HHS Public Access

Author manuscript

Methods Mol Biol. Author manuscript; available in PMC 2017 March 16.

Published in final edited form as:

Methods Mol Biol. 2016 ; 1490: 73–82. doi:10.1007/978-1-4939-6433-8_6.

STarMir Tools for Prediction of microRNA binding sites

Shaveta Kanoria[1], **William Rennie**[1], **Chaochun Liu**[1], **C. Steven Carmack**[1], **Jun Lu**[2], and **Ye Ding**[1,*]

[1]Wadsworth Center, New York State Department of Health, Center for Medical Science, 150 New Scotland Avenue, Albany, NY 12208, USA

[2]Department of Genetics and Yale Stem Cell Center, Yale University, New Haven, CT 06520, USA

Abstract

MicroRNAs (miRNAs) are a class of endogenous short non-coding RNAs that regulate gene expression by targeting messenger RNAs (mRNAs), which results in translational repression and/or mRNA degradation. As regulatory molecules, miRNAs are involved in many mammalian biological processes and also in the manifestation of certain human diseases. As miRNAs play central role in the regulation of gene expression, understanding miRNA-binding patterns is essential to gain an insight of miRNA mediated gene regulation and also holds promise for therapeutic applications. Computational prediction of miRNA binding sites on target mRNAs facilitates experimental investigation of miRNA functions. This chapter provides protocols for using the STarMir web server for improved predictions of miRNA binding sites on a target mRNA. As an application module of the Sfold RNA package, the current version of STarMir is an implementation of logistic prediction models developed with high throughput miRNA binding data from crosslinking immuno-precipitation (CLIP) studies. The models incorporated comprehensive thermodynamic, structural and sequence features, and were found to make improved predictions of both seed and seedless sites, in comparison to the established algorithms

*Correspondence: ye.ding@health.ny.gov; Phone: (518) 486 1719; Fax: (518) 474 3183.

[1]To further assist the users with STarMir input and output features, an online ‘MANUAL’ as well as ‘DEMO OUTPUT’ are provided at the STarMir front page.

[2]As described above, the STarMir computes a logistic probability score from a selection of thermodynamic and structure based features. The logistic probability provides the measure of confidence in the predicted miRNA binding site. A site with a probability of 0.5 indicates a good chance of being a true miRNA binding site. Further, higher probability scores e.g. 0.75 and above, suggests greater likelihood of miRNA binding in vivo.

[3]In general, STarMir based predictions are time consuming due to RNA folding. For current Sfold web server, typical processing time is three minutes for 500 nt, five minutes for 1000 nt, 30 minutes for 2000 nt, two hours for 3000 nt, five hours for 4000 nt and nine hours for 5000 nt. Before a job submission, the users should first check STarMirDB, a database of pre-computed transcriptome-scale prediction results currently available for human, mouse and worm. STarMirDB complements STarMir, and is available at: http://sfold.wadsworth.org/starmirDB.php.

[4]As the underlying models of the STarMir performed very well in cross-species validations, the applications of STarMir are not limited to the species with available CLIP data, but rather can be extended to other species as well.

[5]As the CLIP methodology provides information of miRNA binding, the models developed from the CLIP data are efficient for prediction of miRNA binding sites and may not always be extendable to miRNA functions. In other words, these models do not make predictions for the functional outcome of miRNA binding (i.e. target degradation or translational repression) and the extent of regulation on either the mRNA or the protein.

[6]A specific database is under development, based on a recent study on genetic variants within and near miRNA binding sites [24]. The database will allow users to search for polymorphisms in the context of miRNA binding sites.

