Nile University
جامعة النيل

Bioinformatics – PD

Winter 2019

CIT – 656

Programming for Bioinformatics

Project (4)

RNA Secondary Structure of RBP Target Motifs

Prepared by

Asmaa Abdel Magid El Neanai

Maha Ibrahim Abdel Aleem

Manar Hashem

Nada Gamal Ibrahim Eldawy

Phoebe Magdy Abd El Massieh

Background

RNA binding proteins (RBPs) are crucial for a variety of cell processes within both the nucleus and the cytoplasm. The RBPs bind to particular regions on the target RNA, these regions consist of specific short sequences that is known binding motifs. Beside the primary nucleotide sequence of the RNA, the structure of the RNA target also has a vital role in the RBP-target recognition process. Although it has been settled that it is preferable for the majority of RBPs to bind their motif targets at the single stranded sites (***Li, et al., 2014***), few proteins e.g. Staufen are found to bind particularly to dsRNA as it has a double stranded RNA binding domain (dsRBD) (***Ramos et al., 2000***). While it has been generally accepted that the dsRBDs perceive their focuses in a non-sequence specific manner for their RNA motif targets, late examinations have demonstrated that they recognize both the RNA nucleotide sequence as well as their structure determinants (***Masliah et al., 2012***). In addition, it has been mentioned that "RBPs belonging to other domain families have been shown to bind in a sequence specific manner to preferred RNA secondary structures, such as the yeast protein Vts1, which was experimentally verified to bind to a sequence motif within a loop of a hairpin structure" (***Aviv et al., 2006***).

The binding techniques that is used to predict the protein-RNA binding are divided into in vitro and in vivo methods. The in vitro methods, as RNAcompete, are based on the High Throughput Systematic Evolution of Ligands by Exponential Enrichment (HT-SELEX) (***Ray et al., 2013***). The in vivo methods are based on the CrossLinking and ImmunoPrecipitation (CLIP) that originally was used to determine the binding target of the neural RBP Nova in the mouse transcriptome (***Licatalosi et al., 2008***). In order to increase the methods sensitivity and specificity, various variants have been developed and applied to RBPs in different cell types.

Moreover, from the information mentioned above which demonstrates that the structure of the RNA has an important role in the process of protein-RNA recognition, a large number of the algorithms such as MEMERIS, RNAcontext, RCk, the RBPmotif web server and RNAplfold, that publicly used for searching the RNA motifs consider the RNA primary nucleotide sequence and its structure information (***Li, et al., 2014***).

Depending on the fact that the RBPs prefer to bind to the single-stranded RNA (ssRNA) region of the RNA, the **MEMERIS** tool can predict the RPBs binding sites by using the Expectation Minimization motif discovery algorithm. **RNAcontext** and the **RNAcontext-k-mer** (RCK) algorithm methods is using the uncertainty in the secondary structure of the RNA sequence to predict the RBP sequence and the structure (***Kazan and Morris, 2013***).

To predict the binding sites of RBPs, the **RBPmotif web server** (http://www.rnamotif.org) perform either the de novo motif finding or analysis of structure preferences when there is no previous knowledge or there is a previous knowledge on the RBPs binding motifs, respectively (***Kazan and Morris, 2013***). **RNAplfold** is an RNA folding algorithms that is used to predict the probabilities of the RNA structure being paired or unpaired depending on the used options (***Lorenz et al, 2011***). Referring to the RNAplfold 2.4.13 manual page

(https://www.tbi.univie.ac.at/RNA/RNAplfold.1.html) there are several options for the RNAplfold, one of them is the -u option that is capable to calculate the mean probability that RNA regions are unpaired.

Furthermore, it would be useful to have a bioinformatics tool that can be easily used for predicting the single strandedness probability of the RNA target motifs, and also to be more user friendly. So that, it would be easier and save much time. As a result, the purpose of this program is to determine the probability of whether or not a particular motif occurrence is likely to be in single-stranded RNA.

Usage

Before starting to run the script, the user must download and install the VinnenaRNA Package 2 as script will execute one of its programmes (the RNAplfold programme). The ViennaRNA Package 2 can be downloaded from the homepage of ViennaRNA (https://www.tbi.univie.ac.at/RNA/). To be able to perform the installation step, the user has to follow the instructions presented in VinnenaRNA Package 2 documentation link: https://www.tbi.univie.ac.at/RNA/documentation.html

There are six input parameters the user has to submit to carry out the programme, which are:

1- the path of the RNAplfold programme, as the path vary from one user to another.
2- the name of the FASTA file that contains the sequences.
3- the name of the text file that contains a list of sequence motifs.
4- the number of the base pairs located before and after the motif.
5- the name of the output file.
6- the name of the output figure.

