Vir-Mir db: prediction of viral microRNA candidate hairpins

Sung-Chou Li¹, Cheng-Kai Shiau² and Wen-chang Lin¹,2,*

¹Institute of BioMedical Informatics, National Yang-Ming University, Taipei 112 and ²Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan, Republic of China

Received June 15, 2007; Revised July 17, 2007; Accepted July 26, 2007

ABSTRACT

MicroRNAs have been found in various organisms and play essential roles in gene expression regulation of many critical cellular processes. Large-scale computational prediction of miRNAs has been conducted for many organisms using known genomic sequences; however, there has been no such effort for the thousands of known viral genomes. Some viruses utilize existing host cellular pathways for their own benefit. Furthermore, viruses are capable of encoding miRNAs and using them to repress host genes. Thus, identifying potential miRNAs in all viral genomes would be valuable to virologists who study virus–host interactions. Based on our previously reported hairpin secondary structure and feature selection filters, we have examined the 2266 available viral genome sequences for putative miRNA hairpins and identified 33 691 hairpin candidates in 1491 genomes. Evaluation of the system performance indicated that our discovery pipeline exhibited 84.4% sensitivity. We established an interface for users to query the predicted viral miRNA hairpins based on taxonomic classification, and a host target gene prediction service based on the RNAhybrid program and the 3'UTR gene sequences of human, mouse, rat, zebrafish, rice and Arabidopsis. The viral miRNA prediction database (Vir-Mir) can be accessed via http://alk.ibms.sinica.edu.tw.

INTRODUCTION

MicroRNAs (miRNAs) are endogenous non-protein-coding RNAs that are ~22-nt long. They negatively regulate gene expression by complementary binding to the 3'-UTR regions of target genes (1). Since the first discovery of miRNA in Caenorhabditis elegans, thousands of miRNAs have been computationally and/or experimentally identified in many organisms, including mammals, invertebrates, insects and plants (2). It has been shown that miRNAs can function in various physiological pathways. Plant miRNAs can regulate development in embryos, leaves and floral meristems (3,4). Mammalian miRNAs participate not only in developmental regulation (5) but also in pathogenesis if they are dysfunctional (6). Even with limited information about miRNA gene structures and their target genes' recognition selections, several genome-wide bioinformatics approaches have been recently employed to identify these important molecules (7–9). Establishing such a bioinformatic resource for miRNA will be beneficial for daily experimental research in the biological laboratories.

Similar to many eukaryotic organisms, viruses also encode miRNAs (10,11). To date, at least 82 viral-associated miRNAs have been identified from eight different viruses (2). Because viruses are often parasitic, these viral miRNAs may target important host genes to reduce the host cell defense and to control host cell biogenesis. Human herpes virus 4 (Epstein-Barr virus, EBV) represses several host genes, including those encoding B cell-specific chemokines and cytokines, transcriptional regulators and components of signal transduction pathways, by means of virus-encoded miRNAs (11). Moreover, HIV-1 also has been proposed to enhance viral infection capability by repressing host immune system genes via specific HIV viral miRNAs (10).

MOTIVATION

It would be beneficial for biomedical researchers who study virus–host interactions to identify potential viral miRNAs and their target genes in hosts. Although a new database collecting viral gene targets of host miRNAs has been reported recently (12), but there is no genome-wide miRNA prediction for all completed viral genome sequences. We have modified our previous miRNA discovery pipeline to predict viral miRNA hairpins (13). Here, we present our research effort on candidate hairpin discovery of virus-encoded miRNAs from all virus genomes available from the National Center for

*To whom correspondence should be addressed. Tel: +886226523967; Fax: 886227827654; Email: wenlin@ibms.sinica.edu.tw

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Biotechnology Information (NCBI), National Institutes of Health, USA. It would be beneficial to have a user-friendly web interface to query the predicted viral miRNA hairpins based on their familial taxonomic classification. In addition, a host target gene prediction service is established based on the RNAhybrid program and the parsed 3'-UTR gene sequences of human, mouse, rat, zebrafish, rice and Arabidopsis species to search for potential target genes of user’s interest.