[1]. Their broad applicability was indicated by their good performance in cross-species validation. STarMir is freely available at http://sfold.wadsworth.org/starmir.html

Keywords

miRNA; CLIP; target mRNA; RNA secondary structure; miRNA binding site

1. Introduction

MicroRNAs (miRNAs) are a class of naturally occurring, small non-coding RNAs (ncRNAs) of ~21–25 nucleotide (nt) in length. miRNAss have been found in plants, animals and some viruses. A mature miRNA guides RNA-induced silencing complex (RISC) for target recognition by hybridizing to partially complementary sequences typically in the 3' untranslated regions (3' UTRs) of the target mRNAs, leading to translational repression and/or mRNA degradation of the target mRNA [2,3]. miRNA mediated gene regulation is rather extensive, as one miRNA may regulate hundreds of targets, whereas an individual mRNA can be targeted by multiple miRNAs [4]. miRNAs play important roles in numerous biological processes including development, differentiation, apoptosis and proliferation [3,5]. Additionally, mis-regulation in miRNA activity has been found to be associated with human diseases [6,7]. However, our current understanding of miRNA functions in physiological processes and diseases is rather limited. Identification of miRNA targets is essential for a full characterization of miRNA functions. For plants, identification of miRNA targets is straightforward, as most miRNAs are perfectly complementary to their target sequences [8]. However, in animals, the complementarity between miRNA and target mRNA is imperfect [9], presenting a challenge for binding site identification. Most of the algorithms for miRNA binding site prediction are based on the seed rule, i.e., the nucleotides of the target site forms Watson-Crick (WC) base pairs with nucleotide 2 to 7 or 8 of the 5' end of the miRNA [10]. However, an increasing number of studies show that some miRNA binding sites do not follow the seed rule [11–15]. In addition to seed, several sequence features have been proposed to be important for miRNA target binding. These include sequence conservation, strong base-pairing to the 3' end of the miRNA, local AU content and location of miRNA binding sites [16]. Based on a two-step model (Figure 1) for the hybridization between a miRNA and an mRNA with target secondary structure predicted by Sfold [17,18], the importance of target structure for miRNA target recognition was convincingly demonstrated [19–22]. Another independent mammalian study, established that structure based predictions could be more efficient than seed based predictions [23]. A recent study revealed that genetic variations can influence miRNA:target interactions and alter the structural accessibility of the binding sites as well as the flanking regions [24].

1.1 Identification of miRNA binding sites

Most existing algorithms for miRNA target prediction are primarily based on the seed rule. With the development of the CLIP technique [25], it has become possible to identify short AGO crosslinked sequences that contain miRNA binding sites. CLIP involves UV irradiation of tissues, organisms or cells, to covalently crosslink miRNA targets to the Argonaute (AGO) proteins (the catalytic component of the RISC complex. The crosslinked

RNAs are shortened by partial RNAse digestion to ~50 nt and further amplified by RT-PCR. The shortened RNA fragments are then sequenced for identification of AGO tags containing miRNA binding sites on the target mRNAs. Numerous CLIP studies have been published in the recent years, including HITS-CLIP for mouse brain [25], PAR-CLIP in human cell lines [26], variants of PAR-CLIP [27], and a study in worm [28]. These CLIP studies generate short target fragments containing miRNA binding sites, thereby providing a genome wide map of miRNA target interactions. The high throughput data from the CLIP studies have been successfully utilized in the development of logistic models for making improved miRNA binding site predictions [1]. These models are based on a comprehensive list of sequence, thermodynamic and target structure features that were found to be enriched for miRNA binding sites identified by CLIP, and were validated by intra-dataset, inter-dataset as well as cross-species validations [1]. The models have been implemented into the STarMir application module of the Sfold RNA package, which predicts miRNA binding sites on a target mRNA [29]. This chapter describes a detailed protocol for using STarMir web server. STarMir is a free web service available to all without any registration or email requirement.

Materials

As an application module of the Sfold RNA package (http://sfold.wadsworth.org), STarMir can be freely accessed at http://sfold.wadsworth.org/starmir.html. Through a web browser such as Safari, Internet Explorer or Firefox, the user can use the web service by providing either the miRNA ID and RefSeq ID for the mRNA or by submitting their miRNA and the target sequences.

