

Sfold

Table of Contents

- FUNCTION
- DESCRIPTION
 - RNA folding and applications
 - Rules for siRNA design and **Sfold** design methodology
- INPUT
 - Common information for all modules
 - Module-specific input information
- OUTPUT
 - Sirna**
 - Soligo**
 - Sribo**
 - Srna**
- ALGORITHMS
- REFERENCES

FUNCTION

Sfold predicts probable RNA secondary structures, assesses target accessibility, and provides tools for the rational design of RNA-targeting nucleic acids.

DESCRIPTION

RNA folding and applications

Sfold is based on patent-pending algorithms developed by Ding and Lawrence (2001, 2002, 2003) for RNA folding, prediction of target accessibility, and rational design of RNA-targeting nucleic acids. 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 recent Turner free energy rules.

A focus of **Sfold** is on applications of prediction of target accessibility, and rational design of RNA-targeting nucleic acids. Three modules, **Sirna**, **Soligo** and **Sribo**, provide application tools for the design of short-interfering RNAs (siRNAs), antisense oligonucleotides (oligos), and

trans-cleaving ribozymes, respectively. General folding features and output are available from the fourth module, **Srna**.

Rules for siRNA design and Sfold design methodology

Empirical rules for siRNA duplexes. The following basic design rules are widely used: 1) siRNA duplexes should be composed of 21-nt sense and 21-nt antisense strands, paired so as to each have a 2-nt 3' dTdT overhang; 2) the siRNA sequence should have low to moderate GC content; 3) sequences with more than three Gs or three Cs in a row should be avoided, because polyG and polyC sequences may hyperstack and thus interfere in the siRNA silencing mechanism; 4) AAAA or TTTT should also be avoided for RNA polymerase III mediated promoters because transcription tends to terminate at these sequences. It is note that there is a lack of support in the literature for the significance of target patterns such as AA(N19) or NA(N19). Recently, empirical rules on sequence features of functional siRNAs have been reported by several groups (Reynolds *et al.* 2004; Ui-Tei *et al.* 2004; Amarzguioui and Prydz 2004). Based on the largest published siRNA data set, the rules by Reynolds *et al.* (2004) are the most comprehensive. However, these rules alone do not guarantee siRNA function.

Target secondary structure and accessibility. Single-stranded regions in RNA secondary structure are likely to be accessible for RNA-targeting nucleic acids through base-pairing interactions. Target accessibility has long been established as an important factor for the potency of antisense oligonucleotides (oligos) and *trans*-cleaving ribozymes. Recently, the importance of target structure and accessibility for the function of siRNAs has been demonstrated, using a number of experimental approaches that include oligo library (Lee *et al.* 2002; pp. 502, 503), oligo array (Bohula *et al.* 2003), antisense evaluation of accessibility (Far and Sczakiel 2003), and targeting the same sequence in both structured and unstructured sites (Vickers *et al.* 2003; p.7114, left column, top paragraph). A number of potent siRNAs in these studies do not meet key empirical rules. However, their function is explained by accessibility.

Based on the RNA folding algorithm, we also developed a probability profiling method for the prediction of target accessibility (Ding and Lawrence 2001, 2003). A stochastic approach to accessibility evaluation may be essential to account for the likely existence of a population of structures for mRNAs (Christoffersen *et al.* 1994). The probability profiling approach reveals target sites that are *commonly accessible* in a high proportion of statistically representative structures for the target RNA. Through assignment of statistical confidence in predictions, this approach bypasses the long-standing difficulty in accessibility evaluation due to limited representation of probable structures.

siRNA duplex thermodynamics. It has been reported that functional siRNA duplexes tend to have lower stability on the 5'-antisense end (about four base pairs) than on the 5'-sense end (Schwarz *et al.* 2003; Khvorova *et al.* 2003). It is proposed that both the absolute and relative stabilities of the siRNA duplex ends determine the degree to which each strand participates in the RNAi pathway (Schwarz *et al.* 2003). Furthermore, functional siRNAs tend to have relative instability at the cleavage site, which may facilitate product release and multiple turnovers (Khvorova *et al.* 2003). However, these rules on siRNA duplex stabilities do not guarantee siRNA function

(Khvorova *et al.* 2003), perhaps because they do not address the structure of the target mRNA. It appears to be a consensus view that after the duplex unwinding the antisense strand in the activated RISC needs to bind to the target sequence through complementary base-pairing for target recognition. This would explain the exquisite specificity by siRNAs. When the target sequence is in a heavily structured (i.e., helical) region, the large energy barrier will likely prevent the formation of the hybrid between the antisense siRNA strand and the target sequences. Zamore lab has shown that a single hydrogen bond can decide which of the duplex strands will be incorporated into RISC. Likewise, we believe that secondary structure, accessibility and thermodynamics at the target site are also important.

