A Network Tool for Mining MicroRNA Target Genes in the Chromosome

Xingfeng Lv

College of computer science and technology, Heilongjiang University, Harbin, Heilongjiang, P.R.China, 150080,E-mail: xingfeng99@126.com

Abstract. The research of the chromosomal positions enrichment of genes is a method widely used to characterize a set of genes. In this paper we have developed MTGCFinder, an online tool for mining microRNA-regulated target genes clusters on chromosome. The sliding window algorithm is provided to mine the target genes clusters on chromosome segments. At the same time, it also provides a search for the enrichment of all microRNA-regulated target genes on chromosomes/regions in number of species. In order to further study the relationship between microRNA and human cancer, MTGCFinder also provides the function for mining cancer-related microRNA target genes clusters on chromosome regions, which is helpful for the study of human cancer.

Keywords: microRNA, enrichment on chromosomes, hypergeometric distribution test, sliding window

1 Introduction

MicroRNA (miRNA) is a kind of non-encoding small molecule RNA, whose length is only about 22 bases. It is widely found in animals, plants, single-celled organisms and has very important biological function. MicroRNA plays an important role in some physiological and pathological processes, such as stem cell differentiation, immune response, cancer cell development and metastasis [1]-[3]. In recent years, as an important indicator of miRNA function, the research of target gene (TGs) is becoming more and more in-depth, such as target point prediction, GO node enrichment and KEGG pathway enrichment and genome distribution [4]-[9].

Although studies have confirmed that the genes in the eukaryotic chromosome are not evenly distributed [10]-[13], but so far, there is still no suitable tool for detecting enriched microRNA-regulated target genes sets on chromosomes. At present, there are several software tools that can characterize the gene cluster on chromosome, such as GSEA[14], CROC[15], ChromoScan[16]. Although the software can analyze the degree of gene enrichment on the chromosome, they are not the analysis software for the target gene enrichment of the miRNA. So we developed MTGCFinder, a tool based web for the identification of the enrichment of microRNA-regulated target genes on chromosomes.

* This work is supported by Heilongjiang University Youth Fund.
The main structure of this paper is as follows: Methods and materials are given in Section 2. Functions and features are obtained in Section 3. In Section 4 overview of MTGCFinder is presented. In Section 5 the conclusions of this paper are given.

2 Methods and Materials

MTGCFinder provides miRNA-TGs information of 10 species: human, nematode, zebrafish, chicken, cattle, dog, rat, mouse, monkey, chimpanzee. Among them, the human miRNA-TGs information is from the miRBase, miRGen, and miRNA map database. The miRNA-TGs information of other 9 species (nematode, zebrafish, chicken, cattle, dogs, rats, mice, monkeys, chimpanzees) are from the miRBase database (Table 1: the third column indicates that the data in the database is predicted by different miRNA target. (1) indicates the miRBase database; (2) indicates miRGen database; (3) indicates the miRNA map database). Extraction of the locating information for genes on chromosomes is conducted from the geneCards database using online extraction procedures. Disease related miRNA-TGs information is from the literature mining.

Table 1. The number of miRNA and chromosome for each species

<table>
<thead>
<tr>
<th>Species name</th>
<th>Number of chromosomes</th>
<th>Number of microRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bos taurus</td>
<td>30</td>
<td>125 (1)</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>6</td>
<td>136 (2)</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>39</td>
<td>5 (3)</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>25</td>
<td>219 (4)</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>28</td>
<td>131 (4)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>24</td>
<td>711 (1), 419 (2), 313 (3)</td>
</tr>
<tr>
<td>Macaca mulatta</td>
<td>21</td>
<td>70 (1)</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>21</td>
<td>348 (1)</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>25</td>
<td>80 (5)</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>21</td>
<td>348 (1)</td>
</tr>
</tbody>
</table>

In order to mine the clustering of miRNAs-TGs on chromosome, we use two algorithms, the hypergeometric distribution test and the sliding window algorithm. We developed 5 kinds of mining tools. "by gene set search" and "window sliding" tools are based on sliding window technology; the other three tools ("by miRNA by", "by chromosome disease"), the use of the hypergeometric distribution test to test the miRNA-TGs collection enrichment.
2.1 Hypergeometric Distribution Test

MTGCFinder uses the hypergeometric distribution test to identify the MiRNA-TGs cluster on the chromosome. The formula is shown below.

\[
p = 1 - \sum_{i=0}^{x-1} \left( \begin{array}{c} K \\ i \end{array} \right) \left( \frac{M}{N} \right)^i \left( 1 - \frac{M}{N} \right)^{k-i}
\]

The meaning of each variable in the formula is as follows:
- \(N\) means the number of genes that are regulated by all miRNAs, and have annotation information of the chromosome position;
- \(K\) means the number of genes that are regulated by one or other miRNAs, and have annotation information of the chromosome position;
- \(M\) means the number of all genes that have annotation information of the chromosome position in a chromosome or a region;
- \(X\) means the number of all genes that are regulated by one or other miRNAs, and have annotation information of the chromosome position in a chromosome or a region.

