Assignment 1.

Shilova Liubov (new student, do not have a matriculation number yet)

Ex.1.

In order to make a master table, we do not need subversions of miRBase. Thus, right away I filtered the data and made a list of folders, that have integer value or end with ".0".

In the last version there is an .xls file, called "miRNA.xls" which already contains a lot of information. It is important, that all miRNAs, that are marked as "dead" (were proved to be false positives) are not in this document. So, taking this file, I already had most of needed columns.

What we do not have here is the release. In order to do find that out, I collected the files, that end with ".diff". All versions from 4 to 22 have them and there all newly discovered miRNA are stated.

However, first 3 releases do not contain these files. Thus, I needed original fasta files with newly discovered miRNA.

Opening each of file, that ends with ".fa" or ".fa.gz" in releases 1.0, 2.0 and 3.0, I checked if they contain information about at least one of my four species (they all did) and downloaded them. For checking the release number, where the miRNA appeared first, a file with precursor was enough, so I did not download sequencing data of mature miRNA (it was already in a table from release 22).

In the end, I had the list of fallowing links:

```
['ftp://mirbase.org/pub/mirbase/22/miRNA.xls/miRNA.csv,

 'ftp://mirbase.org/pub/mirbase/10.0/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/11.0/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/12.0/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/13.0/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/14/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/15/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/16/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/17/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/18/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/19/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/20/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/21/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/22/miRNA.diff.gz',
 'ftp://mirbase.org/pub/mirbase/22/miRNA.diff',
 'ftp://mirbase.org/pub/mirbase/4.0/miRNA.diff',
 'ftp://mirbase.org/pub/mirbase/5.0/miRNA.diff',
 'ftp://mirbase.org/pub/mirbase/6.0/miRNA.diff',
 'ftp://mirbase.org/pub/mirbase/7.0/miRNA.diff',
 'ftp://mirbase.org/pub/mirbase/8.0/miRNA.diff',
 'ftp://mirbase.org/pub/mirbase/9.0/miRNA.diff.gz'

 'ftp://mirbase.org/pub/mirbase/1.0/cel.dat.gz',
 'ftp://mirbase.org/pub/mirbase/1.0/hsa.dat.gz',
 'ftp://mirbase.org/pub/mirbase/1.0/mmu.dat.gz',
 'ftp://mirbase.org/pub/mirbase/2.0/precursor.fa.gz',
 'ftp://mirbase.org/pub/mirbase/3.0/precursor.fa']
```

Ex. 2

Mentioning: I worked with data, already downloaded on my PK. Thus, in code I am opening files by writing or finding a path on my computer, rather than on FTP server. It might be not an optimal solution, as it makes difficult to check my program from other servers and I am sorry for that. I have realized it only being too close to deadline. On my PK the data was stored on disk E in folder "RNA". Than the structure of data is the same as on FTP link.

1. We only need 4 species: "hsa", "mmu", "rno", and "cel". Thus, I deleated all raws with miRNAs of other species.
2. Then, going through all ".diff" files, if I find ID of a needed precursor and the word NEW in the same line, I put the number of a release (heading of a foulder) in the according cell of "Release" column in my table.
3. For first 3 versions of miRBase I look for Precursor name in .fa files. If I find it in folder "1.0", I put "1.0" in according cell and don't look for this name anywhere else. Only if I didn't find it in first, I will go to second folder. Thus, I am registering the very first appearance of the name.

