In silico candidates miRNA prediction
From genome Dictyostelium discoideum

Kalaria R.1, Patel N.2
1 Department of Applied mathematics, The M.S. University of Baroda(Gujarat)India
2 Department of Microbiology, Naranlala college, Veer Narmad south Gujarat university, surat(Gujarat)India
* Email-rishee_kal@yahoo.com

Abstract
The discovery of microRNAs, almost 10 years ago, changed dramatically our perspective on eukaryotic gene expression regulation. However, the broad and important functions of these regulators are only now becoming apparent. Dictyostelium discoideum is one of the simplest studied eukaryote that possesses true multicellularity and provides unique advantages for studying fundamental cellular processes with powerful molecular genetic, biochemical, and cell biological tools. Small RNAs play crucial roles in regulation of gene expression in many eukaryotes. Different methods for elucidation of microRNA from the genome are described briefly. However, in absence of availability of microRNA sequence data from closely related species, the identification of microRNA in D. discoideum is challenging. Here we report the probable microRNA candidates from the social amoeba D. discoideum. We have used two approaches for prediction of microRNA from the social amoeba D. discoideum. Using comparative genomics approach, we identified 171 putative candidate microRNA and its precursor from D. discoideum genome sequence. In ab initio approach we have identified 79 putative candidate microRNA and its 116 precursor using neural network method.

Keywords: Dictyostelium discoideum, candidate microRNA, retrotransposon

Introduction

Dictyostelium discoideum: THE MODEL SYSTEM

Dictyostelium often referred to as “slime mold” or “social amoeba”, is one of the simplest studied eukaryote that possesses true multicellularity (Raper, 1984). The organism provides unique advantages for studying fundamental cellular processes with powerful molecular genetics, biochemical, and cell biological tools. These processes include cell motility, chemotaxis, cytokinesis, phagocytosis, endocytosis, signal transduction, and aspects of development such as cell sorting, pattern formation, and cell-type differentiation (Parent and Devreotes, 1999; Thomason et al., 1999; Gerisch and Weber, 2000). Recently, Dictyostelium was also described as a suitable host for pathogenic bacteria to study the process of infection conveniently (Skiwian et al., 2002). In addition, Dictyostelium has many of the experimental conveniences of S. cerevisiae and is probably the best experimentally manipulatable protozoan. A series of cell biological assays and molecular genetics tools can been used to study gene function (Eichinger et al., 1999).

MicroRNAs (miRNAs) are ~22-nucleotide (nt) noncoding RNA molecules generated from stem-loop hairpin structures called miRNA precursors (premiRNAs) of ~80 nt. These premiRNAs are first transcribed as longer RNAs (Lee et al., 2002) and are then processed by the RNAase III enzyme, Drosha (Lee et al., 2003). Another RNAase III enzyme, Dicer (Grishok et al., 2001; Huttunen et al., 2001; Ketting et al., 2001), cuts these ~80-nt premiRNAs to release ~22-nt mature miRNA. MicroRNAs (miRNAs) were first discovered in 1993, when the miRNA lin-4 was determined to downregulate expression of the gene lin-14 in Caenorhabditis elegans by the Ambros and Ruvkun laboratories (Lau et al., 2001). However, since there is no homolog to lin-4 in other species, this discovery was considered to be unique. Specific and potent silencing of genes by double stranded RNA (RNAi) was discovered in 1998, and the discovery of the miRNA let-7 in 2000, with homolog’s in other species including humans, showed that miRNAs are quite common in eukaryotes. There are now known to be multiple types of small noncoding RNA, with miRNAs being the largest family of noncoding RNAs involved in gene silencing. They are highly conserved by evolution, suggesting a fundamental biological function. The turn of the century brought the realization that miRNAs form a large new class of ncRNAs that provide a ubiquitous and powerful mechanism for RNA-mediated control of gene expression. It has been estimated that miRNA regulates up to one-third of human genes (Griffiths-Jones S, 2004).

Recent studies have implicated miRNAs in such fundamental processes as cell development, differentiation, communication and death. Therefore, it is not surprising that miRNAs have been found to be involved in such disparate areas as hematopoiesis, insulin secretion, nervous system patterning and human cancer development. D. discoideum has long served as a laboratory model system for development as it exhibits a variety of developmental stages (Bonner, 1959; Loomis, 1982; Maeda et al., 1997; Kessin, 2001). microRNAs target in
particular developmental genes, in contrast, genes involved in functions common to all cells, such as gene expression, have very few microRNA target sites, and seem to be under selection to avoid targeting by microRNAs.