Sfold siRNA design methodology. The Sfold siRNA design method is based on factors supported by scientific evidence. More specifically, for siRNA screening, Sfold 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 Sfold and is expected to improve the chance of success. By integrating target accessibility evaluation, thermodynamic properties and sequence features for siRNA duplexes, Sfold provides a unique combination of tools for siRNA design.

Scoring of siRNAs. Sfold 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.

BLAST search for specificity. To avoid potential non-specific effects, the user should perform a nucleotide-nucleotide BLAST search on NCBI server (<http://www.ncbi.nlm.nih.gov/BLAST/>) against UniGene database of selected organism to minimize homology to non-targeted genes. Currently, the rules for siRNA specificity are not yet fully established. It is prudent to allow at least 2 mismatches between a siRNA and all non-target genes. The user needs to be aware of the possibility of multiple entries under the same or different names for the same gene. A BLAST search also needs to be performed for a scrambled control to ensure that it is not complementary to any gene in the target organism.

A note to all users: please do not design siRNAs or other RNA-targeting nucleic acids based on a *single* structure from module **Srna**. We believe that there may be an ensemble of structures for mRNAs and viral RNAs. Please refer to FAQ # 3 on **Sfold** front page.

INPUT

Sfold is available through web server at two sites: <http://sfold.wadsworth.org> and <http://www.bioinfo.rpi.edu/applications/sfold>. On the server home page, a user can select the application module of primary interest. From the output page for the selected module, links for output are provided for the other modules, with default input settings as described below.

Common information for all modules

Folding conditions. Folding temperature is 37°C. Ionic conditions are 1M NaCl, with no divalent ions.

Job mode and limits. A job can run in either interactive mode or batch mode. Current limits are 200 bases for an interactive job, and 5,000 bases for a batch job. For a batch job, a correct e-mail address is required, for notification of job completion.

For a sequence exceeding the length limit, the user needs to truncate the sequence for folding. It is quite possible that some portions of the secondary structure of a truncated sequence will be different from the structure of the full sequence. This in turn can change the predicted accessibility of some sites on the target. To computationally address this issue, we suggest two strategies:

(1) Centering. For a decision on where to truncate the sequence, we suggest that the region of primary interest be located far way from the sites of the truncation, e.g., near the center of the truncated sequence. Our computational experience suggests that the accessibility of a site is more likely to be maintained when the site is located near the center of the truncated sequence. This can be explained by the fact that, to a great extent, RNA secondary structure is determined by nearest-neighbor interactions. The regions near the 5' and 3' ends of the shortened sequence, i.e., the sites of truncation, are most likely to have different predictions for local structures and accessibility.

(2) Folding of two overlapping sequences. For a region X of primary interest, the user selects two truncated sequences 5'-A-X-B-3', and 5'-C-X-D-3', such that both contain X and both are within the length limit. After folding both sequences, the user selects common accessible sites in X that are predicted by the sequences. This process takes more work, however, the reward is increased confidence in accessible sites, because the location of truncation has little effect on the predicted accessibility.

Sequence input format. Sequences in raw format, in FASTA format, or in GenBank format are accepted. A sequence can be entered in the input window; alternatively, a sequence file can be uploaded from your computer directory (folder). Note that any character other than A, C, G, T, or U will be edited out. An option is provided if the RNA sequence to be folded is the reverse complement of the input sequence.

Module-specific input information

Sirna. The length of siRNAs with 3' TT overhangs is fixed at 21 nt.

Soligo. The user has the option to set the length of the oligos. The default length is 20 nt.

Sribo. The user can specify an NUH cleavage triplet for hammerhead ribozymes. The default triplet is GUC.

Srna. None.

