International Journal of u- and e- Service, Science and Technology
Vol.9, No. 4 (2016), pp.195-202
http://dx.doi.org/10.14257/ijunesst.2016.9.4.20

MTGCFinder:A Web Tools for Mining MicroRNA Target Genes Clusers on Chromosome

Xingfeng Lv

College of computer science and technology, Heilongjiang University, Harbin, Heilongjiang, P.R. China, 150080.

xingfeng99@126.com

Abstract

The research of the chromosomal positions enrichment of genes is a method widely used to characterize a set of genes. Several software tools are now available for this analysis, such as GSEA, CROC, ChromoScan. So far, however, there is still no suitable tool for detecting enriched microRNA-regulated target genes sets on chromosomes. Here, for the first time, 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, and according to the needs of the specific issues, users can set the window size themselves. 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. Background

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.

The main structure of this paper is as follows: Methods and meterials are given in Section 2. Database design and platform implementation are obtained in Section 3. In Section 4 the function of MTGCFinder is presented. In Section 5 the conclusions of this paper are given. As shown in Figure 1, a detailed flow was as follows: (1) miRNA and target genes and the information of genes' chromosome location are

ISSN: 2005-4246 IJUNESST

inputted.(2) Hypergeometric tests were used to determine whether the miRNA target genes set were significantly enriched in the chromosome regions and the corresponding chromosomal region was considered as target genes enrichment chromosomal region, if the p value was <0.01.(3)The enrichment chromosome regions are showed in the web browser.