Methods

3.1 Web protocol for using STarMir

This protocol outlines input and output of STarMir web service, provided through Sfold web server (http://sfold.wadsworth.org). The user can start by pointing a web browser to http://sfold.wadsworth.org/starmir.html.

3.2 STarMir Input Page

Figure 2 illustrates the main page with manual sequence entry option selected for both miRNA and target sequences. The user can input the sequence information for a single or multiple miRNAs and a single target mRNA for predicting miRNA binding sites. Upon job submission, a link is provided to the user for tracking the progress of the job and to access the prediction results. A detailed description of input is given below.

3.2.1. Model—STarMir predicts miRNA binding sites based on three models for human, mouse and worm. These models have been trained on V-CLIP data for human (*Homo sapiens*) [26], HITS-CLIP data for mouse (*Musmusculus*) [25] and ALG-1 CLIP data for worm (*Caenorhabditis elegans*), respectively [28]. The two mammalian models were cross-validated and can be broadly used for other species [1].

3.2.2. Species—The user needs to select a species for prediction. If user enters RefSeq ID of the target mRNA and selects one of the three modeling species, the species information

will be used for retrieving pre-stored evolutionary conservation information in predictions. If the mRNA sequence information were entered manually, the selection of species would not have any effect on predictions. Furthermore, if 'Other' is selected, conservation information cannot be used in predictions by our models.

3.2.3. miRNA—miRNA information can be provided in two ways. For the default option, one or more miRNA IDs can be entered (an example of miRNA ID is shown in Figure 2), for which the sequences are retrieved from an internal database developed using release 20 of the miRBase [30]. An alternative is to enter one or more miRNA sequences into the input box in FASTA format, or upload a FASTA file (Figure 2). Although there is no limit on the number of miRNA sequences that can be entered, each sequence must not be longer than 55 nt in length. Any characters other than A, T, G, C and U in the entered miRNA sequence are removed.

3.2.4. mRNA—The target mRNA information can be entered in three different ways. The default method is to enter the RefSeq ID in the provided input box (Figure 2), for which the sequence will be retrieved from our internal database of mRNA sequences for human and mouse. If the RefSeq ID of the mRNA is provided, evolutionary conservation information [31] will be used to make more accurate miRNA binding site predictions [1]. Alternatively, by selecting the 'Manual sequence entry' option, one can enter the sequence information in raw or FASTA format or upload a FASTA file. If the sequence is uploaded using a FASTA file, the file must not contain more than one sequence. As in the case of miRNA, any character in the mRNA sequences other than A, T, G, C and U is removed. As the current limit of the web server on the length of the mRNA sequence is 5000 nt, longer sequences will be truncated to 5000 nt starting from the 5' end.

3.2.5. mRNA region—For manual sequence entry, the user needs to inform the server if the entered sequence represents an entire mRNA or a single region, i.e. 3' UTR, CDS or 5' UTR, through a region dropdown box directly above the sequence input box. If the case of an entire mRNA, the nucleotide positions for the start and end of the coding region must be specified in the boxes shown below the input window (the first nucleotide of an entered sequence is counted as 1).

3.2.6. mRNA name—If the user provides a RefSeq ID for the mRNA, the CDS start and end would be retrieved from our internal mRNA database and the binding sites will be predicted for all three mRNA regions. RefSeq ID would be considered as the name of the sequence for output. However, for a manually entered sequence, the user can enter the name of the sequence.

3.2.7. Email address—Provision of an email address is optional. If an email address is provided, the user receives a notification once the job is completed. Alternatively, the user can check for job status using the link provided after job submission.

