BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 10(8)

miRAFinder and GeneAFinder scripts: large-scale searching for miRNA and related information in indexed literature abstracts

Olga Berillo [1]*, Mireille Régnier[2] & Anatoly Ivashchenko[1]

[1]National nanotechnology laboratory, al-Farabi Kazakh National University, Almaty, Kazakhstan; [2]INRIA/LIX team AMIB, École Polytechnique, Palaiseau, France; Berillo Olga – Email: devolia18@mail.ru; *Corresponding author

Received July 03, 2014; Revised July 14, 2014; Accepted July 16, 2014; Published August 30, 2014

Abstract:
In recent times, information on miRNAs and their binding sites is gaining momentum. Therefore, there is interest in the development of tools extracting miRNA related information from known literature. Hence, we describe GeneAFinder and miRAFinder scripts (open source) developed using python programming for the semi-automatic extraction and arrangement of updated information on miRNAs, genes and additional data from published article abstracts in PubMed. The scripts are suitable for custom modification as per requirement.

Availability: miRAFinder and GeneAFinder scripts are free and available for download at http: //sites.google.com /site/malaheenee/software.

Keywords: python script, gene information, miRNA information, semi-automatic extraction.

Background:
The number of abstracts for biological articles at the PubMed **[1]** database has increased over timeline due to steady advancement in biomedical research. There are a number of online servers that extract information in a specific manner from abstract archived databases. The MedlineRanker web server allows a flexible ranking of Medline for a given topic of interest **[2]**. MedEvi imposes positional restriction on occurrences matching multi-term queries, based on the observation that term with semantic relations **[3]**. The FNeTD method for clustering achieved PubMed abstracts using revealing frequent "phrases" or "words" and identifying "nearer terms" of the domain **[4]**. The MiSearch is an adaptive biomedical literature search tool that ranks citations based on a statistical model **[5]**. Genomics researches are having a major impact on biological and medical sciences although the function of many genes remains unknown. In recent times, PubMed database search shows about 285 thousand articles associated with breast cancer researches. While, one of causes of tumorigenesis is the suppression of gene expression via microRNAs (miRNAs) **[6].** There are more than 2500 human miRNAs and some of them are potential therapeutic targets for neoplastic diseases **[7].** Database search shows more than 32 thousand articles devoted to miRNAs. Short nucleotide sequences of miRNAs can be used as biomarkers for cancer diagnostics as they circulate in biological liquids. Therefore, it is important to reveal comparative information about each studied miRNA in the literature. It should be noted that more 13 thousand articles are associated with miRNA participation in tumourigenesis as per the PubMed database search. Thus, specific search for refined information from archived databases is often time consuming and tedious in nature. Hence, we

describe miRAFinder and GeneAFinder scripts written in python to simplify such tasks during biological investigations.

Software development and usage:
Python was used to write scripts for this purpose.

Software execution command:
The following commands were used for execution.
cd folder/python miRAFinder.py -g pathway_1 -f pathway_2 > file.txt Where pathway_1 is directory to gene dictionary file (genedic.txt) and pathway_2 is directory to file with abstracts (abstract.txt).