The script for this task is called "Exersise2-Copy1.py"

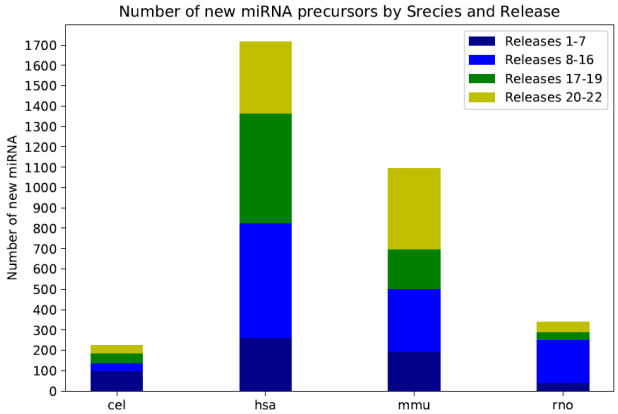
The master table is called "Master_table.csv"

Ex. 3

a. Showing the absolute number of miRNA precursors per species and miRBase version

Fig. 1 (a) shows the number of miRNA precursors of cel (Caenorhabditis elegans), hsa (Homo Sapiens), mmu (Mus musculus), rno (Rattus norvegicus) that were newly discovered in different releases of miRBase.
It can be easily seen that the biggest quantity of precursors was identified for humans, which is logical, because humans are mostly interested in studying themselves. The second popular is Mus musculus as a very popular model organism. C. elegans and R. norvegicus are also good model organisms, but they were less popular in all of the releases.
Also, last releases are very huge comparing to the first ones. In the last 5 releases more than a thousand of new miRNAs was found for humans, which is twice as much as was found before (releases 1 to 16). However, it seems like C. elegance and R. norvegicus were studied less, and in last years the quantity of miRNAs discovered in these species rose only slightly.

a)



b)

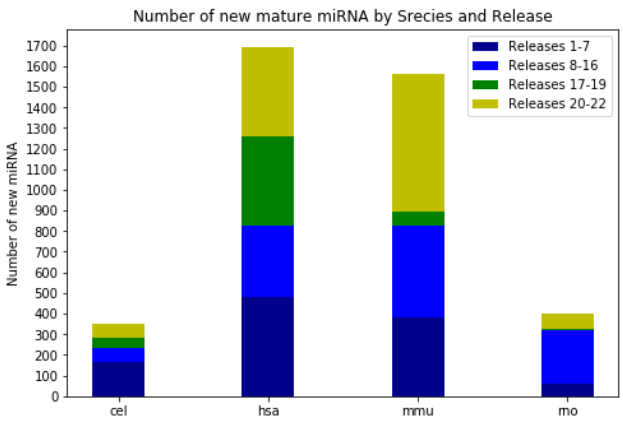


Fig. 1. Number of new precursers (a) and mature miRNA (b) discovered in different species in different releases.

b. Showing the absolute number of mature miRNAs per species and miRBase version

Fig. 1 (b) shows the number of mature miRNA of cel (Caenorhabditis elegans), hsa (Homo Sapiens), mmu (Mus musculus), rno (Rattus norvegicus) that were newly discovered in different releases of miRBase.
Here, as for precursors, H. sapiens is the most popular subject to study and, consequently, has the biggest number of new miRNA.
However, if we compare this to the quantity of miRNA precursors, we can see, that for H. sapiens the number of mature miRNAs is almost the same as the number of precursors. Ideally, as each precursor should have 2 miRNAs from each string. So, the number of mature RNAs should be twice larger than the number of precursors. It is approximately so for M. musculus. Nevertheless, it seems like for many precursors of H. sapiens mature miRNAs were not found. This may question how real are the precursors, that we have predicted in recent years

c. Showing for each miRBase version the ratio of added precursors and added mature miRNAs per species.

Continuing the thought from the previous subtask, we can compare the mature miRNA/ precursor miRNA ratio. Ideally, it should be around 2, as with one miRNA precursor two mature miRNA should be described. However, for latest releases, this ratio is much smaller. Again, this means, that some precussor miRNAs were described without one or two of the mature miRNAs, that should exist in order to make it real precursor.