**DATA GENERATION**

**Predicting hairpins from virus genome scaffold sequences using Srnaloop**

During miRNA maturation, the full-length pre-miRNA transcript needs to form a hairpin (stem-loop) structure. The secondary structure is folded via intramolecular base pairing and has been the most significant criterion for computational identification of miRNAs (13–16). The Srnaloop program, developed by Grad et al. (14), has been used to identify putative hairpin secondary structures from viral genomic sequences. All the viral genome sequences were obtained on 4 May 2006 from NCBI, National Institutes of Health, USA. The NCBI viral1.genomic.fna file comprises the genomic sequences of 2266 viruses, and the file size is 41 MB. The optimized parameters for Srnaloop were as reported previously (13), except for the parameters ‘–90’ and ‘–17’. These parameters, modified based on known viral pre-miRNAs, are specific for identifying hairpins that are up to 90 bases long and have a score of at least 17.

**Sequence and structural features filter**

To reduce the number of falsely predicted miRNAs based on hairpin structure alone, we have included several additional selection filters, namely GC content, minimum free energy of the core of hairpin structure (core mfe), minimum free energy of the whole hairpin (hairpin mfe) and the ratio of core mfe to hairpin mfe (ch ratio) as defined in a previous publication (13). To distinguish authentic miRNA candidates, we also employed additional classification criteria specific to known miRNAs. We first investigated the features of known viral miRNAs as well as their precursors and then determined the reference range values of the selected features. The reference range values are listed in Table 1. Therefore, a candidate hairpin with sequence and structural features within the optimized reference range values was considered to be a positive miRNA hairpin.

| Table 1. Distribution range of sequence and structure features from known viral miRNAs |
|---------------------------------|-----------------|---------------|---------------|--------------|
|                                 | GC content     | core mfe      | hairpin mfe   | ch ratio     |
| Distribution                    | ~34–65         | ~37.5 to ~14.5| ~22 to ~59.2  | ~51–100      |
| Reference value                 | ~38–65         | ~37.5 to ~19.4| ~50.9 to ~28.2| ~51–98       |

From the known viral pre-miRNA information, we evaluated the distributions of quantifiable sequence and structure features, namely GC content, core mfe, hairpin mfe and the ratio of core mfe to hairpin mfe. Because extreme values existed, we adopted the reference value to be used in the sequence and structure feature filter.

**Application of open reading frame (ORF) feature filter**

When predicting miRNAs in organisms other than viruses, a ‘conservation pattern’ is usually used to find evolutionarily conserved miRNA candidates (13–16). However, this feature is not applicable for most viral miRNA prediction here. Cullen found only a few miRNAs located in protein-coding transcripts (17), which is used as a criterion to predict eukaryotic miRNAs (13,14). Therefore, we screened our candidate hairpins to check if they overlap the ORFs belonging to the same virus. The sequences of the ORFs were extracted from viral1.protein.gff file downloaded from NCBI on 4 May 2006. Because viruses encode functional genes redundantly in a highly compact genomic region, we cannot exclude the possibility that viral miRNAs may overlap other viral protein-coding genes. This concern is more significant for small genome RNA viruses. Therefore, we still considered the hairpins overlapping ORFs to be positive miRNA hairpins, but we marked the NP accession numbers of the overlapped ORF protein in the dataset for reference.

**DATABASE STATISTICS**

**Candidate hairpins predicted by Srnaloop and additional feature filters**

We identified roughly 514874 candidate hairpins from all viral genomic sequences by means of Srnaloop from 2266 viral genomes. They were then divided into 5P and 3P stem arms as putative miRNA candidates as done previously (13), because an miRNA could be located in either the 5’ arm or 3’ arm of the stem-loop precursors. Following the screen with the sequence and structure feature filter, 33691 candidates survived. Among them, 5306 candidate hairpins do not overlap any ORF of the correspondent virus, implying they have higher likelihood to be authentic, but not absolutely as described earlier.

