN page. The URL is https://github.com/bioperl/bioperl-live/blob/master/README.md. The CPAN package is included with the Perl installation.

Shell

STarMir can be run on the command line within a console window. It was developed using bash and thus should run with any shell.

Configuration

Sfold

To configure sfold, the user must enter the '`Sfold-main/bin`' directory. This directory contains the Sfold executable, and a document called '`Running_Sfold`', which provides instructions for configuring and running Sfold. The document also

explains how to use the testing utility to confirm that the Sfold package is installed correctly.

STarMir

STarMir does not have an automated configuration utility. The user must manually edit a few parameter files located in the '`Sfold-main/STarMir`' directory. The README file in that directory contains relevant information.

The file '`starmir_param.pl`' needs to be edited. The user needs to set the `$RNA_bindir` path to the location of the RNAhybrid executable. Typing "which RNAhybrid" on the command line will display the bin directory of the RNAhybrid executable. Only the directory (path) is needed, not the program name.

The user must also set the path to `disruptEn`, a program that calculates the free energy required to open a local structure. This program is part of the Sfold package, and the binary can be found in the `Sfold-main/bin` directory. To ensure proper functionality, use the full path, such as: "`/home/williamrennie/development/Sfold_main/bin/`". The final slash is important and must be included. Additionally, the line "`$SFOLDBIN="/home/bill/Desktop/Sfold-main/bin/"` must be set to the same `Sfold-main/bin` directory specified in the previous above.

The steps above complete the installation and configuration of the programs required for STarMir.

Procedures for executing STarMir

Executing Sfold

Sfold must be run first for predicting target structures. Running Sfold is straightforward, and the default parameters are sufficient for STarMir to use predicted structures. It has many configuration options which can be viewed by running Sfold without any arguments. Sfold should always be executed using the full path to the executable, for example "`/home/bill/Sfold-main/sfold`".

- The input to Sfold is a file containing a single RNA sequence in FASTA format.
- Sfold produces a directory of output files, some of which are required by STarMir. The user shall select the directory, which need to be passed to STarMir.
- An example command to execute Sfold is "`/home/williamrennie/Sfold-main/bin/sfold -o myoutputdir myseq.fasta`"
- Running Sfold can take anywhere from a few minutes to a few hours depending on both the computational power of the host computer and the length of the target sequence.

Executing STarMir

• STarMir is a system of Perl scripts and helper application that predict and rank miRNA binding sites on a target mRNA.
• The main command script is located in the distribution's `Sfold-main/STarMir` directory of the distribution. The user can run the program through that script.
• The script MUST be executed from the directory that contains it. Otherwise, it may not find all the necessary supporting scripts and may not run correctly.
• The main command script "`starmir_research.pl`", requires nine arguments. Arguments that point to files MUST give the full path to the file (absolute path). The code has no facility for deducing the path to the file. The arguments, in the order they appear, are the following:

Required arguments (in order)

1. <working directory>: The directory where output files will be stored. By default, the intermediate output files will also be saved here. This argument must end with a backslash (\) to set an absolute path to the directory.

2. <miRNA file>: A file containing one or more miRNA sequences in FASTA format (absolute path).

3. <mRNA file>: A file containing a single mRNA sequence in FASTA format (absolute path).

4. <sfold output directory>: The directory containing Sfold results for the mRNA sequence can be found (absolute path, must end with a forward slash (/).

5. <target species>: The species name used by RNAhybrid. Must be one of fly, human or worm (these are only species supported by RNAhybrid).

6. <model species>: The species used by the STarMir prediction model. It must be one of human, mouse, or worm, STarMir predicts miRNA binding sites using models built for human (Homo sapiens), mouse (*Mus musculus*) and worm (*Caenorhabdites elegans*). These models were trained on V-CLIP data for human (Hafner *et al* 2010), HITS-CLIP data for mouse (Chi *et al* 2009), and ALG-1 CLIP data for worm (Zisoulis *et al* 2010). The human and mouse models were cross-validated and can be broadly used for other species [12].

7. <CDS start>: The start of the coding region.

8. <CDS end>: The end of the coding region.

Optional argument

An optional ninth argument can be either 1 or 0. Setting this argument to 1 deletes all the intermediate files, whereas the default (0) preserves the intermediate files.
The command line, which must be run in the same directory as the scripts is:

```
<path to script directory>/starmir_research.pl <working
directory> <miRNA file> <mRNA file> <sfold output
directory> <target species> <model species> <CDS start>
<CDS end> <optional 1 for deleting intermediate files>
```

For best practice, use absolute path names for all input files and directories. Below is an example of STarMir command line run with intermediate files deleted:

```
./starmir_research.pl Data/myseq mirnas.fasta myseq.fasta
~/runs/sfoldDataMyseq/ human human 234 874 1
```

Output files

A successful local installation and execution of STarMir will generate separate final output files for miRNA binding sites in the 5′ UTR, the coding region and the 3′ UTR, and for both seed and seedless sites. These output files are prefixed with 'Final-'. If no prediction is made for a specific site, e.g.., in the case of the lack of a single seed, the corresponding file will not be generated. For the seed output file, several seed-specific features (e.g., Seed_Access for seed accessibility) are provided. In each file, the binding sites are listed in the descending order of their logistic probabilities. In addition, a file containing miRNA:target hybrid conformations in text format is generated.