N.B., the extension of the figure has to be of supported formats: eps, pdf, pgf, png, ps, raw, rgba, svg, svgz.

There are two outputs, which are:

1- a text file that contains four columns for **the sequence ID**, **motif's start position**, **motif's end position** and **the motifs unpaired probabilities**, respectively.
2- a histogram figure for the motifs unpaired probabilities frequencies.

- **An example command to run the programme:**

```
python  RBP_Prob_prediction.py "/usr/local/bin/"  omega_singleexon.fasta CaceresHurst.txt 70 DataSheet.txt histogram.png
```

The four inputs here are: **"/usr/local/bin/"** (the path of the RNAplfold), **omega_singleexon.fasta** (the name of the fasta input file), **CaceresHurst.txt** (the name the text input file), **70** (the number of base pairs), respectively.
The last two arguments are: **DataSheet.txt** (the name of the output file) and **histograme.png** (name of the histogram figure).

Implementation

The script is divided into three main parts that are:

1- Import modules and packages part

This part contains the imported modules and packages each of them have specific functionalities, such **Bio.SeqIO, re, sys, os, glob, subprocess, numpy , and matplotlib. pyplot**.

The **Bio.SeqIO packag,** is imported as it provides a simple interface to input/output file format and deals with sequences as SeqRcord objects (https://biopython.org/wiki/SeqIO). To support the regular expressions that may presented in the motifs sequences the **re module is imported** (https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial). The **sys** and **os modules** are important as it provide functions and parameters that deal with filenames, paths, and directories (https://thomas-cokelaer.info/tutorials/python/module_os.html). Files of particular extension is needed to be removed before executing an external programme through this script, consequently the **glob module** is imported to find this particular extension (https://docs.python.org/3/library/glob.html). In order to execute the RNAplfold programme (an external programme) through the script the **subprocess module** is imported (https://docs.python.org/3/library/subprocess.html). Finally the **numpy** and **matplotlib.pyplot** modules are imported so as to convert the output list to array (https://docs.scipy.org/doc/numpy/user/basics.creation.htm) and plot a histogram figure (https://matplotlib.org/3.1.0/api/_as_gen/matplotlib.pyplot.hist.html) for the results, respectively.

2- Functions part

The Functions part contains three functions created that is available to be run in any script. The names of the functions are: **DeL_AllFiles**, **Execute_RNAplfold**, and **Histogram**.

As it is intended to remove files of a certain extension the **DeL_AllFiles** function was created. The **DeL_AllFiles** function has one string parameter that is extension of the file.

In order to execute the RNAplfold progamme through the script the **Execute_RNAplfold** function had been created. The **Execute_RNAplfold** function has two parameters the length of the motif (string), and the sequence which is the input of the RNAplfold programme, the return is the motifs unpaired probabilities.

The third function that was created is the **Histogram** function that is used to plot and save a histogram containing figure. The **Histogram** has two parameters, the list of the unpaired probabilities and a variable to specify the name of output figure.

3- The script body part

This part is the main part of the script in which the input files is parsed, the sys.argv[] is used to get the inputs arguments, the created functions are called, and three nested loops is created. For three nested loop, the first loop is looping over the list of the fasta file, the second loop is Looping over the motifs list file, and the third loop is Looping

over the sequences file for finding & determining the motifs and their start & end locations by using regular expression (re.finditer()) and "match" function.

Hint: The _lunp and .ps files remains from the previous execution of RNAplfold programme, so that if there is an error in running the RNAplfold programm that leads it to not forming a new _lunp file that overwrites the old file if present, the user will not parse the old file.

References

Li, X., Kazan, H., Lipshitz, H. D., & Morris, Q. D. (2014). Finding the target sites of RNA-binding proteins. *Wiley Interdisciplinary Reviews: RNA*, *5*(1), 111-130.

Ramos, A., Grünert, S., Adams, J., Micklem, D. R., Proctor, M. R., Freund, S., & Varani, G. (2000). RNA recognition by a Staufen double-stranded RNA-binding domain. *The EMBO journal*, *19*(5), 997-1009.

Masliah, G., Barraud, P., & Allain, F. H. T. (2013). RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cellular and Molecular Life Sciences*, *70*(11), 1875-1895.

Aviv, T., Lin, Z., Ben-Ari, G., Smibert, C. A., & Sicheri, F. (2006). Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. *Nature structural & molecular biology*, *13*(2), 168.

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., ... & Na, H. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, *499*(7457), 172.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., ... & Darnell, J. C. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, *456*(7221), 464.

Kazan, H., & Morris, Q. (2013). RBPmotif: a web server for the discovery of sequence and structure preferences of RNA-binding proteins. *Nucleic acids research*, *41*(W1), W180-W186.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26.