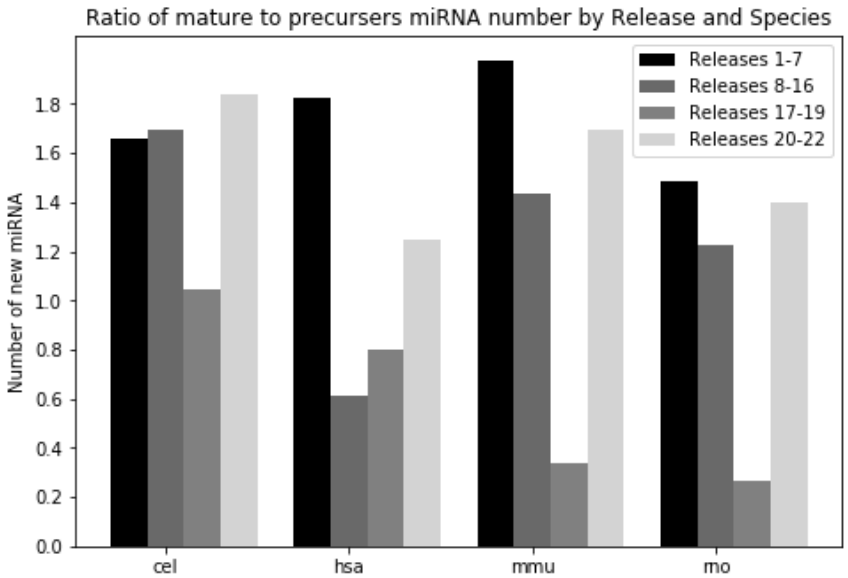


Fig. 2. The Ratio between number of mature miRNAs and number of precursors miRNAs pro species and release group

Ex. 4.

Further, the initial master table was extended with columns with different nucleotides content in precursor and mature miRNA. The resulting table was called “Extanded_Master_table.csv”. First, let us look at precursors (Fig. 3 and 4). Here, we can make a few observations:

1. The content varies depending on a species. For example, for C. elegance the GC content is visibly lower, than for other species.
2. In the last releases for some species (H. sapiens and M. musculus) the average GC content has increased significantly, whereas the AU content decreased.
   High GC content decreases the free energy of a molecule, as there are 3 bands between G and C. Subsequently, the program, which calculates the free energy of a molecule is more likely to pick the molecule as a possible miRNA precursor. However, real precursors do not have that high GC content. Therefore, we found one more reason for big quantity of false positives in miRNA data.