Investigate the small RNA (18–26 nt) profile of D. discoideum during growth and development (Andrea Hinas et al 2007) For this reason, They cloned and sequenced pooled small RNAs from growing single cells and from two different multicellular stages (16 and 24 h of development). The specific characteristics of different small RNA classes determine their accessibility to different cloning methods. Most clones represented small RNAs derived from the DIRS-1 retrotransposon (68.7% and 37.5% in the 5'-ligation-dependent and 5'-ligation-independent libraries, respectively). Skipper is the second most abundant retrotransposon in the D. discoideum genome. In contrast to DIRS-1, Skipper is flanked by direct LTRs, and no antisense RNA has yet been reported to be transcribed from this retrotransposon. Computational approach to predict microRNA from genomic sequence provides a valuable complement to the experimental methods. It was the discovery of the C. elegans miRNA let-7 that led to computational approaches for discovering other novel miRNAs. This principal approaches leads to identify the microRNA in most plants and animals.

Most of the Computational approaches developed are based on simple homology search using BLASTN(Hongyu 2003). Due to conservation of microRNA sequence in closely related species this approaches were able to identify many orthologues in numerous species. But in some of the cases just performing homology search were not able to identify the orthologues due to cross- species sequence divergence, so the alternative approaches were developed which uses the RNA FOLD prediction to identify the sequences likely to form stable stem-loop structure. For example the RNA structure prediction program MFOLD (Zuker 2003) which predict RNA structure by free energy minimization was used to identify novel miRNAs in C.elegans.

**Methodology:**

**Sequence and annotation data:**

The complete set of genome of D. discoideum was was retrieved from Dictybase database This database not only provides access to genomic data including functional annotation of genes, gene products and chromosomal mapping, but also to extensive biological information such as mutant phenotypes and corresponding reference material. Release miRBase (13.0) database(Griffiths-Jones S, et al 2007) contains 9539 entries representing hairpin precursor miRNAs, expressing 9169 mature miRNA products, in 103 species. NCBI standalone ( blast Blast2.2.21) BLAST uses a heuristic algorithm that seeks local as opposed to global alignments and is therefore able to detect relationships among sequences that share only isolated regions of similarity (Altschul et al. 1990).it is used to run BLAST searches against private, local databases, or downloaded copies of the NCBI databases.

**Total RNA extract from the Retrotransposons and the Skipper region of genome**

Total small RNA of 12-30nt extract from the dysm(2495 rna sequence),dynl(3654 RNA sequence)and skipper(9 RNA sequence) were taken from paper “The small RNA repertoire of Dictyostelium discoideum and its regulation by components of the RNAi pathway” ( Andrea Hinas et al 2007 ).

**EumiR:**

Eukaryotic microRNA precursor prediction web server of IGIB. It extract the positive precursor from query sequences based on the SVM model library. The query interface has the form for querying multiple sequences which have potential to form hairpin secondary structures, each delimited by a line break.in EumiR result interface we get positive and negative sequences EumiR has better accuracy and sensitivity as compare to mirabel, BayesmiRNAfind.

**Mipred:**

MiPred(Peng jiang et.al 2007): classification of real and pseudo microRNA precursors using random forest prediction model with combined features. To distinguish the real pre-miRNAs from other hairpin sequences with similar stem-loops (pseudo pre-miRNAs), a hybrid feature which consists of local contiguous structure-sequence composition, minimum of free energy (MFE) of the secondary structure and P-value of randomization test is used. Besides, a novel machine-learning algorithm, random forest (RF), is introduced. The results suggest that our method predicts at 98.21% specificity and 95.09% sensitivity. When compared with the previous study, Triplet-SVM-classifier, our RF method was nearly 10% greater in total accuracy. Further analysis indicated that the improvement was due to both the combined features and the RF algorithm. Given a sequence, MiPred decides whether it is a pre-miRNA-like hairpin sequence or not. If the sequence is a pre-miRNA-like hairpin, the RF classifier will predict whether it is a real pre-miRNA or a pseudo one. There were basically two approaches used for finding the candidates microRNA from the genome.
A) Comparative genomics approach