All the hairpin information can be accessed from the web interface. We compared the sequences of these putative viral candidate miRNAs with the sequences of known miRNAs in miRBase (release 9.0). The positive match criterion required at least an 18-nt matched length with >90% identity. With this high stringency of sequence similarity, only seven candidates with unique IDs of 821661, 1341202, 1343051, 1408482, 1430283, 3325175 and 3663618 showed positive matches. For example, candidate 821661, identified from Cercopithecine monkey herpes virus 15, is similar to ebv-miR-BART19; candidates 1341202, 1343051 and 325175, identified from chimpanzee cytomegalovirus, are similar to hcmv-miR-UL112, hcmv-miR-UL148D and hcmv-miR-UL112, respectively. It is possible that more matches could be found if the stringency of sequence similarity was lowered.
To assess the efficacy of our pipeline system, we performed both sensitivity and specificity tests. Because it is much easier to identify hairpin structures in sequences of known hairpins than from entire genomic sequences, we calculated the number of predicted candidate hairpins after each filter procedure in our pipeline to determine the recovery rate of reported miRNAs. Originally, the test dataset consisted of 64 known pre-miRNAs belonging to five viruses from miRBase. As shown in Table 2, 63 known miRNAs remained after the Srnaloop step, 54 after the sequence and structure feature step, and 50 after the ORF comparison step of the selection. These results imply that the recovery rate is close to 98% after Srnaloop hairpin prediction, 84% after Srnaloop hairpin prediction and sequence and structure feature filter and 78% after additional ORF filter.

For the specificity examination on the pipeline, the negative dataset generation procedure was similar to the one described by Sewer et al. (18). This procedure is based on the fact that the fraction of miRNA-encoding sequences in the genome is very small; therefore, randomly extracted sequences are extremely unlikely to encode miRNAs. We randomly extracted 19,200 (64 x 300) sequence fragments (90 bp in length) from viral genomic sequences. These 19,200 randomly chosen sequences of 1.8 Mb were applied to our discovery pipeline under the same hairpin identification parameters and sequence and structural filter criteria. As a result, we obtained 533 presumed false positives from three independent experiments (175, 176 and 182 predicted candidates, respectively), corresponding to an average of 178 false positives. This would give an estimated 1% false positive prediction in a randomly generated 20,000 sequences. With over 40,000,000 nt from virus genomes, we could estimate 400 false positive predictions (using non-overlapping 90 bp length windows). Therefore, the best specificity we could achieve is close to 98%, and the actual prediction rate would lower depending on the viral genome structures (DNA versus RNA viruses) and gene organizations.

**WEB INTERFACE FOR VIR-MIR db**

Due to the large number of prediction results, we constructed a user-friendly web interface to present the viral candidate miRNAs. As shown in Figures 1 and 2, the

---

**Table 2. System performance test and the number of candidate hairpins**

| Known viral miRNAs (species) | Number of miRNAs | Number of miRNAs after Srnaloop filter | Number of miRNAs after sequence and structure filter | Number of miRNAs after ORF filter |
|-----------------------------|------------------|---------------------------------------|-----------------------------------------------------|---------------------------------|
| HCMV                        | 11               | 11                                    | 8                                                   | 6                               |
| KSHV                        | 13               | 12                                    | 8                                                   | 7                               |
| RLCV                        | 16               | 16                                    | 14                                                  | 14                              |
| SV40                        | 1                | 1                                     | 1                                                   | 0                               |
| EBV                         | 23               | 23                                    | 23                                                  | 23                              |
| Total (Sensitivity)         | 64 (100%)        | 63 (98.4%)                            | 54 (84.4%)                                          | 50 (78.1%)                      |
| Total number of predicted miRNA candidate hairpins | 2266 virus | 514,874 | 33,691 | 5,306 |

To test the performance of our system, we calculated the number of known viral pre-miRNAs remaining after each filter process. Originally, there were 64 known pre-miRNAs belonging to five viruses. The number of hairpins predicted from 2266 virus genomes is also listed.