3.3 STarMir Output

Upon job completion, the results are presented through both an interactive viewer and downloadable files. An illustration of a typical output as an interactive viewer is shown in Figure 3, with ‘CDS-seedless’ tab and ‘hsa-let-7a-3p’ selected for display. The results are categorized as seed and seedless sites for each of the three target mRNA regions, i.e. 3' UTR, CDS and 5' UTR. Each tab represents prediction results for one or all miRNAs, which can be selected from the dropdown menu. The sites are presented in the descending order of their logistic probability scores. The output presents comprehensive sequence, thermodynamic and target structure features including the logistic probability score as the measure of confidence for a predicted site. Additionally, a link is provided to the graphic representation of the hybrid conformation. Hybrid diagrams for a 7mer-m8 seed site and a seedless site are shown in Figure 4. The PDF of the hybrid diagram is also available for visualization and download. Further, a file providing definitions of the features is available via the link for ‘Feature definitions’ below the result table. The results are also provided as downloadable tab-delimited text files, which present all site features calculated by STarMir. The features marked with an asterisk (*) are the ones that are used in the prediction model. The prediction models are based on the features that were enriched in the CLIP experiments [1]. A text file is provided for each of the six categories, as shown in different tabs. Alternatively, all the results can also be downloaded as a compressed archive, including a text version of the hybrid conformation diagrams for each site and a file showing the probability of each nucleotide in the site to be unpaired or single-stranded.

Acknowledgments

The Computational Molecular Biology and Statistics Core at the Wadsworth Center is acknowledged for supporting computing resources for this work. This work is supported in part by the National Science Foundation (DBI-0650991 to Y.D.), National Institutes of Health (GM099811 to Y.D. and J. L.).

References

1. Liu C, Mallick B, Long D, Rennie WA, Wolenc A, Carmack CS, Ding Y. CLIP-based prediction of mammalian microRNA binding sites. Nucleic Acids Res. 2013; 41(14):e138. [PubMed: 23703212]
2. Ambros V. The functions of animal microRNAs. Nature. 2004; 431(7006):350–355. [PubMed: 15372042]
3. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004; 116(2):281–297. [PubMed: 14744438]
4. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009; 19(1):92–105. [PubMed: 18955434]
5. Harfe BD. MicroRNAs in vertebrate development. Curr Opin Genet Dev. 2005; 15(4):410–415. [PubMed: 15979303]
6. Esau CC, Monia BP. Therapeutic potential for microRNAs. Adv Drug Deliv Rev. 2007; 59(2–3): 101–114. [PubMed: 17462786]
7. Erson AE, Petty EM. MicroRNAs in development and disease. Clin Genet. 2008; 74(4):296–306. [PubMed: 18713256]
8. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP. Prediction of plant microRNA targets. Cell. 2002; 110(4):513–520. [PubMed: 12202040]
9. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003; 115(7):787–798. [PubMed: 14697198]

10. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005; 120(1):15–20. [PubMed: 15652477]
11. Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature. 2008; 455(7216):1124–1128. [PubMed: 18806776]
12. Vella MC, Choi EY, Lin SY, Reinert K, Slack FJ. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. Genes Dev. 2004; 18(2):132–137. [PubMed: 14729570]
13. Didiano D, Hobert O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol. 2006; 13(9):849–851. [PubMed: 16921378]
14. Loeb GB, Khan AA, Canner D, Hiatt JB, Shendure J, Darnell RB, Leslie CS, Rudensky AY. Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. Mol Cell. 2012; 48(5):760–770. [PubMed: 23142080]
15. Lal A, Navarro F, Maher CA, Maliszewski LE, Yan N, O'Day E, Chowdhury D, Dykxhoorn DM, Tsai P, Hofmann O, Becker KG, Gorospe M, Hide W, Lieberman J. miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. Mol Cell. 2009; 35(5):610–625. [PubMed: 19748357]
16. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136(2):215–233. [PubMed: 19167326]
17. Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. 2003; 31(24):7280–7301. [PubMed: 14654704]
18. Ding Y, Chan CY, Lawrence CE. Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res. 2004; 32(Web Server issue):W135–W141. [PubMed: 15215366]
19. Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y. Potent effect of target structure on microRNA function. Nat Struct Mol Biol. 2007; 14(4):287–294. [PubMed: 17401373]
20. Long D, Chan CY, Ding Y. Analysis of microRNA-target interactions by a target structure based hybridization model. Pac Symp Biocomput. 2008:64–74. [PubMed: 18232104]
21. Hammell M, Long D, Zhang L, Lee A, Carmack CS, Han M, Ding Y, Ambros V. mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. Nat Methods. 2008; 5(9):813–819. [PubMed: 19160516]
22. Liu C, Rennie WA, Mallick B, Kanoria S, Long D, Wolenc A, Carmack CS, Ding Y. MicroRNA binding sites in C. elegans 3' UTRs. RNA biology. 2014; 11(6):693–701. [PubMed: 24827614]
23. Malhas A, Saunders NJ, Vaux DJ. The nuclear envelope can control gene expression and cell cycle progression via miRNA regulation. Cell Cycle. 2010; 9(3):531–539. [PubMed: 20081371]
24. Liu C, Rennie WA, Carmack CS, Kanoria S, Cheng J, Lu J, Ding Y. Effects of genetic variations on microRNA: target interactions. Nucleic Acids Res. 2014; 42(15):9543–9552. [PubMed: 25081214]
25. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature. 2009; 460(7254):479–486. [PubMed: 19536157]
26. 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. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010; 141(1):129–141. [PubMed: 20371350]
27. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat Methods. 2011; 8(7): 559–564. [PubMed: 21572407]
28. Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, Yeo GW. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. Nat Struct Mol Biol. 2010; 17(2):173–179. [PubMed: 20062054]
29. Rennie W, Liu C, Carmack CS, Wolenc A, Kanoria S, Lu J, Long D, Ding Y. STarMir: a web server for prediction of microRNA binding sites. Nucleic Acids Res. 2014; 42(Web Server issue):W114–W118. [PubMed: 24803672]

30. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014; 42(vol Database issue):D68–D73. [PubMed: 24275495]

31. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15(8):1034–1050. [PubMed: 16024819]

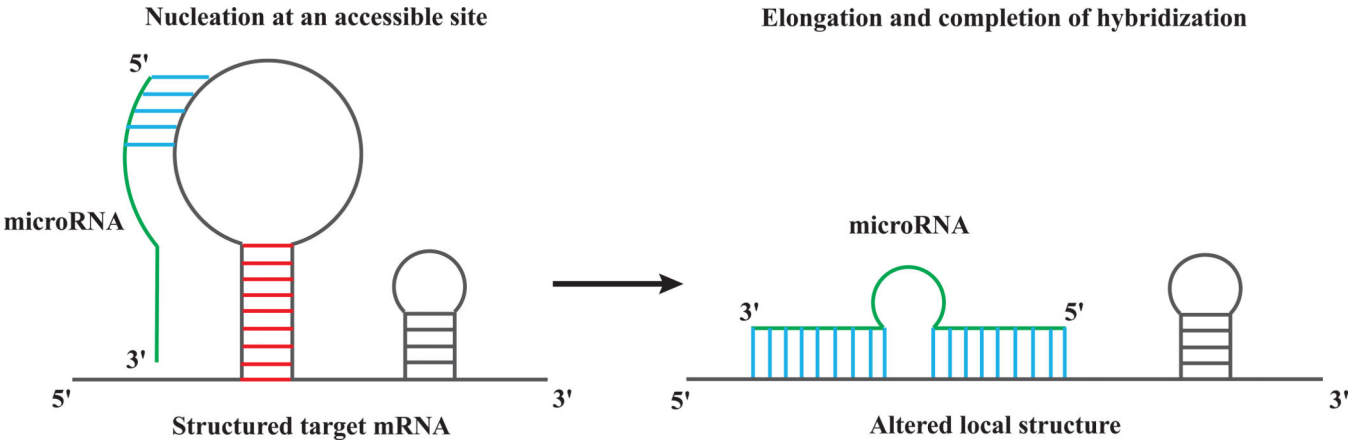

Figure 1.

A two-step hybridization model: nucleation at an accessible target site, followed by hybrid elongationto disrupt local target secondary structure and form the complete microRNA-target duplex [19].