Using standalone blast, taking mature miRNA from mirbase as database and genome short seq.(upto 2mb bp) as query sequences. Taking small query seq from genome (upto 2mb bp). Blast with our training data set which are mirbase mature seq. (GUI output for each chromosome done). Retrieve blast hit from result and input in run matcher tool (emboss package) for finding homology and extending predicted miRNA seq to find precursor seq.Run Matcher identifies local similarities in two input sequences using a rigorous algorithm based on Bill Pearson's align application, version 2.0u4 (Feb. 1996). It finds the best local alignments between two sequences. The substitution matrix, gap insertion and extension penalty are specified. . In the Runmatcher miRNA homology candidates file is given as input and as an output we get probable miRNA precursor candidates. In standalone blast take different database and query sequence to get different output for the conformation of the result. Whatever result come then go for clustalW for finding the common and unique candidates microRNA, then go for the secondary structure prediction of RNA of the validation and checking its thermodynamics stability using online web server like MFOOLD, CENTROIFOLD etc.

B) Ab initio approach

In ab initio approach for chromosome 1, we use IGIB online web server for Eukaryotic microRNA (EuMiR) prediction, where we give 150nt sequence in iterative manner to server and getting output as positive or negative precursor sequences. Then the positive sequences from EuMiR were checked for secondary structure in MFOOLD for its thermodynamics stability and further the sequences showing the correct secondary structure and thermodynamic stability were classified as real or pseudo precursor sequences using MiPred web server.Next for rest of the chromosomes, we created small java code for giving input sequence in EuMiR web server and retrieve the positive precursor sequences. The subsequent steps were same as described about for chromosome 1. Further, using the positive sequence as database and dynl and dysm as query RNA sequences Standalone BLAST was performed to find the sequences giving 100% identity to dynl and dysm small RNA sequences.

Result and Discussion:

A) Comparative genomics approach

For elucidation of microRNA from the D. discoideum genome, in this approach mature microRNA sequences from mirbase was taken as a reference database. The D. discoideum genome was subdivided into overlapping fragments of 2 megabases and matched with the reference database using standalone blast and the hits obtained were further used for prediction of miRNA precursors. Standalone blast was used with default settings for this purpose. A total of 171 sequences were predicted as putative miRNA from the D. discoideum genome by this search method. The hits obtained from the blast search using mirbase miRNA sequences as reference database were taken as input for Run matcher. In the Run matcher program the hits were matched to the D. discoideum genome and these matched sequences were then extended on both sides to give the Precursor microRNA from genome. For finding the precursor miRNA sequences Run matcher gives the sequences both showing homology to the input miRNA and reverse complement sequences of input miRNA sequences. A total of 171 sequences were predicted as precursor miRNA from the D. discoideum genome. A java code was written for converting the output of Run matcher in fasta format and giving number to the putative precursor miRNA (eg. miRNA:‘N’). From the probable miRNA precursor candidates generated by Run matcher, the putative miRNA precursors from the retrotransposons (DYSM and DYNL) cDNA sequence library were identified by blast search. For this blast search the miRNA precursor candidates and DYSM and DYNL cDNA library sequence were used alternatively as database and the query sequences.

After the blast search the hits identified were aligned using ClustalW software along with the dysm and dynl cDNA library sequences. From all the sequences, the common sequences and unique sequences were separated out using the result of the ClustalW alignment (Figure 2). A tree was constructed from the result of multiple sequence alignment and analyzed and then from this tree sequences were selected for secondary structure prediction. There are many online web tool for use for secondary structure prediction like mfold (energy minimization value for RNA 2nd structure prediction), genebee.msu.su (for energy calculation), Centroifold, etc. The ΔG values of the secondary structure prediction in Mfold of these sequences are listed in Table 1 below.
Figure 1: An overview of different steps involved in microRNA prediction from genome

Figure 2: A tree generated from multiple sequence alignment of final putative precursor microRNA

Figure 1: An overview of different steps involved in microRNA prediction from genome
Table 1: ΔG’ values for secondary structure in Mfold