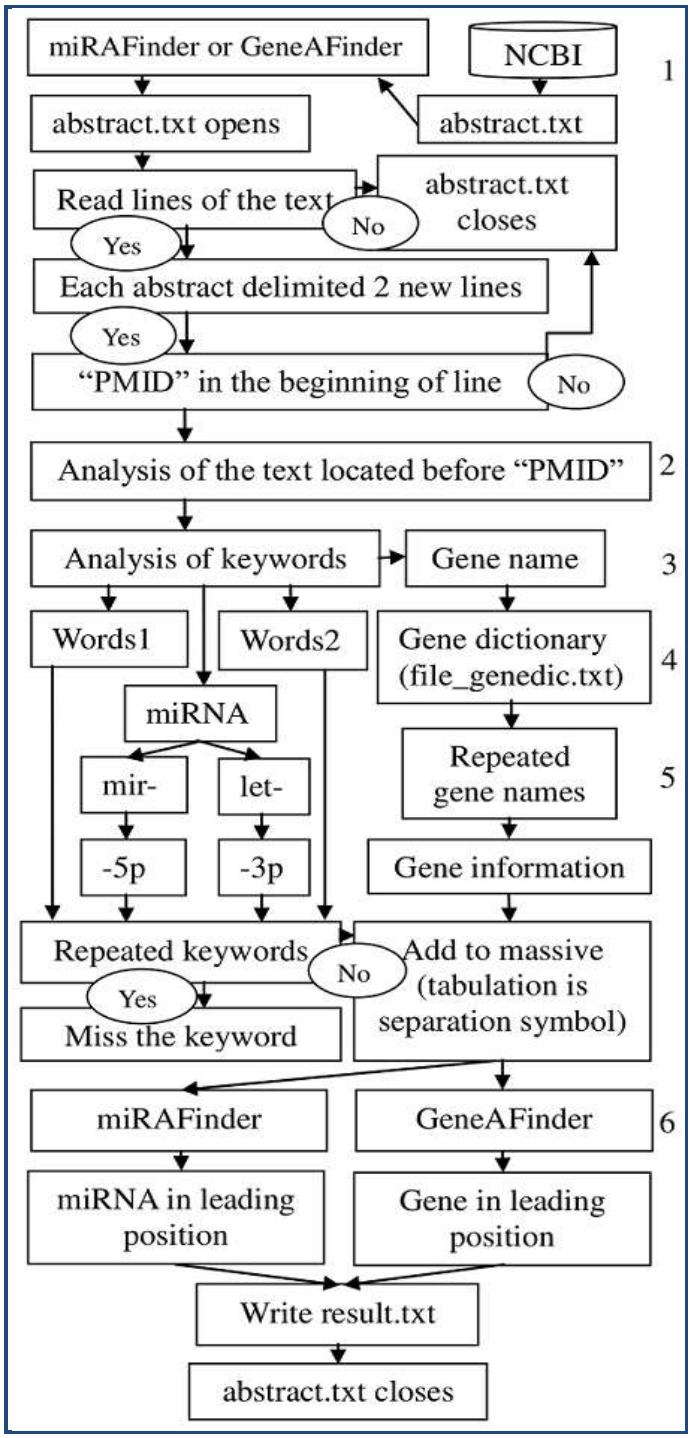

Figure 1: The scheme of miRAFinder and GeneAFinder scripts.

Note: 1 - Preparation of "abstract.txt" file throught the downloading the abstracts from PubMed's site in text format; 2 – treatment of each abstract text; 3 – search for keywords in different groups (gene, miRNA, specific word, etc.); 4 – use of dictionaries with gene data; 5 – identification different miRNA types (mir- or let-); 6 – definition of data presentation depend on miRAFinder or GeneAFinder script.

Input data:
The article abstracts at PubMed **[7]** were downloaded in text format and was saved as "abstract.txt" **(Figure 1).** The list of human genes was extracted from the HGNC **[8]** database and was saved as "genedic.txt" file. The lists of keywords of each word group are make up in depend on subject of research.

Output:
The miRAFinder script processes information of abstracts from the PubMed database. This script allows to find miRNA names and keywords as shown in **Figure 1.** The result of searching for contents of the following data: PMID (publication medicine identification number) of the article, miRNA name, disease localization (organ or tissue), keywords (methods, change fold, cellular processes, functions, animal species, types of cells, biological liquids, etc.) and genes. Some examples are shown in **Table 1 (see supplementary material**). The obtained data on gene names are important as they can be mentioned in abstract as host or target genes of miRNAs. The script uses the list of genes from HGNC database for the correct identification of gene names in the abstract (for an exception of different abbreviations). Thus, the quantity of the found genes is regulated by keyword structure of the dictionary created for each separate searching (human genes, mouse genes, rat genes, etc.). The GeneAFinder script is similar to miRAFinder **(Figure 1)** and it allows to find specific information and gene names in the abstract of PubMed. It is possible to find the list of important keywords for general characteristic in the text of abstract using the GeneAFinder script. Some examples are shown in **Table 2** (**see supplementary material**). The PubMed database was used in this study due to its comfortable data type format for application in the scripts. There are a possibility to use various gene dictionaries and their characteristics to retrieve a different data set.

Caveat & future development:
The need for the effective analysis of miRNA data using computer scripts is gaining momentum in recent times. The miRAFinder and GeneAFinder scripts described in this article help to extract specific information from archived abstracts in NCBI PubMed. We showed that the scripts scan through thousands of abstracts within reasonable time frames. The information gleaned through such approach finds utility in miRNA analysis of specific diseases (e.g. cancer).

