*S*TarMir

Overview

STarMir is a tool for predicting potential binding sites for one or multiple microRNAs (miRNA) in a target RNA sequence (typically mRNA). It is based on the logistic prediction models developed with miRNA binding data from crosslinking immunoprecipitation (CLIP) studies. Each of the candidate sites is assigned a logistic probability as a measure of confidence in the predicted site (Liu et al. 2013, *Nucleic Acids Res* **41**, e138).

For a given pair of miRNA:target mRNA, STarMir first predicts target secondary structures (Ding and Lawrence 2003 *Nucleic Acids Res*, **31**, 7280-7301). Potential miRNA binding sites are then predicted by the RNAhybrid program (Rehmsmeier et al. 2004, *RNA*, **10**, 1507-1517) for either seed matches or seedless sites with a hybrid stability of −15 kcal/mol or lower. For each site, a comprehensive list of sequence and structure-based features are computed as described in Liu et al. (2013). These features are used by our logistic model with parameters specific for the site type (seed or seedless) and the target region (5' UTR, CDS or 3' UTR) to compute a logistic probability as a measure of confidence in the predicted site. In general, a probability of 0.5 indicates a fairly good chance of microRNA binding. A high likelihood of microRNA binding is predicted by a high probability, e.g., 0.75 or higher. In addition, for each site, STarMir also outputs all of the site features along with a diagram of miRNA:target hybrid conformation.

Submitting input

Model for prediction

This selects the model for making binding site predictions. Currently available models are: a model trained on human (*H. sapiens*) V-CLIP data (Kishore et al. 2011, *Nat. Meth.* **8**, 559- 564), a model trained on mouse (*M. musculus*) HITS-CLIP data (Chi et al. 2009, *Nature* **460**, 479-486) and a model for *C. elegans* based on analysis and modeling of worm ALG-1 CLIP data (Zisoulis et al, *Nat Struct Mol Biol*, **17**, 173-179). The good performance of the mammalian models in cross-species validation suggests that they can also be applicable to other mammalian species (Liu et al. 2013).

Species for prediction

This selects the species of the target sequence. This information is only utilized if the user enters a RefSeq ID for the target mRNA instead of manually entering the sequence. The selected species will be used by the server to employ pre-stored evolutionary conservation information in modeling computation, if conservation data is available in our database. When the selected species is “other”, it is not possible for the model to incorporate conservation information in predictions. The choice of species has no effect if the mRNA sequence information is entered manually. Furthermore, the good performance of the models in cross-species validation suggests that the STarMir applications are not limited to the species with available CLIP data and prediction models but can also be extended to other species.

microRNA sequence(s)

In the default option, one or more miRNA IDs can be entered, e.g., hsa-let-7a-3p, mmu-mir-128-1, cel-mir-90. For this option, the sequences are retrieved from an internal database built from miRBase release 20. Alternatively, one or more miRNA sequences can be pasted into the input box in FASTA format, or uploaded from a FASTA file (Fig.1). To load from a file, clicking the **Upload FASTA file** checkbox, will make a **Browse** button visible. Clicking it would permit selection of a file for uploading. There is no limit on the number of miRNA sequences that can be entered. However, each miRNA sequence must be less than 55 nts in length. Any character in the miRNA other than A, T, C, G and U will be removed.

Single target sequence

There are three methods for entering the target sequence as described below:

- The default method for selecting a target sequence is to enter the RefSeq ID in the provided input box. The sequence would then be retrieved from our database of mRNA sequences, which currently contains ~19,000 sequences from NCBI RefSeq Build 36.3 for human and ~12,000 sequences from Build 37.2 for mouse. If the sequence is specified using the RefSeq ID and is present in our mRNA database, the models will utilize evolutionary conservation information for improved predictions. If this option is used, the information about the CDS start and end positions, will be retrieved from our database, and binding sites will be reported for all three regions
- Sequence information may be entered manually by selecting the "Manual sequence entry" option and entering the sequence in raw or FASTA format in the text area provided. If the sequence is entered manually, mRNA region information needs to be provided to the server through the region dropdown box directly above the sequence input box. The user needs to indicate whether the sequence entered represents an entire mRNA or a single region (3′ UTR, CDS or 5′ UTR). If the sequence represents the entire mRNA, the nucleotide positions for the start and end of the coding region must be specified in the boxes provided below the input window.
- In order to enter sequence from a file in FASTA format, the user can begin by selecting the **Manual sequence entry** option. Then, clicking the **Upload FASTA file** checkbox will make the **Browse** button visible. This can be used to select a file for uploading. If the file sequence represents the entire mRNA, the nucleotide positions for the start and end of the coding region must be specified.

All characters other than **A**, **C**, **G**, **T** and **U** will be deleted from the input sequence. The current web server limit is 5,000 nts. For longer sequences, only the 5,000 nts starting from the 5' end will be used for carrying out the analysis.

.

Email address (optional)

A link to the job retrieval URL with status update information is provided during job submission.

Although an email address is optional, if it is provided, it will be used to email a notice to the user when the job is completed.

Output

Overview

Demo output of sample sequence data can be accessed by clicking the “DEMO OUTPUT” button in the STarMir menu line or directly at http://sfold.wadsworth.org/starmirdemo/starmir.html. The output results are presented to the user through both an interactive viewer and downloadable files.