**Databases currently available are list below:**

**Predicted viral microRNA hairpins - classified by NCBI Taxonomy list:**

(Accumulated Numbers of Hairpin Candidates)

Deltavirus (0)
dsDNA viruses, no RNA stage (2,633)
dsDNA viruses (379)
environmental samples (0)
Retro-transcribing viruses (386)
Satellites (59)
ssDNA viruses (410)
ssRNA negative-strand viruses (167)
ssRNA positive-strand viruses, no DNA stage (2325)
unclassified viruses (374)

**Supplementary Information:**

Known virus microRNAs extracted from miRBase

Search protocol: (eg. NC_007605: Search by GenBank Id., RefSeq Acc., or viral scientific name)

Figure 1. Viral miRNA candidate hairpin query web interface. Users may query the putative miRNAs of a specific virus by using the hierarchical menu or by using the search function with GenBank identifier, RefSeq accession number or viral scientific name.
classification and arrangement of the viral candidates are according to the taxonomy table of NCBI, so users can query the viral miRNAs more easily. Users may query the putative miRNAs from a specific virus by using the hierarchical menu or by using the simple search function. A keyword search can be performed, but users are recommended to use it with the GenBank identifier or RefSeq accession number for better search result, e.g. 30844336 or NC_003663 for Cowpox virus.

The numbers in the right parentheses of each classification level indicate the numbers of candidate hairpins found in that level. For example, Atadenovirus has 18 candidate hairpins, which are cumulated from its four sub-classes of viruses, including Bovine adenovirus, Duck adenovirus A, Ovine adenovirus D and Possum adenovirus. We identified 16 candidate hairpins from Duck adenovirus A and no candidate hairpin from Bovine adenovirus D. In some cases, like Bovine adenovirus 4, 5, 8 and Possum adenovirus, we did not record the number of candidate hairpins in their right parentheses, which means there are no genomic sequences available for analysis at these classification levels.

After selecting a specific virus type, as illustrated in Figure 3 for human herpes virus 4, users may further sort the candidate hairpins according to the miRNA genomic location (default), score calculated by Srmaloop, or core mfe calculated byRNAfold. Moreover, for known pre-miRNAs, the sequences and locations of mature miRNAs are shown according to the known miRNA information. For example, both ebv-mir-BHRF1-1 and ebv-mir-BHRF1-3 have only one mature miRNA in their 5’ arms.

![Figure 2](https://example.com/fig2.png)

**Figure 2.** Expanded view of taxonomy table. The classification and arrangement of viral miRNA candidate hairpins are based on the taxonomy table of NCBI. The numbers in the right parentheses of each classification level indicate the numbers of candidate hairpins found in those levels. By clicking on a link, users can see the expanded table. The numbers in parentheses indicate the number of candidate hairpins found in that particular viral genome.

![Figure 3](https://example.com/fig3.png)

**Figure 3.** Viral miRNA candidate hairpin web interface. Here, users can view all the predicted viral miRNA hairpin information for a particular virus. They may sort the candidate hairpins according to genomic location (default), score calculated by Srmaloop or core mfe calculated byRNAfold. The YP accession number in the NP reference protein column denotes the protein-coding ORF sequence overlapped by this candidate hairpin. Click the field header to sort on a field. Click it again to reverse the sorted list.
However, the pre-miRNA ebv-mir-BHRF1-2 has two mature miRNAs in its 5' arm and 3' arm. In addition, users may download the entire sequences of predicted candidate hairpins via the link at the bottom of each page. Besides, we also extracted all known viral miRNAs from miRBase into an independent table that also can be retrieved at http://alk.ibms.sinica.edu.tw/cgi-bin/miRNA/known_viral_miRNA.cgi.