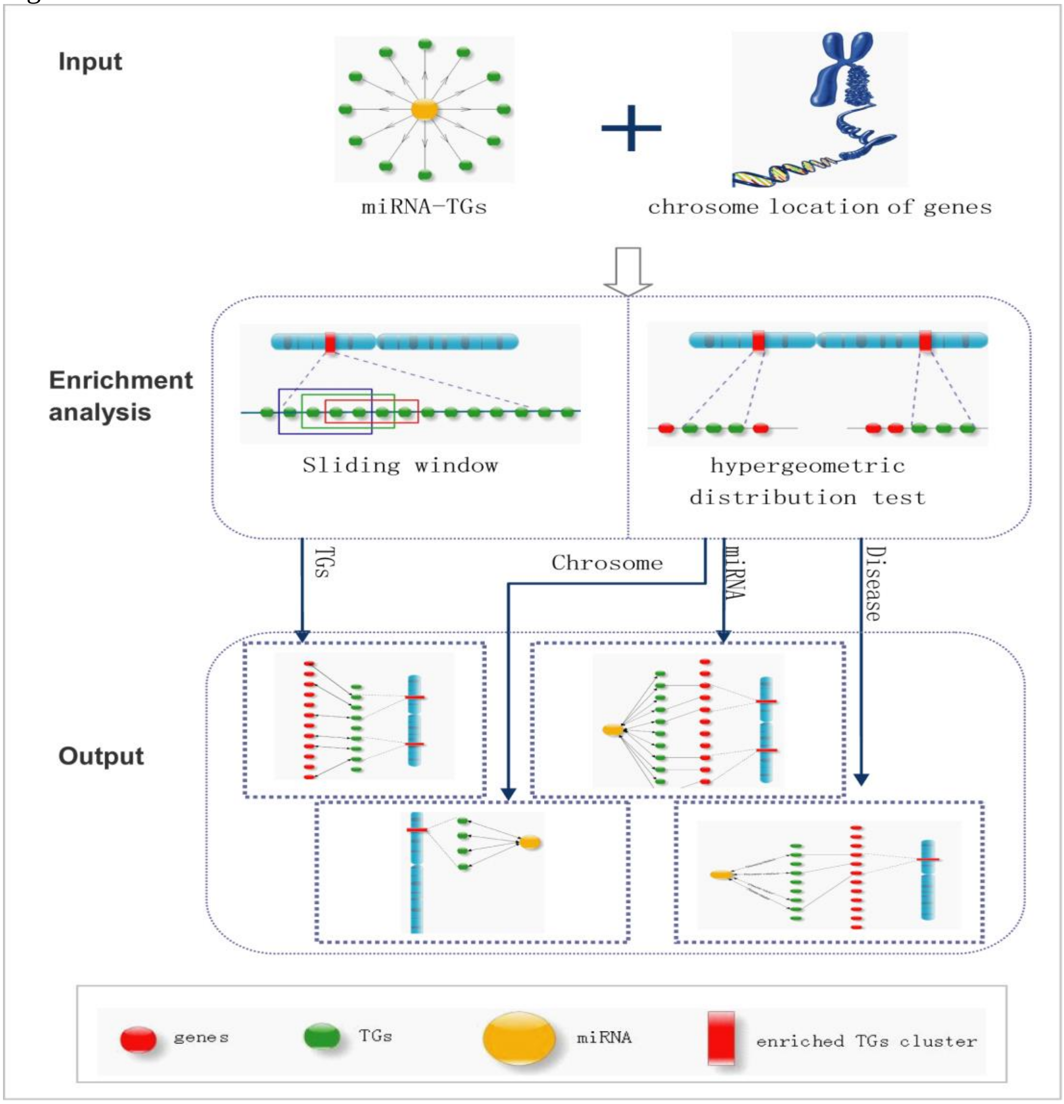


Figure 1. Algorithm Flow of Methods used to Identify the Chromosomal Regions which miRNA Target Genes Enriched

2. Material and Methods

2.1. Material

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 miRNAmap database. The miRNA and its target gene data are shown in Table 1.

Table 1. Data of Human miRNA

database	Number of miRNA	Number of miRNA-TGs
miRBase	711	433,617
miRNAMap	470	500,621
miRGen	494	502,810

The miRNA-TGs information of other 9 species (nematode, zebrafish, chicken, cattle, dogs, rats, mice, monkeys, chimpanzees) are from the miRBase database.It is shown in Table 2. 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 2. The Number of miRNA and Chromosome for Each Species

Species name	Number of chromosomes	Number of microRNAs
Bos taurus	30	125
Caenorhabditis elegans	6	136
Canis_familiaris	39	5
Danio rerio	25	219
Gallus gallus	28	131
Homo sapiens	24	711,419,313
Macaca mulatta	21	70
Mus musculus	21	348
Pan troglodytes	25	80
Rattus norvegicus	21	348

2.2. Methods

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.

(1) Hypergeometric Distribution Test

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

$$p = 1 - \sum_{i=0}^{x-1} \left(\frac{K}{i} \right) \left(\frac{M}{N} \right)^{i} \left(1 - \frac{M}{N} \right)^{k-i}$$

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) 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.

Sliding window is show in Figure 2. The red region represents a chromosome segment,the blue line represents the amplified chromosome segment,and the blue line of the green circle represents the gene on the chromosome. The blue, green, red box represents a sliding window on the chromosome.

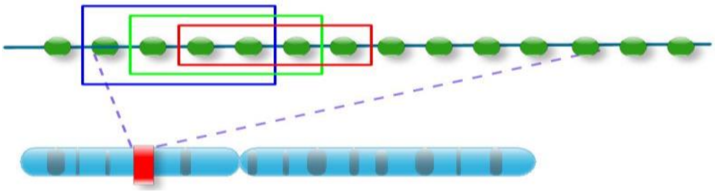


Figure 2. Sliding Window

First,the value of p which in window width is computer by using the hypergeometric distribution. In the second step, window backward movement and calculate the p value ,and 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. Database Design and Platform Implementation

3.1. Requirement Analysis

The software of MTGCFinder is applied to a biological study of no computer background, and MTGCFinder provides a user-friendly interface. The researchers do not need to be familiar with the structure of the database and the program language. They can use this tool to carry out a flexible study. At the same time, we provide all of the program and the source code download, so that the MTGCFinder is also suitable for the familiar computer language of bioinformatics research.

MTGCFinder is able to calculate the enrichment of all target genes on chromosome and chromosome segments of microRNA regulation, because the current target prediction database has some false positive rate and false negative rate. The target gene information of the human species is provided by MTGCFinder, and the results of the database can be predicted by different target points.