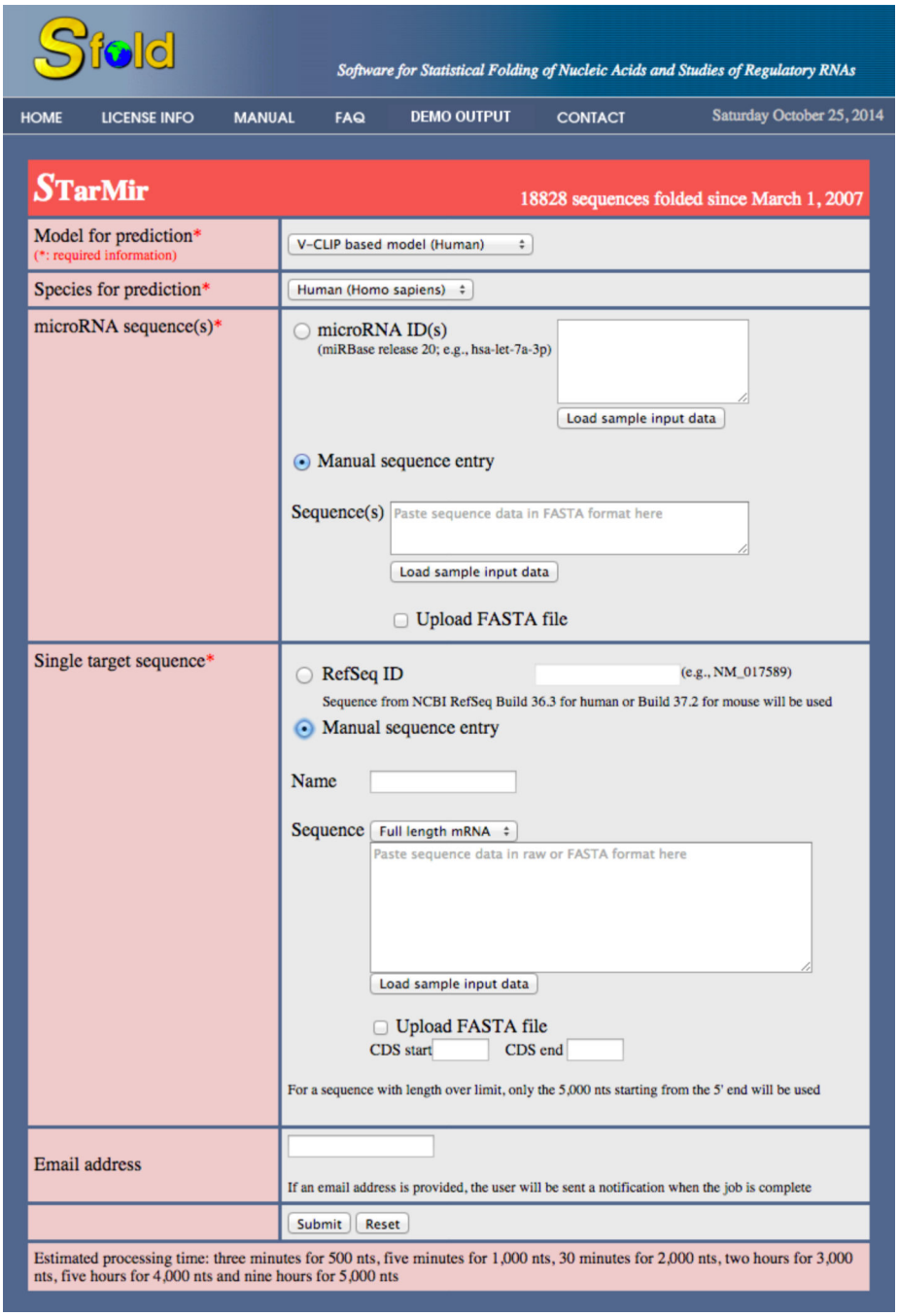

Figure 2.

STarMir input page displaying the input requirements for submitting miRNA and target mRNA sequences.