**Predicting potential viral miRNA target genes using the RNAhybrid program**

It is of researchers' interest to learn about viral miRNAs and their host target genes. To assist biologists, we provide an integrated interface to identify possible host target genes of the viral miRNA in selected host genomes using the RNAhybrid program (19,20). Presently, as shown in Figure 4, we provided the 3'-UTR region sequences of reference genes from human, mouse, rat, zebrafish, rice and Arabidopsis as the search target dataset. We retrieved 3' UTRs based on the coding sequence positions acquired individually from the human.rna.gbff file (from NCBI on 19 December 2005), mouse.rna.gbff file (from NCBI on 19 December 2005), rat.rna.gbff file (from NCBI on 19 December 2005), zebrafish.rna.gbff file (from NCBI on 19 June 2006) and plant1 ~ 5.rna.gbff file (from NCBI on 5 August 2006).

Users may scan the 3'-UTR region sequences of reference genes for possible viral miRNA targets. When operating RNAhybrid, the pipeline first calculates the optimal free energy of a putative miRNA when the entire putative miRNA binds to a perfectly complementary target site, then it calculates the minimum free energy of RNA duplex (mfe of the miRNA/mRNA duplex). An alignment for which the RNA duplex mfe is more than 66% of its correspondent optimal free energy is regarded as a positive alignment as described by Krek et al. (21). A higher stringent 85% parameter is recommended. A critical issue on the target prediction is that it dramatically depends on the seed region located in the mature miRNA. However, it is difficult to correctly predict the mature miRNA following the hairpin prediction pipeline. Therefore, we have a helix constraint option to let users select different seed match region to be used in the target prediction pipeline. It is suggested this information should be used for preliminary exploration purpose without the mature miRNA information.

**DISCUSSION**

The available predicted viral miRNA candidate hairpins will certainly be beneficial to biologists who study interactions between viral miRNAs and host genes. There are many completed viral genome sequences in the database, but it is not a simple task for biologists to execute the entire bioinformatic pipeline to scan for putative miRNA hairpins. We have identified 33691 candidate hairpins from more than 2000 virus genome sequences. Here, we provide an easy-to-use web interface...
for examining predicted viral miRNA hairpins based on the previously published pipeline. To make this tool more user-friendly, we also included the target gene search function. Many functionally related miRNAs cluster together to facilitate their expression and functional regulation. For example, about half of the miRNAs identified in Drosophila are reported to exhibit this clustering phenomenon; they are initially co-transcribed from one polycistrionic transcript and further processed into distinct individual mature miRNAs (1,2,23). This phenomenon has been reported in EBV (11) and also was observed in our present results. As illustrated in Figure 3, two predicted candidate hairpins from this study are clustered in the transcript of BART together with several known EBV pre-miRNAs (11). This phenomenon is also found in the transcript of BHRF1, which contains two newly predicted candidate hairpins and three previously known pre-miRNAs, ebv-mir-BHRF1-1, ebv-mir-BHRF1-2 and ebv-mir-BHRF1-3. This clustering phenomenon indicates that these candidate hairpins may be authentic viral miRNAs.