3.2. Data Processing

The underlying data of MTGCFinder is divided into two parts, a part is stored in the MySQL database, which is used for on-line query, and the other part is placed on the server as a text file, which is used for on-line calculation.

Download the human miRNA-TGs data from the miRGen miRNAmap miRBase database, download the miRNA-TGs data from the other 9 species from the miRBase database. Using the Java program to process the miRNA-TGs data, using the hypergeometric distribution test to detect whether each miRNA has a target gene cluster on each chromosome.

After the treatment, there are 8 columns as follows: the name of database, the position of miRNA, the number of miRNA, the name of miRNA, chromosome number, gene set, P value, FDR value. Other species data due to the only data source, only the miRBase database to download the data, does not include: database name, the same as the other. In addition, we have also calculated the target gene cluster of each miRNA in each chromosome region. The disease related miRNA-TGs data sources from the literature mining, the same use Java program to deal with the above format (because the data source is the literature mining, the result does not include the database name). These results are stored in MySQL database As the basic data of MTGCFinder.

Another part of the data for online computing is currently only human. The human species data were derived from 3 different target databases, and the online calculation program for each target point forecast database needs 3 files.

3.3. Database Design

MTGCFinder includes 3 databases: data database, IDconvert database, otherSpecies database. Data database stores human data, include 13 data tables. IDconvert database stores the data of individual species, include 73 data tables. The otherSpecies database stores the information of 9 species of nematode, zebrafish, chicken, cattle, dog, rat, mouse, monkey and chimpanzee, and has 9 data tables.

3.4. Platform Implementation

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 3 is overview of MTGCFinder.