2.2 Sliding Window Algorithm

Sliding window algorithm is a common method to study the characteristics of molecular sequences. For example, in the study of Proutski et al., the sliding window algorithm is used to analyze the change of nucleic acid sequence [17]. In the sliding window algorithm, the data is divided by a certain moving average standard. For instance, take the number of bases as the window width, the window slides on the sequence or sequence fragment [10].

In the online sliding window algorithm of MTGCFinder which is written by the Java program and based on hypergeometric distribution test, the gene is taken as the input, the window width \(W\) user defined is sliding on the whole chromosome (\(W\): 3-20 gene size). Each time slide for the displacement of a gene and use the super geometric distribution algorithm to calculate the p value of the window. Repeat this step until the end of the chromosome. Finally, in order to find the maximum target gene set, merge all of the interfacing windows with a notable enrichment (\(P<0.05\)), and to recalculate the p value of the chromosome region after the merger.

3 Functions and features

MTGCFinder is a tool based on the hypergeometric distribution test and sliding window algorithm for mining the clusters of microRNA target genes on chromosomes. Compared with other online analysis tools, MTGCFinder has the following advantages.

MTGCFinder can return the excavated target gene cluster in a graphical way by using color markers in the hot region of the chromosome. This intuitive way
to view results can not only benefit the researcher with non-biological background, but also can be used to predict the same chromosome hot region predicted by different target prediction database which is good for users to find a more reliable TGs clusters.

MTGCFinder provides 5 kinds of query functions, including "Sliding window", "by miRNA name Search", "by chromosome location Search", "by gene set Search", and "by disease name Search". Among them, the search method of the sliding window algorithm is a kind of flexible and efficient mining method of miRNA-TGs clusters, and also the core algorithm of MTGCFinder.

MTGCFinder can only identify “Ensemble Transcript ID”, but the user can conduct the mutual conversion for 10 kinds of common gene ID by the ID trans tool provided by MTGCFinder, including Ensemble Transcript ID, Ensemble Gene ID, Ensemble Protein ID, EMBL(GenBank ID), Unigene, Entrez gene ID, Associated Gene name, Uniprot/TrEMBL Accession, Uniprot/Swissprot ID, Uniprot/Swissprot Accession.

4 Overview of MTGCFinder

We use Java program for data processing to the collected information of miRNA-TGs, and the processed data is used to build a local database or placed on the server side by using JSP technology, MySQL database, Apache-Tomcat server and other tools, and finally complete the construction of the network tools. This tool is open source software, and provides the download of all the source code.

The user interface of this network platform is friendly, and supports all major browsers. The tested successful browsers include: IE7.0, IE ie8.0, Firefox, Safari, opera, Google. It does not support the IE6.0 browser with low version. Figure 1 is overview of MTGCFinder.

![Figure 1. Overview of MTGCFinder](image-url)
According to the different query, the results are not the same, but roughly similar. The results mainly include the visualization results of excavated chromosome on which target gene clusters, the chromosome location, the number of genes and the gene number in the cluster, and the N, K, M, and X values in the calculation process of the excavated target genes. The results of the MTGCFinder can return to a good visual result page. The results of the target gene cluster are marked in red on the chromosome. Figure 2 is the result of 4 search methods.

Fig. 2. The result of 4 search methods

A, B, C, D in Figure 1 is the results page based on different user input. A: the TGS cluster excavated by sliding window using hypergeometric distribution test according to genetic set the user defined. B: all miRNA-TGS clusters on the chromosome (or region) which the user interested. C: all TGS clusters regulated by miRNA which the user interested. D: The disease-related TGS clusters of miRNA and the disease which the user inputs.
Conclusions

MTGCFinder is used to excavate the target genes clusters of microRNA on chromosome. Through the MTGCFinder network platform, users can easily excavate the clustering information of miRNA target gene which the user interested. MTGCFinder can excavate the miRNA-TGs clustering from 4 angles such as miRNA, chromosome, target gene, and the disease. The user interface of MTGCFinder is friendly. At present, MTGCFinder only provides the query of the prediction database for three targets of human, while for other species it only provides the prediction query of miRBase database, and uses the sliding window algorithm only in the human species. In the future work, we want to add prediction database for more species and target points, and use sliding window algorithm to all species. This will make the MTGCFinder function more perfect, and can accurately find the chromosomal location of the target gene cluster the user's interested, which is conducive to biological research.

References