Publication quality hybrid diagrams in PDF format can also be produced using the "`create_PDF.pl`" script, located in the STarMir subdirectory of the Sfold GitHub depository. The input file for this script is "`Total-En-Hyb-Fil-`*`mRNA id`*`.out`" file (where "mRNA id" corresponds to the target name). This file will be located in the output directory specified during the execution of STarMir. The resulting PDF file will be named "`SiteX.pdf`", where X refers to the specific site number provided for the `create_PDF.pl` script. For example, the command "`./create_pdf.pl ../../OutputDir2/TotalEn-Hyb-Fil-NM_017589.4.out 3`" will generate the PDF diagram for site 3, saved as "`Site3.pdf`" in the same directory in which the command was executed.

For each binding site, STarMir provides a comprehensive list of site features. Several are unique to STarMir: structure-based free-energy measures, a logistic probability, and a score for the miRNA:target pair. The logistic probability is a measure of confidence for a predicted site. The score is a measure of predicted regulatory efficacy of the miRNA on the target, based on a linear combination of the contributions from both seed and seedless binding sites. Although 8-mer sites are often considered the most effective among all types of sites, they do not necessarily ensure high scores which are predictive of effective regulation. On the other hand, large numbers of seedless sites can influence the combined score, potentially leading to a strong regulatory impact.

Please see Rennie *et al* (2025) for output examples (Springer chapter preprint Sfold_Chapter_2025.pdf is in the Sfold GitHub repository).

REFERENCES

In research publications, the users of ***S*fold** are requested to cite appropriate articles. For ***S*rna**, please cite Ding and Lawrence 2003, and Ding and Lawrence 2005; for ***S*irna**, please cite Shao *et al* Roninson and Ding 2007; for ***S*oligo**, please cite Ding and Lawrence 2001, and Shao *et al* 2006; for ***S*ribo**, please cite Shao *et al* Schneider and Ding 2007; for ***S*TarMir**, please cite Long *et al* 2007 and Liu *et al* (2013).

References for ***S*fold**

William Rennie, Shaveta Kanoria, Jun Lu and Ye Ding (2025, to appear). Sfold Tools for microRNA Target Prediction. *Methods Mol Biol.*, Springer Protocols.

Liu, C., Mallick, B., Long, D., Rennie, W.A., Wolenc, A., Carmack, C.S., and Ding, Y. (2013) CLIP-based prediction of mammalian microRNA binding sites. *Nucleic Acids Res*, **41** (14), e138

Long, D., Lee, R., Williams, P., Chan, C.Y., Ambros, V., and Ding, Y. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.* **14**, 287-294.

Shao, Y, Wu, S., Chan, C.Y., Klapper, J.R., Schneider, E., and Ding, Y. (2007) A structural analysis of *in vitro* catalytic activities of hammerhead ribozymes. *BMC Bioinformatics* **8**, 469.

Shao, Y., Anil, M., Chan, C.Y., Lawrence, C.E., Roninson, I., and Ding, Y. (2007) Effect of target secondary structure on RNAi efficiency. *RNA* **13**, 1631-1640.

Shao, Y., Wu, Y., Chan, C.Y., McDonough, K., and Ding, Y. (2006) Rational design and rapid screening of antisense oligonucleotides for prokaryotic gene modulation. *Nucleic Acids Res.* **34**, 5660-5669.

Ding, Y., Chan, C.Y. and Lawrence, C.E. (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* **11**, 1157-1166.

Ding, Y. and Lawrence, C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* **31**, 7280-7301.

Ding, Y., and Lawrence, C.E. (2001) Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond, *Nucleic Acids Res.* **29**, 1034-1046.

Other references

Markham, N and Zuker, M (2008) UNAFold: software for nucleic acid folding and hybridization**.** *Methods Mol Biol* **453**:3-31. doi: 10.1007/978-1-60327-429-6_1.

Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol*. **288**, 911-940.

Xia, T., SantaLucia, J. Jr, Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., Turner, D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719-35.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., Zamore, P.D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*. **115**, 199-208.

Khvorova, A., Reynolds, A., Jayasena, S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*. **115**, 209-216.

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., Khvorova, A. (2004) Rational siRNA design for RNA interference. *Nat Biotechnol*. **22**, 326-30.

Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. *RNA* **10** (10):1507-1517. doi:10.1261/rna.5248604

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr., Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141** (1):129-141. doi:10.1016/j.cell.2010.03.009

Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460** (7254):479-486. doi:10.1038/nature08170

Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, Yeo GW (2010) Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. In: *Nat Struct Mol Biol*, **17** (2): 173-179. doi:10.1038/nsmb.1745

For all ***S*fold** users, we wish you luck and success in your scientific endeavors!

§§§ Final update by Ye Ding and William Rennie, April 1, 2025