STarMirDB: a database of microRNA binding sites

William Rennie¹, Shaveta Kanoria¹, Chaochun Liu¹, Bibekanand Mallick¹,², Dang Long¹,³, Adam Wolenc¹, C. Steven Carmack¹, Jun Lu⁴ and Ye Ding¹,*

*Correspondence:
Email: ye.ding@health.ny.gov; Phone: (518) 486 1719; Fax: (518) 474 3181

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

² Present address: RNA Biology and Functional Genomics Laboratory, Department of Life Science, National Institute of Technology, Rourkela-769008, Odisha, India

³ Present address: Biotechnology Department, Faculty of Chemistry, Danang University of Science and Technology, 54 Nguyen Luong Bang St., Danang, Vietnam

⁴ Department of Genetics and Yale Stem Cell Center, Yale University, New Haven, CT 06520, USA
Abstract

microRNAs (miRNAs) are an abundant class of small endogenous non-coding RNAs (ncRNAs) of ~22 nucleotides (nts) in length. These small regulatory molecules are involved in diverse developmental, physiological and pathological processes. miRNAs target messenger RNAs (mRNAs) for translational repression and/or mRNA degradation. Predictions of miRNA binding sites facilitate experimental validation of miRNA targets. Models developed with data from CLIP studies have been used for predictions of miRNA binding sites in the whole transcriptomes of human, mouse and worm. The prediction results have been assembled into STarMirDB, a new database of miRNA binding sites available at http://sfold.wadsworth.org/starmirDB.php. STarMirDB can be searched by miRNAs or mRNAs separately or in combination. The search results are categorized into seed and seedless sites in 3′ UTR, CDS and 5′ UTR. For each predicted site, STarMirDB provides a comprehensive list of sequence, thermodynamic and target structural features that are known to influence miRNA: target interaction. A high resolution PDF diagram of the conformation of the miRNA:target hybrid is also available for visualization and publication. The results of a database search are available through both an interactive viewer and downloadable text files.

Keywords

microRNA, CLIP, binding site prediction, seed site, seedless site
Introduction

miRNAs are a class of single-stranded, non-coding RNAs of ~22 nucleotides in length. They have been discovered in plants, animals as well as in some viruses \(^1\)\(^-\)\(^3\). miRNAs play essential roles in cell proliferation, differentiation, development, and are associated with human diseases \(^2\), \(^4\). A mature miRNA can guide miRNA-induced silencing complex (miRISC) for target recognition by sequence complementarity between the miRNA and sequences typically in the 3’ untranslated regions (3’ UTRs) of the cognitive messenger RNAs (mRNAs). Successful target binding usually results in translational repression and/or mRNA degradation \(^5\). Each human miRNA is predicted to be able to regulate several hundred different mRNAs \(^6\).

Computational prediction algorithms have proven to be valuable in the discovery of new miRNA targets. Most of the existing algorithms are based on the seed rule, i.e., the target site within 3’ UTR forms Watson-Crick (WC) pairs with bases at positions 2 through 7/8 of the 5’ end of the miRNA \(^7\). However, numerous exceptions to the seed rule have been well-documented \(^8\)-\(^13\). Other sequence features have been proposed based on their enhancement of targeting specificity. These include sequence conservation, strong base-pairing to the 3’ end of the miRNA, local AU content and location of miRNA binding sites (near either end of the 3’ UTR is favorable) \(^14\). The importance of target structural accessibility for miRNA target recognition has been supported by numerous studies \(^15\)-\(^21\).

In recent years, experimental methods based on cross-linking immunoprecipitation (CLIP) have been developed. For human and mouse studies, these include HITS-CLIP \(^22\), PAR-CLIP \(^23\) and variations of such techniques \(^24\). The CLIP approach has also been successful in worm \(^25\). The CLIP studies have provided high throughput quality datasets for regions of mRNAs containing miRNA binding sites. These data allowed us to develop models for improved predictions of miRNA binding sites \(^18\), \(^26\). The models are based on a comprehensive list of sequence, thermodynamic and target structure features that were enriched for miRNA binding sites identified from CLIP data, and were validated by intra-dataset, inter-dataset as well as cross-species validations. For human, mouse and worm, we have used these models to carry out transcriptome-scale predictions of both seed and seedless sites in the 3’ untranslated region (3’ UTR), coding sequence (CDS) region, and 5’ untranslated region (5’ UTR) of mRNAs. The results have been assembled into STarMirDB, a new database application module of the Sfold
RNA package^{27,28}. In this article, we describe this new resource. The unique tools of STarMirDB shall complement the existing miRNA target resources for computational predictions and experimental target data. Examples of these include, but are not limited to, TargetScan^{29}, Diana-microT^{30}, TarBase^{31}, StarBase^{32}, miRecords^{33} and miRTarBase^{34}.