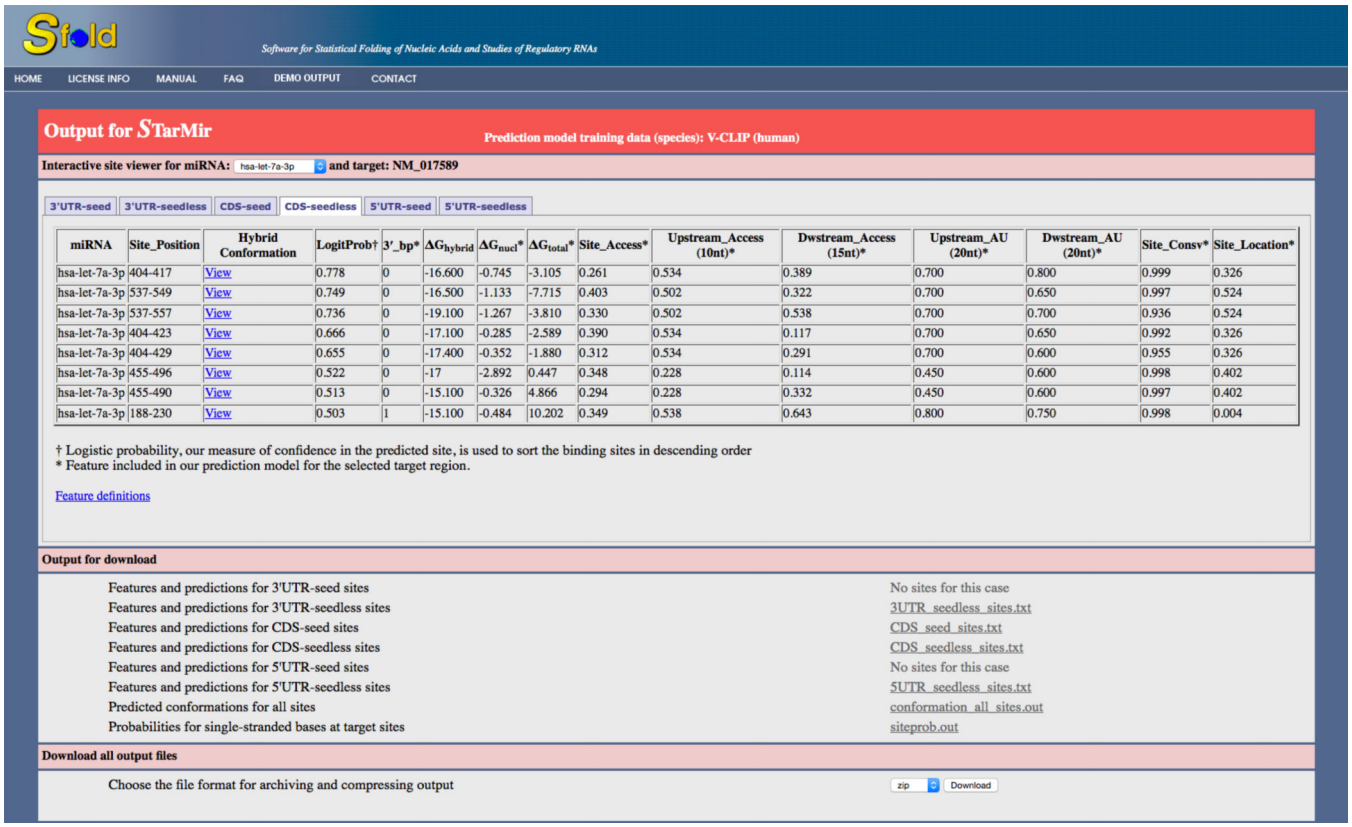

Figure 3.

STarMir output page showing the interactive site viewer with ‘3' UTR-seedless’ tab and ‘has-miR-501-3p’ selected for display. The download links for the text files are also shown.