References:
[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] Fontaine JF *et al. Nucleic Acids Research* 2009 **37**: W141 [PMID: 19429696]
[3] Kim JJ *et al. Bioinformatics* 2008 **24**: 1410 [PMID: 18400773]
[4] David MR & Samuel S, *Bioinformation* 2012 **8**: 20 [PMID: 22359430]
[5] States DJ *et al. Bioinformatics* 2009 **25**: 974 [PMID: 18326507]

[6] Palanichamy JK & Rao DS, *Front Genet.* 2014 **5**: 54 [PMID: 24672539]
[7] Costa PM & Pedroso de Lima MC, *Pharmaceuticals (Basel).* 2013 **6**: 1195 [PMID: 24275848]
[8] http://www.genenames.org/

Edited by P Kangueane
Citation: **Berillo *et al.*** Bioinformation 10(8): 539-543 (2014)

BIOINFORMATION

open access

Supplementary material:

Table 1: An example of received data through miRATarget script

PMID	miRNA	Keywords 1	Keywords 2	Keywords 3	Gene
23425975	miR-125b	bladder, thyroid, hepatocellular, ovarian, breast	cancer, carcinomas, squamous, carcinoma, metastasis	cell, migration, target, human, down, luciferase, cell line	MMP13
23060446	miR-218	bone, breast, mammary	cancer, stromal, tumor	signaling, differentiation, cell, target, pathway, down, up, normal, abnormal	SOST, DKK2, SFRP2
23384942	miR-7	brain, breast, bone	cancer, metastasis, disease, tumor	cell, high, low, level, stem cell, down, up, in vitro, target	KLF4, CD24, CD44
22310293	miR-9	brain, hepatocellular, breast, liver	oncogene, lymphoma, tumour, disease, cancer, carcinomas	in vivo, normal, cell, target, immune, signal, decrease, increase, level	DICER1
19509158	miR-21	oral, breast, colon	cancer, squamous, carcinomas, tumor	cell, apoptosis, experimental, microarray, quantitative, pcr, cell line, normal, level	TPM1, PTEN
22024162	miR-34	oral, colon, breast, lung	cancer, pathological, oncogenic, carcinoma	cell, decrease, level, high, cell cycle, apoptosis, pathway	LEF1
22542163	miR-30b	oral, mammary	squamous, cancer, tumour, metastasis	up, cell, hsa-, level, stage, quantitative, real-time, pcr, normal, percentage, increase	MIR30B
23166327	miR-155	oral, thyroid, breast, renal, gastric, liver	oncogenic, tumor, squamous, carcinoma, cancer, carcinomas	down, cell, up, human, cell line, level, decrease, apoptosis, increase, utr	CDC73
19074899	let-7i	ovarian, breast	cancer, disease	microarray, target, human, early, translation, high, real-time, pcr, increase, cell, decrease, stage, in situ	
20647319	let-7	ovarian, breast	cancer, disease, oncogene, tumor	level, increase, abnormal	KRAS, T, BRCA1, BRCA2
22322863	miR-182	ovarian, breast	tumour, carcinoma, metastasis, oncogenic, cancer, oncogene, pathological	high, early, normal, cell, cell line, increase, in vitro, in vivo	BRCA1, MTSS1, HMGA2

Table 2: An example of received data through GeneATarget script

PMID	Gene	miRNA	Keywords 1	Keywords 2	Keywords 3
20818337	HER-2	mir-106b, mir-500, mir-17	breast	triple, basal-like, luminal a, her2, luminal b	rat, high, up
23112837	HER2	mir-200, mir-200f, mir-200c	breast	basal, carcinomas, subtype, her2, triple, squamous	metastasis, human, low, cell line, rat, decrease, increase, down, up, cell
23748853	HER2	mir-17, mir-34a, mir-155, mir-373, mir-10b, mir-93	breast	her2, triple, carcinomas	serum, rat, cell, human, metastasis, increase
24059244	HER2	mir-17, mir-106b	breast	triple, her2, stromal, subtype, basal-like	rat, up, cell, high, increase
21435948	HER2	let-7	breast	triple, subtype, her2, luminal a, luminal	rat, oncogene
22388088	HER2	let-7	breast	her2, triple	cell, metastasis, human, oncogene, rat

22586447	HER2	mir-608	breast	her2, subtype, carcinoma, luminal, triple	human, rat, increase, low
22631664	HER2	mir-21, mir-205, mir-342	breast	her2, triple, fibroadenoma	metastasis, high, increase, down
22323552	HER2	mir-210	breast	triple, her2	high, cell, rat, hypoxia, low
22118463	HER2	let-7	colon, breast	her2, triple	cell, stem cell, oncogene, human, rat