<table>
<thead>
<tr>
<th>Sequence Number</th>
<th>ΔG’ values</th>
<th>Length of sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Common sequences</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence 8: DYSM_unique_1697</td>
<td>-32.54</td>
<td>23 aa</td>
</tr>
<tr>
<td>Sequence 11: DYSM_unique_2137</td>
<td>-13.56</td>
<td>26 aa</td>
</tr>
<tr>
<td>Sequence 15: DYNL_unique_653</td>
<td>-51.5</td>
<td>20 aa</td>
</tr>
<tr>
<td>Sequence 18: DYNL_unique_1135</td>
<td>-45.78</td>
<td>21 aa</td>
</tr>
<tr>
<td>Sequence 23: DYNL_unique_2124</td>
<td>-34.67</td>
<td>22 aa</td>
</tr>
<tr>
<td>Sequence 31: miRNA-94</td>
<td>-20.3</td>
<td>21 aa</td>
</tr>
<tr>
<td>Sequence 38: miRNA-168</td>
<td>-14.4</td>
<td>21 aa</td>
</tr>
<tr>
<td><strong>Unique sequences</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence 1: DYSM_unique_528</td>
<td>-46.90</td>
<td>19 aa</td>
</tr>
<tr>
<td>Sequence 2: DYSM_unique_761</td>
<td>-27.67</td>
<td>20 aa</td>
</tr>
<tr>
<td>Sequence 3: DYSM_unique_878</td>
<td>-35.54</td>
<td>20 aa</td>
</tr>
<tr>
<td>Sequence 4: DYSM_unique_1048</td>
<td>-52.50</td>
<td>21 aa</td>
</tr>
<tr>
<td>Sequence 5: DYSM_unique_1241</td>
<td>-30.78</td>
<td>21 aa</td>
</tr>
<tr>
<td>Sequence 6: DYSM_unique_1431</td>
<td>-55.1</td>
<td>21 aa</td>
</tr>
<tr>
<td>Sequence 7: DYSM_unique_1537</td>
<td>-31.45</td>
<td>22 aa</td>
</tr>
<tr>
<td>Sequence 9: DYSM_unique_1913</td>
<td>-21.70</td>
<td>24 aa</td>
</tr>
<tr>
<td>Sequence 10: DYSM_unique_2096</td>
<td>-64.3</td>
<td>25 aa</td>
</tr>
<tr>
<td>Sequence 12: DYSM_unique_2192</td>
<td>-44.89</td>
<td>26 aa</td>
</tr>
<tr>
<td>Sequence 13: DYSM_unique_2345</td>
<td>-22.54</td>
<td>27 aa</td>
</tr>
<tr>
<td>Sequence 14: DYNL_unique_377</td>
<td>-35.34</td>
<td>19 aa</td>
</tr>
<tr>
<td>Sequence 16: DYNL_unique_919</td>
<td>-32.70</td>
<td>21 aa</td>
</tr>
<tr>
<td>Sequence 17: DYNL_unique_1055</td>
<td>-29.50</td>
<td>21 aa</td>
</tr>
</tbody>
</table>
**Sequence 19: DYNL_unique_1191**  
-26.90  
21 aa

**Sequence 20: DYNL_unique_1198**  
-37.7  
21 aa

**Sequence 21: DYNL_unique_1393**  
-24.54  
21 aa

**Sequence 22: DYNL_unique_2119**  
-18.20  
22 aa

**Sequence 24: DYNL_unique_2370**  
-23.60  
23 aa

**Sequence 25: DYNL_unique_2890**  
-28.18  
25 aa

**Sequence 26: DYNL_unique_2911**  
-40.10  
26 aa

**Sequence 27: miRNA-101**  
-18.1  
17 aa

**Sequence 28: miRNA-73**  
-14.6  
20 aa

**Sequence 29: miRNA-100**  
-31.2  
21 aa

**Sequence 30: miRNA-91**  
-20.3  
21 aa

**Sequence 32: miRNA-114**  
-39.1  
21 aa

**Sequence 33: miRNA-120**  
-19.1  
21 aa

**Sequence 34: miRNA-121**  
-38.2  
21 aa

**Sequence 35: miRNA-126**  
-22.6  
21 aa

**Sequence 36: miRNA-127**  
-51.3  
21 aa

**Sequence 37: miRNA-153**  
-51.3  
21 aa

**Sequence 39: miRNA-160**  
-21.9  
21 aa

**Sequence 40: miRNA-154**  
-19.7  
21 aa

**Sequence 41: miRNA-23**  
-22.1  
21 aa

**Sequence 42: miRNA-24**  
-22.1  
21 aa

**Sequence 43: miRNA-25**  
-22.1  
21 aa

**Sequence 44: miRNA-58**  
-17.9  
21 aa

---

**B) Ab initio approach**

In this approach the IGIB EumiR (eukaryotic microRNA precursor prediction server) webserver was used for prediction of microRNA precursor from the *D. discoideum* genome sequence. It is based on neural network and support vector machine for eukaryotic microRNA prediction from the raw sequences. Java programming was used for taking each chromosome sequence as input in EumiR webserver and after processing the output html
file was saved in a separate folder for each chromosome. Java programming was also to extract the positive microRNA precursor and save in other file, giving name as Candidates precursor miRNA ‘N’ and write it in Fasta format. The sequences giving positive result in EulMiR were selected and analysed for their secondary structure and thermodynamics stability for secondary structure using RNAfold web server for secondary structure prediction tool (Figure 4).