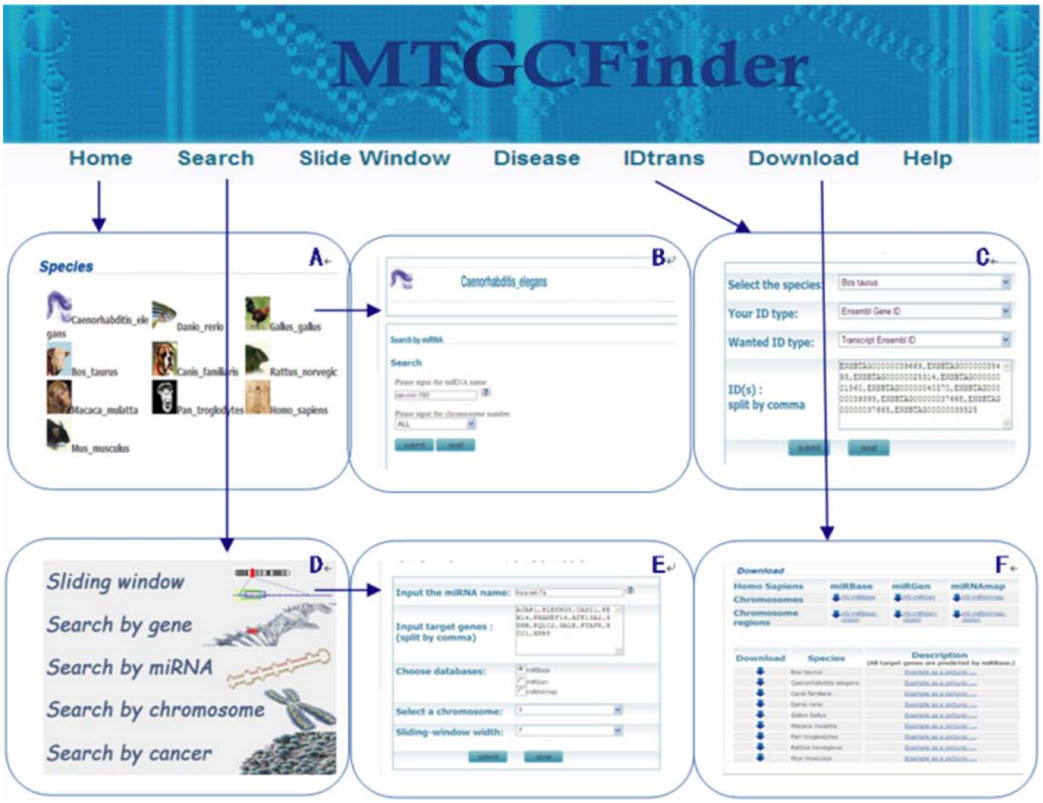


Figure 3. Overview of MTGCFinder

Chart A page can be seen when the user enters into the home page. Click on the species icon to enter the species search tool (Chart B). Search tools include 5 types, as shown in Chart D, for example, click the "Sliding window" tool to enter the Chart E, also from the top of the page menu to directly enter into. IDtrans tool is shown in Chart C. Chart F is the download page, including the description download of miRNA-TGs clusters for human and other 9 species, as well as the information download of 8 kinds of cancer.

4. Functions of MTGCFinder

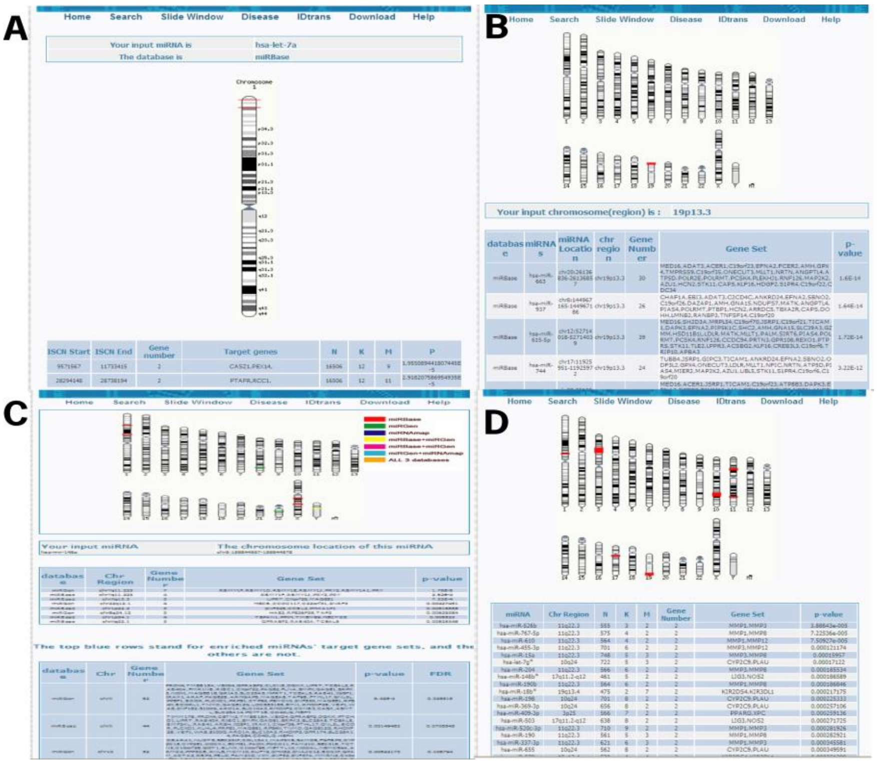
MTGCFinder is a tool based on the hypergeometric distribution test and sliding window algorithm for mining the clusters of microRNA target genes on chromosomes. Compared with other online analysis tools, MTGCFinder has the following advantages.

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 diease 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 Accsession, Uniprot/Swissprot ID, Uniprot/Swissprot Accession.

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 4 is the result of 4 search methods.



Your input miRNA is	hsa-let-7a
The database is	miRBase

ISCN Start	ISCN End	Gene number	Target genes	N	K	M	P
9571567	11733415	2	CASZ1,PEX14,	16506	12	9	1.955089441807445E-5
28294148	28738194	2	PTAFR,RCC1,	16506	12	11	2.918207586954935E-5