**Generation of transcriptome-scale data for STarMirDB**

The database currently contains records for three species, *H. sapiens* (human), *M. musculus* (mouse) and *C. elegans* (worm). For human and mouse, we used complete mRNA sequences from NCBI RefSeq build 36.3 and 37.2, respectively. For worm, 3’ UTR sequences were obtained from the Wormbase version WS-190. The current release of STarMirDB includes 38,745 transcripts for human, 34,631 for mouse and 22,926 3’ UTRs for worm. miRNA sequences were obtained from miRBase release 18^{35}. We collected 1,921 miRNA sequences for human, 1,157 for mouse and 368 for worm.

Our CLIP based models were used to make transcriptome scale predictions of both seed and seedless binding sites^{18}. For each site, a comprehensive list of sequence, thermodynamic and target structure features are computed (Table 1). A logistic probability is provided as a measure of confidence in the predicted site. The number of binding sites is astronomical, so that in the database we only included those with a probability of 0.5 or higher. This filter also helps assure a reasonable response time for database search queries. In the case of interest in those low confidence sites with probabilities under 0.5, the user can use the STarMir web server that presents all predicted sites for single or multiple miRNAs and a target mRNA^{36}. The database can be searched by one or more miRNAs or targets, separately or in a combination. For worm, we provide a user interface that allows developmental stage specific search of miRNA binding sites within the 3’ UTR of transcripts. This interface is activated when *C. elegans* is selected as the species for database search. Additionally, for *C. elegans*, all the prediction data for miRNA binding sites within the 3’ UTR of transcripts are also provided as downloadable files.

**Input of database query**

STarMirDB presents a collection of predicted miRNA binding sites on mRNAs through a web interface that enables both search and retrieval of data and visualization of the conformation of predicted miRNA:target hybrid for each predicted site. The web interface has four input fields:
species, miRNAs, mRNAs, and logistic probability threshold. The requirements for each input field are described below in detail.

To start the database search, the user should first select the species from a dropdown menu. Currently three species are included in the database: human (Homo sapiens), mouse (Mus musculus) and worm (C. elegans). Next, one or a set of miRNAs can be selected from the miRNA scroll down list, which displays all available miRNAs assembled for the selected species. Additionally, one or more miRNA names can be entered in the text box. The database follows the naming convention used by miRBase, i.e., all the miRNAs can be identified by their miRNA name/identifier (e.g., hsa-let7-5p for human, mmu-let7a for mouse, and cel-mir-1018 for worm).

Target mRNA information has to be entered into the provided text box. For human and mouse, either Genbank accession number or Gene symbol, as assigned by the HUGO Gene Nomenclature Committee (HGNC), can be provided. For worm, Wormbase ID is required. For search result display through an interactive site viewer, a user can choose to display only the most relevant site features for each binding site, or the complete list of features. The most relevant features are considered by us to be the most informative. They were selected from those used in the development of the prediction models. A user may choose to input merely miRNAs while leaving the target input box blank. In this case, the database server will retrieve predicted sites for the entire transcriptome assembled for the species. The user can also choose to input merely mRNA IDs, which will prompt the database server to identify all miRNAs assembled for the species that have binding sites on those mRNAs. This can be useful, e.g., when the question is whether an mRNA is targeted by any miRNA. A database search is typically instantaneous. However, if the database is queried with only miRNA(s) without target information, the search will take minutes. Finally, the user can use a drop down menu to filter out miRNA binding sites with logistic probabilities below the specified threshold.

**Output of database query**

Relevant data in the database are retrieved in response to a specific database query and are available through both an interactive site viewer and downloadable files. For the interactive site viewer, the data is classified into three mRNA regions (5’ UTR, CDS and 3’ UTR) and seed and seedless sites. To facilitate online viewing, the number of sites displayed in the interactive area is limited to top-ranked sites according to the decreasing order of their logistic probabilities. By
default, 100 binding sites are displayed. Alternatively, the user can choose to display the top 250, 500 or 1,000 sites. The results of a search are also available for download as text files, wherein all of the retrieved binding sites are listed. The interactive viewer presents the results with either the most relevant site features or all of the site features as specified by the user in the input page. The downloadable text files provide all site features. In the text files, features marked with an asterisk are those used in the model computations of the logistic probabilities. In addition to comprehensive sequence, thermodynamic and target structural features (Table 1), a high resolution PDF diagram of the conformation of the miRNA:target hybrid is also provided. The diagram was developed to be high quality so that the user can choose to use them for publication purposes. When both the miRNA and the mRNA were included in the CLIP study for the prediction model development, an indicator field named “CLIP” will be given a value of 1 if the predicted site is support by the CLIP data, and 0 otherwise. CLIP studies are limited to abundant miRNAs and expressed transcripts. When either the miRNA or the mRNA was absent in the CLIP study, a value of “NA” is assigned to the CLIP indicator. In the database, less than 1% of sites have a CLIP indicator value of 0 or 1. Thus, our prediction data complement the CLIP data. A file providing definitions of site features is available via the link for ‘Feature definitions’ under the table listing predicted sites.