OUTPUT

The output page for each module includes both graphical representations and relevant text files. Many of the graphics are interactive through a web browser, and the user can download a colored or black and white plot in PDF or PostScript format. For monochromic printers, the black and white version is recommended. All of the graphs, with the exception of RNA structure diagrams, are first generated by the public domain software Gnuplot and then post-processed by Perl scripts to provide additional features and to improve visualization. Links to output from other modules with default input settings are provided. Output can be accessed on the server for up to 72 hours after job completion.

Sirna

Graphical output

Probability profiling for visual display of accessibility prediction. For prediction of target accessibility, a complete probability profile of single-stranded regions is generated for the entire target RNA. Sites with high probabilities of being single-stranded are predicted to be accessible. At nucleotide position i , the profile shows the probability that nucleotides i , $i+1$, $i+2$, and $i+3$ are *all* unpaired. In other words, the profile is for consecutive fragments with a width W of 4 bases. Although the profile can be generated for any W , $W=4$ has been found to be particularly useful (Ding & Lawrence 2001). Currently, colored or black and white profile plots are available in PDF and PostScript formats.

For RNA sequences longer than 200 bases, the regional probability profile allows the user to examine either any region of 200 bases, by specifying the starting position, or an adjacent region of 200 bases, by clicking << or >>.

Loop profiling. Probability profiles for all types of loops (loop probability profiles) are also produced. The user can select among hairpin, bulge, interior, multi-branched, or exterior loops, for generation of *Hplot*, *Bplot*, *Iplot*, *Mplot* and *Extplot*, respectively. The probability that nucleotide i is in a loop of selected type is plotted at position i (here we are interested in individual bases with $W=1$, not fragments of several bases). For RNA sequences longer than 400 bases, the loop profile is displayed for any region of 400 bases, with the starting position specified by the user. At the present, we do not know whether a certain type of loop is more favorable than other types for binding by complementary nucleic acids.

siRNA internal stability profiling. The internal stability profile for every possible siRNA duplex is available through the interactive graphic window for siRNA ends and internal stability profiling. On the profile, the user can use the stabilities for positions 1 and 19 to make a visual comparison of the stabilities of siRNA duplex ends. The stability profiles are not included in the Zip or tar.gz file for **Sfold** output, because of large number of profiles for long mRNAs. However, from the interactive profiling window, the user can download the profiles for the selected target sites.

Text 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. **Sirna** 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 **Sirna** 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 ($k=2, \dots, 7$), 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 (kcal/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 *filtered.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) $30\% \leq \text{GC } \% \leq 60\%$;

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) $\text{GC } \% = \text{GC count in siRNA (excluding overhangs)} / 19 \times 100\%$;

- 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 *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 4 is used for making profile plot for individual bases for ribozyme application.

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 (for *Hplot*)

Column 4: the probability that this nucleotide is in a bulge loop (for *Bplot*)

Column 5: the probability that this nucleotide is in an interior loop (for *Iplot*)

Column 6: the probability that this nucleotide is in a multi-branched loop (for *Mplot*)

Column 7: the probability that this nucleotide is in the exterior loop (for *Extplot*)

Column 8: sum of columns 3 through 7 (this is the same as column 4 of file *sstrand.out*)

Other output

Links to output from other modules with default input settings are provided.

Output downloading

With the exception of siRNA internal stability profiles, all of the output and sampled structures in Zip or compressed tar (tar.gz) format are available for downloading. After the compressed file has been uncompressed, a directory with the job ID as the name is created under the current directory. Under the job ID directory, there are seven subdirectories and a file *readme.txt* to describe the files in the seven subdirectories.

Job and system information

Links to information on the job and system usage are provided.

Soligo

Graphical output

Same as the output for **Sirna**, except for internal stability profiling for siRNAs.

Text 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) $40\% \leq \text{GC} \% \leq 60\%$;

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.

Files *sstrand.out* and *loopr.out* are the same as described for **Sirna**.