Your input chromosome(region) is : 19p13.3

database	miRNAs	miRNA Location	chr region	Gene Number	Gene Set	p-value
miRBase	hsa-miR-663	chr20:26136836-26136857	chr19p13.3	30	MED16,ADAT3,ACER1,C19orf23,EFNA2,FCER2,AMH,GPX4,TMPRSS9,C19orf35,ONECUT3,MLLT1,NRTN,ANGPTL4,ATP5D,POLR2E,POLRMT,PCSK4,PLEKHJ1,RNF126,MAP2K2,AZU1,HCN2,STK11,CAPS,KLF16,HDGF2,S1PR4,C19orf22,CDC34	1.6E-14
miRBase	hsa-miR-937	chr8:144967165-144967186	chr19p13.3	26	CHAF1A,EBI3,ADAT3,C2CD4C,ANKRD24,EFNA2,SBNO2,C19orf26,DAZAP1,AMH,GNA15,NDUFS7,MATK,ANGPTL4,PIAS4,POLRMT,PTBP1,HCN2,ARRDC5,TBXA2R,CAPS,DOHH,LMNB2,RANBP3,TNFSF14,C19orf20	1.64E-14
miRBase	hsa-miR-615-5p	chr12:52714018-52714039	chr19p13.3	39	MED16,SH2D3A,MRPL54,C19orf70,JSRP1,C19orf21,TICAM1,DAPK3,EFNA2,PIP5K1C,SHC2,AMH,GNA15,SLC39A3,GZMM,HSD11B1L,LDLR,MATK,MLLT1,PALM,SIRT6,PIAS4,POLRMT,PCSK4,RNF126,CCDC94,PRTN3,GPR108,REXO1,PTPRS,STK11,TLE2,LPPR3,ACSBG2,KLF16,CREB3L3,C19orf6,TRIP10,APBA3	1.72E-14
miRBase	hsa-miR-744	chr17:11925951-11925972	chr19p13.3	24	TUBB4,JSRP1,GIPC3,TICAM1,ANKRD24,EFNA2,SBNO2,ODF3L2,GPX4,ONECUT3,LDLR,MLLT1,NFIC,NRTN,ATP5D,PIAS4,MIER2,MAP2K2,AZU1,UBL5,STK11,S1PR4,C19orf6,C19orf20	3.22E-12

Your input miRNA	The chromosome location of this miRNA
[illegible]	[illegible]

database	Chr Region	Gene Number	Gene Set	p-value
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

The top blue rows stand for enriched miRNAs' target gene sets, and the others are not.

database	Chr	Gene Number	Gene Set	p-value	FDR
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

miRNA	Chr Region	N	K	M	Gene Number	Gene Set	p-value
hsa-miR-526b	11q22.3	555	3	2	2	MMP1,MMP3	3.88643e-005
hsa-miR-767-5p	11q22.3	575	4	2	2	MMP1,MMP8	7.22536e-005
hsa-miR-610	11q22.3	564	4	2	2	MMP1,MMP12	7.50927e-005
hsa-miR-455-3p	11q22.3	701	6	2	2	MMP3,MMP12	0.000121174
hsa-miR-15a	11q22.3	748	5	3	2	MMP3,MMP8	0.00015957
hsa-let-7g*	10q24	722	5	3	2	CYP2C9,PLAU	0.00017122
hsa-miR-204	11q22.3	566	6	2	2	MMP3,MMP8	0.000185534
hsa-miR-148b*	17q11.2-q12	481	5	2	2	LIG3,NOS2	0.000186589
hsa-miR-190b	11q22.3	564	6	2	2	MMP1,MMP8	0.000186846
hsa-miR-18b*	19q13.4	475	2	7	2	KIR2DS4,KIR3DL1	0.000217175
hsa-miR-198	10q24	701	8	2	2	CYP2C9,PLAU	0.000225333
hsa-miR-369-3p	10q24	656	8	2	2	CYP2C9,PLAU	0.000257106
hsa-miR-409-3p	3p25	566	7	2	2	PPARG,XPC	0.000259136
hsa-miR-503	17q11.2-q12	638	8	2	2	LIG3,NOS2	0.000271725
hsa-miR-520c-3p	11q22.3	710	9	2	2	MMP1,MMP3	0.000281926
hsa-miR-190	11q22.3	561	5	3	2	MMP1,MMP8	0.000282921
hsa-miR-337-3p	11q22.3	621	6	3	2	MMP1,MMP3	0.000345581
hsa-miR-655	10q24	562	8	2	2	CYP2C9,PLAU	0.000349591

Figure 4. The Result of 4 Search Methods

A, B, C, D in Figure 4 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.