Interactive Viewer

The interactive viewer permits the user to examine the binding site predictions in detail. The sites are divided into six groups according to mRNA regions (5′ UTR, CDS and 3′ UTR) and site types (seed or seedless) and are presented in a six-tabbed pane. Tabs are provided at the top of the display area to select one of the six groups. The output is provided in a tabular format with one line per site. Each line displays the details of the following features for the binding site: the position of the site on the target, the logistic probability of the site and various thermodynamic, sequence and structural features used by the logistic model. Features used for the calculation of the logistic probabilities by the model are marked with "*". A file for definitions of the features with references is available by clicking the link for “Feature definitions” under the result table. Some features may be excluded for a specific model due to a lack of feature enrichment (see Liu et al. (2013), Supplementary Tables 1-5). In the result table, the binding sites are presented in descending order of their logistic probabilities.

Additionally, a link to a graphic representation of the hybrid conformation (third field in each row) and a PDF of the diagram is available for visualization and download. Clicking the diagram links to a downloadable high-resolution image of the hybridization diagram in a PDF format.

Output for download

This section provides links to the downloadable files containing the data displayed in the interactive viewer. The text files include all site features calculated by STarMir with the features used in the prediction model marked with an asterisk (*). The prediction models exclude features that were not enriched in CLIP data analysis. A text file is available for each of the six categories represented by the tabs

Additionally, there is a file representing a simplified text version of the hybrid conformations for each site, and a file presenting the probability that each nucleotide in the site is unpaired (i.e., single-stranded).

Download all output files

This section allows all the files to be downloaded as either a zip or tar/gzip archive.

Output from other application modules

The dropdown menu within this section allows access to predictions generated by Srna as well as other modules of the Sfold.

Description of all site features for website viewer and download files

Site ID	Predicted sites are sequentially numbered along the target sequence
Target	Accession number of the target mRNA
miRNA	Name of the microRNA (miRNA)
Target_Len	Length of the target
Site_Position	Start and end position of the target region (site) predicted to be bound by miRNA
Seed_Position	Start and end position of the target sub-region complementary to the miRNA seed (i.e. positions 2-7/8 of the miRNA)
Seed_Type	6mer, offset 6mer, 7mer-A1, 7mer-m8, and 8mer seed sites (Bartel 2009, *Cell*, **136**, 215-33)
Site_Access	A measure of structural accessibility as computed by the average probability of a nucleotide being single-stranded (i.e., unpaired) for the nucleotides in the predicted binding site
Seed_Access	A measure of structural accessibility as computed by the average of single-stranded probabilities of the nucleotides in the target sub-region complementary to the miRNA seed
Upstream_Access (# nt)	A measure of structural accessibility as computed by the average of single-stranded probabilities for the block of nucleotides upstream of the predicted binding site (# is the block size)
Dwstream_Access (# nt)	A measure of structural accessibility as computed by the average of single-stranded probabilities for the block of nucleotides downstream of the predicted binding site (# is the block size)
Upstream_AU (# nt)	Percentage of AU for the block of nucleotides upstream of the binding site (# is the block size)
Dwstream_AU (# nt)	Percentage of AU for the block of nucleotides downstream of the binding site (# is the block size)
Site_Location	Relative starting location of the predicted binding site along the length of the sequence (e.g., for 3′ UTR, 0 indicates the 5′ end of the UTR, and 1 corresponds to the 3′ end)
3′_bp	Presence of contiguous Watson Crick base pairing for miRNA nucleotide positions 12-17 (sites with 3′_bp are also called 3′ compensatory/supplementary sites)
Site_Consv	Conservation score by the PhastCons program for the binding site
Seed_Consv	Conservation score by the PhastCons program for the target sub-region complementary to the miRNA seed
Offseed_Consv	Conservation score by the PhastCons program for nucleotides within the target site, but outside the seed complementary region
dG_hybrid = ΔG_{hybrid}	A measure of stability for miRNA:target hybrid as computed by RNAhybrid (Rehmsmeier et al. 2004, *RNA* **10**, 1507-1517)
dG_nucl = ΔG_{nucl}	A measure of the potential of nucleation for miRNA:target hybridization (Long et al 2007, *Nat. Struct. Mol. Biol.* **14**, 287-294)
dG_total = ΔG_{total}	A measure of the total energy change of the hybridization (Long et al 2007, *Nat. Struct. Mol. Biol.* **14**, 287-294)
LogitProb	Probability of the site being an miRNA binding site as predicted by our nonlinear logistic model
Target_Mismatch	Nucleotides in the target binding site that are not base paired with the miRNA
Target_Match	Nucleotides in the target binding site that are base paired with the miRNA

Mir_Match	Nucleotides in the miRNA that are base paired with the target mRNA
Mir_Mismatch	Nucleotides in the miRNA that are not base paired with the target mRNA
Hybrid Conformation	The last four fields above present information for the miRNA:target hybrid conformation predicted by RNAhybrid. In each of the fields, spaces are included so the fields can be easily aligned to produce a simple diagram of the hybrid conformation as illustrated below:

```
Target_Mismatch: U    UUUCC     U          A
Target_Match:     GACU     AUGUA     CUACCUC
Mir_Match:        UUGA     UACGU     GAUGGAG
Mir_Mismatch:                  UGGAU       A
```