Sribo

Graphical output

For every site of the selected cleavage triplet (e.g., GUC) on the target RNA, the probability profile for individual bases ($W=1$) is produced for the region that includes the triplet and the two flanking sequences of 15 bases each. Thus, column 4 (not column 5 or 6) of file *sstrand.out* is used as input data for profiling in ribozyme applications. **Sfold** does not address the issue of

optimal length for flanking sequences (hammerhead's binding arms); the length of 15 bases was selected for the purpose of covering the normal range of length with some cushion. We recommend selection of cleavage sites for which *both* flanking sequences are at least partially accessible, because antisense hybridization is believed to start with nucleation at a location of several unpaired bases, and then elongation occurs by "unzipping" the adjacent helix on the target (Milner *et al.* 1997). It is unclear if accessibility of the cleavage triplet is important for cleavage. We note that profiling of the target RNA only addresses the accessibility of the target. It is also important to assess the folding of a designed ribozyme whose binding arms are determined by the cleavage triplet and its flanking sequences. To address the issue of ribozyme folding, the user can run module *Srna* to fold the ribozyme. Output from this module is helpful for a confidence assessment about the degree of correct ribozyme folding.

Loop profiling is the same as described for **Sirna**.

Text output files

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

Srna

This module provides tools and statistics to statistically characterize the Boltzmann ensemble through the sampled structures.

Representation of the structure sample

A two-dimensional histogram (*2Dhist*) displays base pair probabilities computed from a statistical sample of structures. In the *2Dhist*, base pair probabilities are shown by solid squares in the upper left triangle, with the nucleotide positions on both axes. The areas of the solid squares are proportional to the frequencies of the base pairs in the sampled structures. The current sample size is 1,000 structures, which is sufficiently large for producing stable estimates for *2Dhist* and probability profiles (Ding & Lawrence 2001). Of course, the user can experiment with this, by folding the same sequence twice and comparing results to assess the power of statistical sampling.

2Dhist has an option for the display of base pair probabilities. When this option is selected, the probability and positions of the base pair for a solid square can be shown through mouse pointing. For long RNA sequences, the base pair probabilities take some time to load after the *2Dhist* is displayed, we thus have set "no base pair probabilities" as the default display. The lower right triangle could be used to plot the minimum free energy (MFE) structure or any single structure of particular interest. However, particularly for long sequences, a single structure is of little significance from the Boltzmann ensemble perspective. For example, for *E. coli* RNAase P, which has a moderate length of 377 nt, the MFE structure from *mfold* 3.1 has a free energy of -173.6 kcal/mol, and its Boltzmann probability is merely 0.000037. For *E. coli lacZ*, 3113 nt, the MFE is -1234 kcal/mol, and the MFE structure has a Boltzmann probability of 2.84E-49!

The cumulative free energy distribution of sampled structures is similar to the cumulative distribution function (CDF) in probability theory. The MFE in the sample (SMFE) is computed. We note that the SMFE is not necessarily the global MFE for long sequences, because of the MFE structure's small Boltzmann probability and the many competing structures with similar free energies. However, the chance of observing the MFE structure in the sample increases with the size of the sample. For an integer $0 \leq P \leq 100$, the free energy CDF plots the probability that a structure in the sample has a free energy within $P\%$ of the SMFE.

An *ad hoc* representation of the structure sample is given by a table. First, LFE, the largest free energy of structures in the sample is computed. The free energy range covering all structures in the sample, i.e., [SMFE, LFE], is then divided into ten equally spaced intervals. For each free energy interval, the structure with the lowest free energy is selected as the representative. For the representative, the table presents its associated free energy interval, the frequency with which structures in the sample fall into the energy interval (the frequencies for the ten intervals add up to 1.0), the free energy of the structure, and the secondary structural diagram. The structural diagram is currently available in PNG, PDF, and PostScript formats. The PNG format has the capabilities for enlargement and local display. The structural diagrams are produced with further modifications of the modified *naview* program (Zuker, Mathews & Turner 1999; Brucoleri & Heinrich 1988; <http://www.bioinfo.rpi.edu/~zukerm/rna/node3.html#SECTION00033>). The GCG connect file is also provided. We note that this is a rather crude representation of the structure sample and the Boltzmann ensemble, mainly because structures in a common free energy interval can have substantially different structural features. An appealing method for an efficient statistical presentation of the Boltzmann ensemble is to classify the structures in the sample (Ding & Lawrence 2003). A classification algorithm based on a suitable distance measure is under development.

Text output files

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.