Illustration of database search

For an illustration of the database search, Figure 1 shows the input screen for a query starting with ‘Human (V-CLIP; NCBI RefSeq Build 36.3)’ selected in the species dropdown menu. From the dropdown list of miRNAs, hsa-7a-5p, hsa-7b-5p, and hsa-7d-5p were selected. In addition, two miRNAs, hsa-let-7c and hsa-let-7e were manually entered. For mRNA targets, accession numbers NM_0000024, NM_000021, NM_000017, and Gene symbol Lin54 were entered. Next, the option of “Show predictions with the most relevant features in interactive viewer” was selected. Finally, a logistic probability threshold of 0.6 was selected. The “Search” button was then clicked for submitting the query input information for processing by the database server.

Upon completion of data retrieval by the database server, the user is presented with an interactive site viewer (Figure 2). By default, the list of the top 100 sites is displayed in decreasing order of logistic probabilities. An alternative number of sites can be selected from a dropdown menu. The tab for “3’ UTR-seedless” was selected for presenting seedless sites in
the 3’ UTR of the target. For example, the first entry in the site table has a logistic probability of 0.8669, which indicates a high confidence in this predicted site. A rather low value of −21.26 kcal/mol for ΔG_{total} indicates a high structural accessibility at the target site\textsuperscript{17}. In the “Hybrid Conformation” column, a link is provided for a high resolution PDF diagram of the conformation of the miRNA:target hybrid at the predicted site. Clicking this link will open the diagram in a new tab or window, depending on the configuration of the user’s web browser. Multiple windows/tabs facilitate comparison of hybrid conformations for multiple binding sites. Figure 3 shows hybrid diagrams for a seed site and a seedless site.

Under the interactive site viewer, links are provided for downloading files of the query results for the six combinations of regions and site types (Figure 2). For site feature information, the downloadable files provide all site features whereas the interactive viewer displays either all or the most relevant features as selected by the user in the query input page. The user can initiate a new search by clicking on the link at the bottom of the page.

Conclusions

STarMirDB is a new bioinformatics resource for facilitating miRNA target studies. The current release of database includes 96,302 mRNAs and 3,446 miRNAs for human, mouse and worm. It will be periodically updated and likely extended to other species. It presents predictions for all three mRNA regions and for both seed and seedless sites. Importantly, it presents a probability for each site as an indicator of confidence in the prediction. In addition to use for visualization and publication, high quality diagrams of miRNA:target hybrids can facilitate design of nucleotide mutations for experimental validation of binding sites. The option for search by developmental stage shall be useful for studies of miRNAs in worm. The unique tools from STarMirDB will complement the existing miRNA target resources for computational predictions and experimental target data. The database can retrieve miRNA binding sites for single or multiple miRNAs and/or one or more targets. For example, this capability will be useful for elucidating miRNA regulation of genes of interest. It will also be useful in miRNA overexpression and knockout studies, wherein differentially expressed genes can be further examined by prediction and validation of miRNA binding sites.

We have also developed STarMir, a web server for prediction of miRNA binding sites\textsuperscript{36}. STarMir and STarMirDB are complementary tools. While the database allows fast search of pre-
computed results, STarMir makes predictions for any miRNA:mRNA pair from any species of interest. For example, the user can use STarMir in making predictions for a new isoform absent in the current database release.

The provision of extensive predictions of seedless sites (i.e., non-canonical sites) is a major feature for both the database and the web server. The functionality of seedless sites has been demonstrated by numerous studies based on diverse methods, which include reporter assays, nucleotide mutation analysis, analysis of microarray data, analysis of proteomics data, and phenotypic analysis\textsuperscript{11, 37-43}. However, a study primarily based on microarray data failed to find support for functional seedless sites\textsuperscript{44}. Further experimental investigations will be helpful for addressing this lack of consensus. Our tools will facilitate experimental testing of predicted seedless sites, especially those with high logistic probabilities.

**Acknowledgements**

The Bioinformatics 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.).