Figure 3: Graphical analysis of EumiR result for positive sequence

![MicroRNA precursor sequence Positive sequence graph](image)

Subsequently the sequences showing appropriate secondary structure and minimum free energy were selected and classified as real or pseudo precursor sequences using Mipred web server. It is based on the statistical calculation of reference dataset, which predicts whether the candidate precursor microRNA is real or pseudo.

Figure 4: RNAFOLD WEB SERVER for RNA secondary structure prediction

![RNAfold WebServer](image)
Subsequently the sequences showing appropriate secondary structure and minimum free energy were selected and classified as real or pseudo precursor sequences using Mipred web server. It is based on the statistical calculation of reference dataset, which predicts whether the candidate precursor microRNA is real or pseudo.

Figure 5: Snapshot of MiPred Result: Classification of Real and Pseudo microRNA Precursors using Random Forest Prediction Model with Combined Features

<table>
<thead>
<tr>
<th>Prediction Result:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence Name:</strong></td>
</tr>
<tr>
<td><strong>Sequence Content:</strong></td>
</tr>
<tr>
<td><strong>Length:</strong></td>
</tr>
<tr>
<td><strong>Pre-miRNA-like Hairpin?</strong></td>
</tr>
</tbody>
</table>
| **The Secondary Structure:** | .....

```
((((((((((((((((((((((((C.(((((((((C.((((((((((((((((((((((...)))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))
```
| **MFE:** | $-64.60$ |
| **p-value (shuffle times:1000)** | 0.001 |
| **Prediction result:** | Real microRNA precursor |
| **Prediction confidence:** | 87.9% |
The real miRNA precursor sequences predicted by MiPred server were then matched by standalone blast with small RNA retrotransposon sequences (dynl & dysm).

**Conclusion:**

The objective of the present work was to perform computational analysis of *Dictyostelium discoideum* genome for identifying probable microRNA and its precursor sequences. MicroRNA analysis was done using sequence based and structure based prediction algorithms by using various web based tools. In comparative genomics approach, we get 171 putative candidate microRNA and its precursor using Mirbase as reference database and query as small subsequence from genome. In *ab initio* approach we get 79 putative candidate microRNA and its 116 precursor using neural network method. Here genomic subsequence were analysed in EumiR web server and positive precursor microRNA sequence were identified based on neural network and support vector machine approach. Then, further positive sequence analyze in RNAFOLD web server for the thermodynamic stability of RNA by secondary structure prediction. Further precursor micro RNA sequence were classify in Mipred web server using statistical approach by Random forest method into real or pseudo precursor microRNA sequence.

**Future Prospects**

*Dictyostelium* often referred to as “slime mold” or “social amoeba”, is one of the simplest studied eukaryote that possesses true multicellularity. The organism provides unique advantages for studying fundamental cellular processes with powerful molecular genetic, biochemical, and cell biological tools. Small RNAs play crucial roles in regulation of gene expression in many eukaryotes. Identification of the small size of miRNAs has been technically challenging, however a number of antisense-based approaches are available that transiently block miRNA function. In the model organism *D. discoideum*, RNAi has for several years been used as a tool. Thus, identification of novel microRNAs will be quite valuable in studying the developmental regulation of *D. discoideum* life cycle. In the present study, comparative genomics and *ab initio* approaches were used for prediction of microRNA from *D. discoideum* genomic sequence. In future the sequences of the putative microRNAs and their flanking sequences can be analyze for identifying the regulatory regions that are involved in controlling the expression of these microRNAs. Further the putative MiRNA precursors identified in the study need to be validated using wet lab techniques.

**Acknowledgements**

We thanks and acknowledge dictybase team for providing the genome sequence data of *D. discoideum*. Fredrik Söderbom et.al for providing the small RNA repertoire sequence which help us in predicting candidates miRNA and RNA@IGIB team for providing access to Eumir tool for prediction of microRNA Precursor sequence.

**References:**