ALGORITHMS

The software is based on a statistical sampling algorithm for prediction of RNA secondary structure. The details of the algorithm and its unique capabilities are presented in Ding & Lawrence (2003). The recent Turner free energies (Xia *et al.* 1998; Mathews *et al.* 1999) are used in the algorithm, with the exception of co-axial stacking. The extension of the algorithm to a probability profiling algorithm for prediction of target accessibility is reported in Ding & Lawrence (2001). A much simpler sampling algorithm for the stacking energy model is presented in Ding (2002). Bayesian inferences on energy parameters and loop numbers for the stacking energy model are described in Ding & Lawrence (1999).

In research publications, the users of Sfold are requested to cite the articles describing the algorithms and the web server (Ding & Lawrence 2003, 2001; Ding, Chan, Lawrence 2004), in addition to including the web server site <http://sfold.wadsworth.org>.

REFERENCES

RNA structure sampling algorithms and applications

Ding, Y., Chan, C.Y. and Lawrence, C.E. (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.* (Web Server Issue, PDF available from Sfold Web site).

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. (2002) Statistical algorithms for folding and target accessibility prediction and design of nucleic acids. *Pending U.S. Patent*, Wadsworth Center, New York State Department of Health; filed on January 28, 2003 by Frommer Lawrence & Haug LLP, 745 Fifth Avenue, New York, NY 10151, (212) 588-0800. (U.S. Provisional Patent Application 60/352,643, filed on January 29, 2002; law office case # 454311-2230.1 WO)

Ding, Y. (2002) Rational statistical design of antisense oligonucleotides for high throughput functional genomics and drug target validation. *Statistica Sinica* **12**, 273-296.

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.

Ding, Y., and Lawrence, C.E. (1999) A Bayesian statistical algorithm for RNA secondary structure prediction. *Computers and Chemistry* **23**, 387-400.

Target RNA secondary structure/accessibility and potency of siRNAs

Far, R.K. and Sczakiel, G. (2003). The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides. *Nucleic Acids Research*, **31**, 4417-4424.

Vickers, T.A., Koo, S., Bennett, C.F., Crooke, S.T., Dean, N.M., Baker, B. (2003) Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis. *J. Biol. Chem.* **278**, 7108-18.

Bohula, E.A., Salisbury, A.J., Sohail, M., Playford, M.P., Riedemann, J., Southern, E.M., Macaulay, V.M. (2003) The efficacy of small interfering RNAs targeted to the type 1 Insulin-like growth factor receptor (IGF1R) is influenced by secondary structure in the IGF1R transcript. *J. Biol. Chem.* **278**, 15991-7.

Lee, N.S., Dohjima, T., Bauer, G., Li, H., Li, M.J., Ehsani, A., Salvaterra, P., Rossi, J. (2002) Expression of small interfering RNAs targeted against HIV-1 rev transcripts in human cells. *Nat. Biotechnol.* **20**, 500-5.

siRNA duplex thermodynamics and features and rational design of siRNAs

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.

Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., Saigo, K. (2004) Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res.* **32**, 936-48.

Amarzguioui, M., Prydz, H. (2004) An algorithm for selection of functional siRNA sequences. *Biochem Biophys Res Commun.* **316**, 1050-8.

Ding, Y. and Lawrence, C.E. (2004) Rational design of siRNAs with the Sfold software. In *RNA Interference: from Basic Science to Drug Development*, ed. Krishnarao Appasani, Cambridge University Press (PDF available from Sfold Web site).

Turner thermodynamic parameters

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.

DNA/RNA stacking energy parameters

Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M., Sasaki, M. (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **34**, 11211-6.

Other references

Brucoleri RE; Heinrich G (1988). An improved algorithm for nucleic acid secondary structure display. *Comput. Appl. Biosci.* **4**, 167-73.

Christoffersen, R.E., McSwiggen, J.A. and Konings, D. (1994). Application of computational technologies to ribozyme biotechnology products. *J. Mol. Structure (Theochem)*, **311**, 273-284.

Milner, N., Mir, K.U. and Southern, E.M. (1997). Selecting effective antisense reagents on combinatorial oligonucleotide arrays. *Nat. Biotechnol.* **15**, 537-541.

Zuker, M., Mathews, D.H. and Turner, D.H. (1999). Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In, *RNA Biochemistry and Biotechnology*, 11-43, J. Barciszewski & B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers, Dordrecht, the Netherlands.

§§§ Updated by Ye Ding, April 15, 2004