**Author Contributions**

Y.D. conceived and supervised the study. W.R. performed development, implementation and deployment of the database. C.L. and B.M. contributed to generation of data for the database; S.K. and J.L. performed testing of database interface. D. L. and A.W. wrote the initial software for the computation of several target site features used by the database, and C.C. provided hardware and system support. W.R., S.K. and Y.D. wrote the paper with contributions from all authors. All authors read and approved the final manuscript.

**Disclosure of Potential Conflicts of Interest**

No potential conflicts of interest were disclosed.

**Figure legends**

**Figure 1.** An illustrative example of the STarMirDB input page. Binding sites are searched for miRNAs selected from a pre-stored list as well as manually entered by a user, multiple targets,
and a specific logistic probability threshold selected by the user. The option of the most relevant site features is selected for output display.

**Figure 2.** STarMirDB output page for the default display of top 100 sites, with the tab selected for displaying seedless sites in the 3’ UTR.

**Figure 3.** Conformation diagrams of miRNA:target hybrids for a seed site (A) and a seedless site (B).
<table>
<thead>
<tr>
<th>Description of site information and features for STarMirDB output</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Site ID</strong></td>
</tr>
<tr>
<td><strong>Target</strong></td>
</tr>
<tr>
<td><strong>Gene</strong></td>
</tr>
<tr>
<td><strong>miRNA</strong></td>
</tr>
<tr>
<td><strong>Target_Len</strong></td>
</tr>
<tr>
<td><strong>Site_Position</strong></td>
</tr>
<tr>
<td><strong>Seed_Position</strong></td>
</tr>
<tr>
<td><strong>Seed_Type</strong></td>
</tr>
<tr>
<td><strong>Site_Access</strong></td>
</tr>
<tr>
<td><strong>Seed_Access</strong></td>
</tr>
<tr>
<td><strong>Upstream_Access (# nt)</strong></td>
</tr>
<tr>
<td><strong>Dwstream_Access (# nt)</strong></td>
</tr>
<tr>
<td><strong>Upstream_AU (# nt)</strong></td>
</tr>
<tr>
<td><strong>Dwstream_AU (# nt)</strong></td>
</tr>
<tr>
<td><strong>Site_Location</strong></td>
</tr>
<tr>
<td><strong>3’_bp</strong></td>
</tr>
<tr>
<td><strong>Site_Consv</strong></td>
</tr>
<tr>
<td><strong>Seed_Consv</strong></td>
</tr>
<tr>
<td><strong>Offseed_Consv</strong></td>
</tr>
<tr>
<td><strong>dG_hybrid</strong></td>
</tr>
<tr>
<td><strong>dG_nucl</strong></td>
</tr>
<tr>
<td><strong>dG_total</strong></td>
</tr>
<tr>
<td><strong>LogitProb</strong></td>
</tr>
<tr>
<td><strong>Target_Mismatch</strong></td>
</tr>
<tr>
<td><strong>Target_Match</strong></td>
</tr>
<tr>
<td><strong>Mir_Match</strong></td>
</tr>
<tr>
<td><strong>Mir_Mismatch</strong></td>
</tr>
<tr>
<td><strong>Hybrid Conformation</strong></td>
</tr>
<tr>
<td><strong>Target_Mismatch:</strong></td>
</tr>
<tr>
<td><strong>Target_Match:</strong></td>
</tr>
<tr>
<td><strong>Mir_Match:</strong></td>
</tr>
<tr>
<td><strong>Mir_Mismatch:</strong></td>
</tr>
</tbody>
</table>
References

44. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife 2015; 4.
Select one or more microRNAs from the dropdown list, or enter one or more microRNA names into the text box.

Enter the Genbank accession number(s) or gene symbols(s) of the target(s)

Select a probability. No site with a logistic probability less than the selected value will be displayed.
### Prediction model training data (species): V-CLIP (Human)