ACKNOWLEDGEMENT

This research is partially supported by the Academia Sinica research grant support. Funding to pay the Open Access publication charges for this article was provided by the National Science Council, Taiwan.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116, 281–297.
2. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res., 34, D140–D144.
3. Reinhart,B.J., Weinstein,E.G., Rhoades,M.W., Bartel,D. and Bartel,D.P. (2002) MicroRNAs in plants. Genes Dev., 16, 1616–1626.
4. Chen,X. (2004) A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. Science, 303, 2022–2025.
5. Olsen,P.H. and Ambros,V. (1999) The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev. Biol., 216, 671–680.
6. Johnson,S.M., Grosshans,H., Shingara,J., Byrom,M., Jarvis,R., Cheng,A., Labourier,E., Reinert,K.L., Brown,D. et al. (2005) RAS is regulated by the let-7 microRNA family. Cell, 120, 635–647.
7. Megraw,M., Setyaputra,P., Corda,B. and Hatzigeorgiou,A.G. (2006) miRGen: a database for the study of animal microRNA genomic organization and function. Nucleic Acids Res. (Database issue):D149–155.
8. Hsu,P.W., Huang,H.D., Hsu,S.D., Lin,L.Z., Tsou,A.P., Tseng,C.P., Stadler,P.F., Wathietl.S. and Hofacker,I.L. (2006) miRNASat: genomic maps of microRNA genes and their target genes in mammalian genomes. Nucleic Acids Res., 34, D135–D139.
9. Setyaputra,P., Corda,B. and Hatzigeorgiou,A.G. (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. RNA, 12, 192–197.
10. Bennasser,Y., Le,S.Y., Yeung,M.L. and Jeang,K.T. (2004) HIV-1 encoded candidate micro-RNAs and their cellular targets. Retrovirology, 1, 43.
11. Pfeffer,S., Zavolan,M., Grasser,F.A., Chien,M., Russo,J.I., Ju,J., John,B., Enright,A.J., Marks,D. et al. (2004) Identification of virus-encoded microRNAs. Science, 304, 734–736.
12. Hsu,P.W., Lin,L.Z., Hsu,S.D., Hsu,J.B. and Huang,H.D. (2007) ViTa: prediction of host microRNA targets on viruses. Nucleic Acids Res., 35, D381–D385.
13. Li,S.C., Pan,C.Y. and Lin,W.C. (2006) Bioinformatic discovery of microRNA precursors from human ESTs and introns. BMC Genomics, 7, 164.
14. Grad,Y., Aach,J., Hayes,G.D., Reinhart,B.J., Church,G.M., Ruvkun,G. and Kim,J. (2003) Computational and experimental identification of C. elegans micro RNAs. Mol. Cell, 11, 1253–1263.
15. Iai,E.C., Tomancak,P., Williams,R.W. and Rubin,G.M. (2003) Computational identification of Drosophila microRNA genes. Genome Biol., 4, R42.
16. Berezikov,E., Guryev,V., van de Belt,J., Wienholds,E., Plasterk,R.H. and Cuppen,E. (2005) Phylogenetic screening and computational identification of human microRNA genes. Cell, 120, 21–24.
17. Cullen,B.R. (2004) Transcription and processing of human microRNA precursors. Mol. Cell, 16, 861–865.
18. Sewer,A., Paul,N., Landgraf,P., Aravin,A., Pfeffer,S., Brownstein,M.J., Tuschl,T., van Nimwegen,E. and Zavolan,M. (2005) Identification of clustered microRNAs using an ab initio prediction method. BMC Bioinformatics, 6, 267.
19. Kruger,J. and Rehmsteiner,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res., 34, W451–W454.
20. Rehmsteiner,M., Stellin,P., Hochsmann,M. and Giegerich,R. (2004) Fast and effective prediction of microRNA/target duplexes. RNA, 10, 1507–1517.
21. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C. et al. (2005) Combinatorial microRNA target predictions. Nat. Genet., 37, 495–500.
22. Du,T. and Zamore,P.D. (2005) microPrimer: the biogenesis and function of microRNA. Development, 132, 4645–4652.
23. Aravin,A.A., Lagos-Quintana,M., Yalcin,A., Zavolan,M., Marks,D., Snyder,B., Gaasterland,T., Meyer,J. and Tuschl,T. (2003) The small RNA profile during Drosophila melanogaster development. Dev. Cell, 5, 337–350.