5. 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.

Acknowledgments

This work was supported by the Laboratory of Software Engineering, College of Heilongjiang province, (Heilongjiang University), P. R. China, by the Heilongjiang University youth fund.

References

[1] M. Lionetti, P Musto and D. Martino "MT: Biological and clinical relevance of miRNA expression signatures in primary plasma cell leukemia", Clin Cancer Res, no. 19, (2013), pp. 3130-42.

[2] C. Lei, C. Xiao-ping and L. Yuan-jian, "The study of microRNA and its single nucleotide polymorphisms", International Journal of Pathology and Clinical Medicine, vol. 29, no. 3, (2009), pp. 211-5.

[3] G.L Papadopoulos, P. Alexiou, M. Maragkakis , M. Reczko and A.G. Hatzigeorgiou, "DIANA-mirPath: Integrating human and mouse microRNAs in pathways", Bioinformatics, vol. 25, no. 15, (2009), pp. 1991-3.

[4] S. Gr.-Jones, "miRBase: the microRNA sequence database", Methods Mol Biol, vol. 342: (2006), pp. 129-38.

[5] M. Megraw , P. Sethupathy, B. Corda , and A.G. Hatzigeorgiou, "iRGen: a database for the study of animal microRNA genomic organization and function", Nucleic Acids Res, vol. 35 (Database issue), (2007), pp. D149-55.

[6] P.W. Hsu, H.D. Huang, S.D. Hsu, L.Z. Lin, A.P. Tsou, C.P Tseng, P.F. Stadler, S. Washietl, and I.L. Hofacker, "miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes", Nucleic Acids Res, vol. 34 (Database issue), (2006), pp. D135-9.

[7] R. Song-wei, S. Wei-hong, Y. Peng-cheng, Z. Yi, and S. Qi-xiang, "The research survey and development trend of microRNA target gene prediction algorithm", Life science, vol. 19, no. 5, (2007), pp. 562-7.

[8] X. Wei, C. Guo-jun, and S. Ning-sheng, "The research progress in search and identification method of microRNA target genes", Science in China Press, vol. 39, no. 1, (2008), pp. 121-8.

[9] M. Riaz, MT V. Jaarsveld and A Hollestelle, "miRNA expression profiling of 51 human breast cancer cell lines reveals subtype and driver mutation-specific miRNAs", Breast Cancer Res, vol. 15, R33, (2013).

[10] F.Tajima, "Determination of window size for analyzing DNA sequences" J Mol Evol, vol. 33, no. 5, (1991), pp. 470-3.

[11] P. Michalak, "Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes", Genomics,vol. 91, no. 3, (2008), pp. 243-8.

[12] A.Coppe, GA Danieli, and S Bortoluzzi, "REEF: searching REgionally Enriched Features in genomes", BMC Bioinformatics, vol. 7, (2006), pp. 453.

[13] L. Yong-qing, X. Jian-zhen, and M. Zhi-gang, "microRNAs agminated on chromosome have more common target genes", Acta Biophysica Sinica, vol. 23, no. 6, (2007), pp. 470-4.

[14] M Thomassen, Q Tan, and TA Kruse, "Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis", Breast Cancer Res Treat, vol. 113, no. 2, (2009), pp. 239-49.

[15] M Pignatelli, F Serras, A Moya, R Guigó, and M Corominas, "CROC: finding chromosomal clusters in eukaryotic genomes", Bioinformatics, vol. 25, no. 12, (2009), pp. 1552-3.

[16] YV Sun, DM Jacobsen, and SL Kardia, "ChromoScan: a scan statistic application for identifying chromosomal regions in genomic studies", Bioinformatics, vol. 22, no. 23, (2006), pp. 2945-7.

[17] V Proutski, and E Holmes, "SWAN: sliding window analysis of nucleotide sequence variability", Bioinformatics, vol. 14, no. 5, (1998) ,pp. 467-8.

Author

Xingfeng Lv is a lecturer at Heilongjiang University now. She obtained her bachelor's degree and master's degree in Heilongjiang University. Her major researches are computer network, computational biology.