<table>
<thead>
<tr>
<th>Target</th>
<th>Gene</th>
<th>miRNA</th>
<th>Site_Start</th>
<th>Site_End</th>
<th>Hybrid Conformation</th>
<th>LogitProb</th>
<th>Site_Consv</th>
<th>AG_Hybrid</th>
<th>AG_aucl</th>
<th>AG_total</th>
<th>CLIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_194282</td>
<td>LIN54</td>
<td>hsa-let-7b-5p</td>
<td>5119</td>
<td>5127</td>
<td>view</td>
<td>0.9126</td>
<td>0.7227</td>
<td>-18.7</td>
<td>-5.935</td>
<td>-17.3</td>
<td>0</td>
</tr>
<tr>
<td>NM_194282</td>
<td>LIN54</td>
<td>hsa-let-7b-5p</td>
<td>5119</td>
<td>5140</td>
<td>view</td>
<td>0.879</td>
<td>0.2959</td>
<td>-21.6</td>
<td>-6.089</td>
<td>-20.19</td>
<td>0</td>
</tr>
<tr>
<td>NM_194282</td>
<td>LIN54</td>
<td>hsa-let-7c</td>
<td>4973</td>
<td>5003</td>
<td>view</td>
<td>0.8733</td>
<td>1</td>
<td>-17.8</td>
<td>-5.183</td>
<td>-13.43</td>
<td>1</td>
</tr>
<tr>
<td>NM_194282</td>
<td>LIN54</td>
<td>hsa-let-7c</td>
<td>4973</td>
<td>5005</td>
<td>view</td>
<td>0.8681</td>
<td>1</td>
<td>-17.9</td>
<td>-5.223</td>
<td>-13.29</td>
<td>1</td>
</tr>
<tr>
<td>NM_000021</td>
<td>PSEN1</td>
<td>hsa-let-7b-5p</td>
<td>1793</td>
<td>1812</td>
<td>view</td>
<td>0.8669</td>
<td>0.5912</td>
<td>-21.1</td>
<td>-7.179</td>
<td>-21.26</td>
<td>0</td>
</tr>
<tr>
<td>NM_194282</td>
<td>LIN54</td>
<td>hsa-let-7c</td>
<td>4973</td>
<td>5000</td>
<td>view</td>
<td>0.8621</td>
<td>1</td>
<td>-15.9</td>
<td>-4.826</td>
<td>-11.53</td>
<td>1</td>
</tr>
<tr>
<td>NM_000024</td>
<td>ADRB2</td>
<td>hsa-let-7a-5p</td>
<td>1524</td>
<td>1544</td>
<td>view</td>
<td>0.8565</td>
<td>0.8208</td>
<td>-16.8</td>
<td>-4.358</td>
<td>-16.44</td>
<td>0</td>
</tr>
<tr>
<td>NM_000024</td>
<td>ADRB2</td>
<td>hsa-let-7b-5p</td>
<td>1516</td>
<td>1539</td>
<td>view</td>
<td>0.8543</td>
<td>0.659</td>
<td>-17.3</td>
<td>-5</td>
<td>-17.45</td>
<td>0</td>
</tr>
<tr>
<td>NM_000024</td>
<td>ADRB2</td>
<td>hsa-let-7a-5p</td>
<td>1524</td>
<td>1539</td>
<td>view</td>
<td>0.8528</td>
<td>0.7788</td>
<td>-15.2</td>
<td>-4.876</td>
<td>-15.89</td>
<td>0</td>
</tr>
<tr>
<td>NM_000024</td>
<td>ADRB2</td>
<td>hsa-let-7c</td>
<td>1524</td>
<td>1544</td>
<td>view</td>
<td>0.8525</td>
<td>0.8208</td>
<td>-16.5</td>
<td>-4.173</td>
<td>-16.14</td>
<td>0</td>
</tr>
<tr>
<td>NM_194282</td>
<td>LIN54</td>
<td>hsa-let-7d-5p</td>
<td>4779</td>
<td>4798</td>
<td>view</td>
<td>0.8524</td>
<td>0.9964</td>
<td>-15.5</td>
<td>-0.462</td>
<td>-8.907</td>
<td>0</td>
</tr>
<tr>
<td>NM_194282</td>
<td>LIN54</td>
<td>hsa-let-7b-5p</td>
<td>4973</td>
<td>5003</td>
<td>view</td>
<td>0.8505</td>
<td>1</td>
<td>-17.8</td>
<td>-5.023</td>
<td>-13.43</td>
<td>1</td>
</tr>
</tbody>
</table>

### Feature definitions

### Download results from the search above
- Features and predictions for 3' UTR-seed sites
- Features and predictions for 3' UTR-seedless sites
- Features and predictions for CDS-seed sites
- Features and predictions for CDS-seedless sites
- Features and predictions for 5' UTR-seed sites
- Features and predictions for 5' UTR-seedless sites

Download options: None found.
\[ \Delta G = -23.2 \text{ kcal/mol} \]

hsa–let–7a–5p

\[
\begin{array}{c}
\text{NM}_000024\\
\text{Fig. 3A}
\end{array}
\]

\[ \Delta G = -16.8 \text{ kcal/mol} \]

hsa–let–7a–5p

\[
\begin{array}{c}
\text{NM}_000024\\
\text{Fig. 3B}
\end{array}
\]