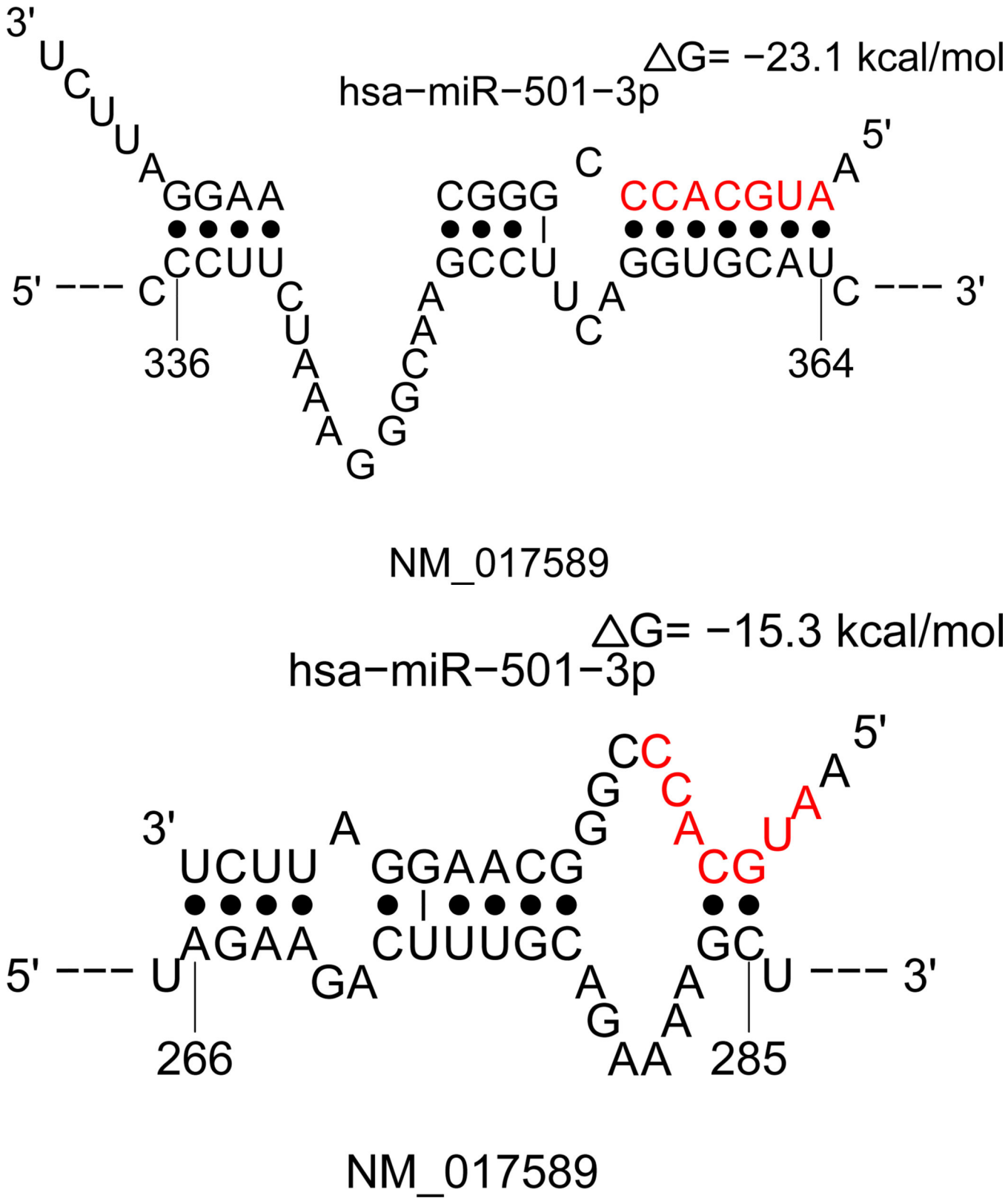

Figure 4.

Schematic representation of the hybrid diagrams for (**a**) a seed site (miRNA seed region (nt 2–7)) and (**b**) a seedless (non-canonical) site. The seed regions are shown in red.