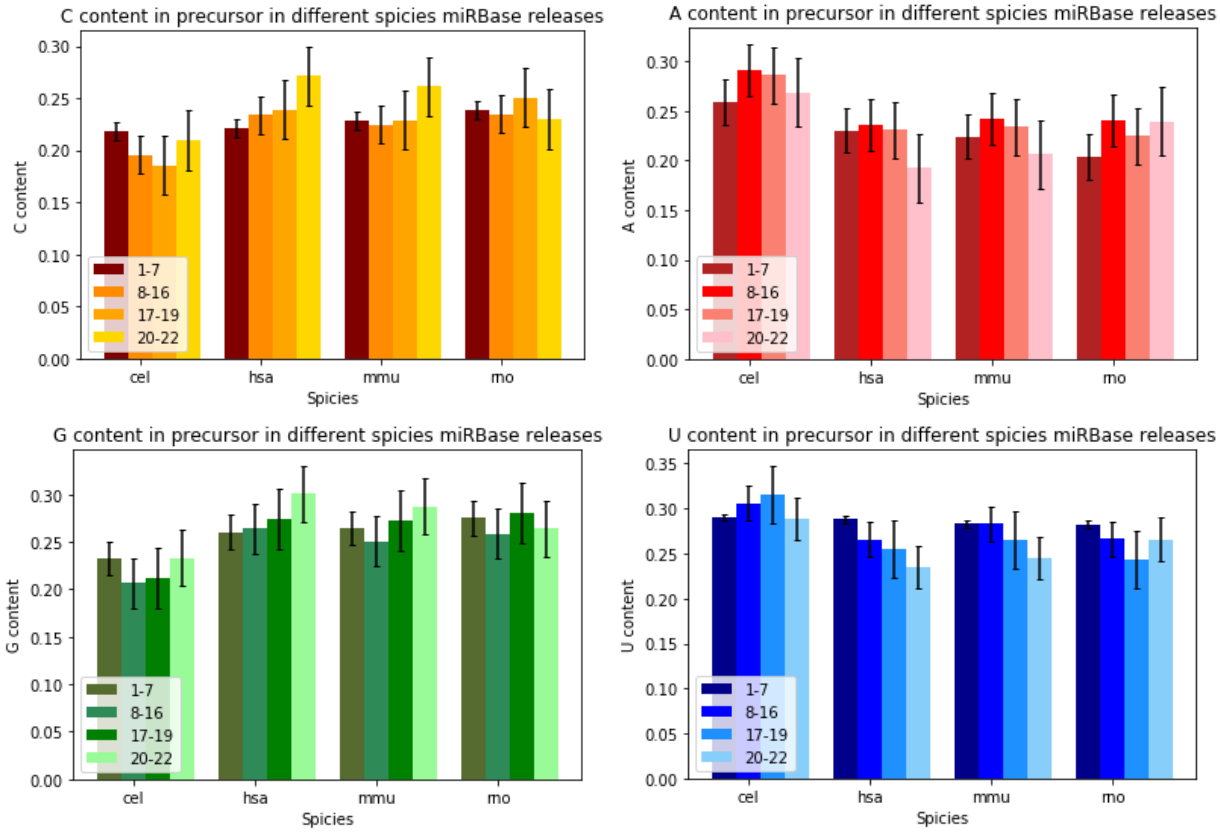


Fig. 3 C, A, G, U content in the precursor of miRNA in different releases and different species. Data presented as Mean ± Standard Deviation. The lighter the color, the later the release.

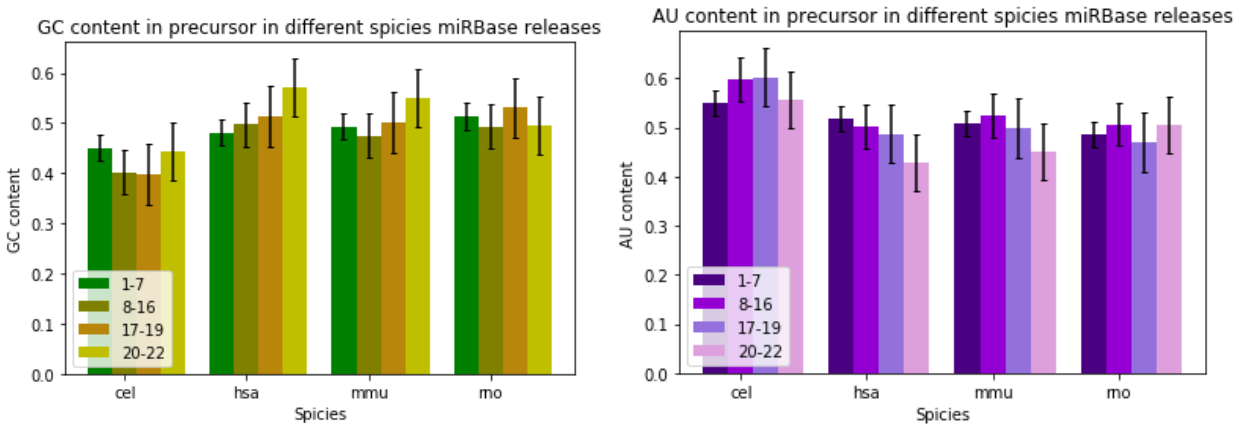


Fig. 4 GC and AU content in the precursor of miRNA in different releases and different species. Data presented as Mean ± Standard Deviation. The lighter the color, the later the release.

Conserning mature miRNA:

1. The GC content of C. ellegans is also lower than of other species in all releases.
2. Later releases have higher GC content and lower AU content.

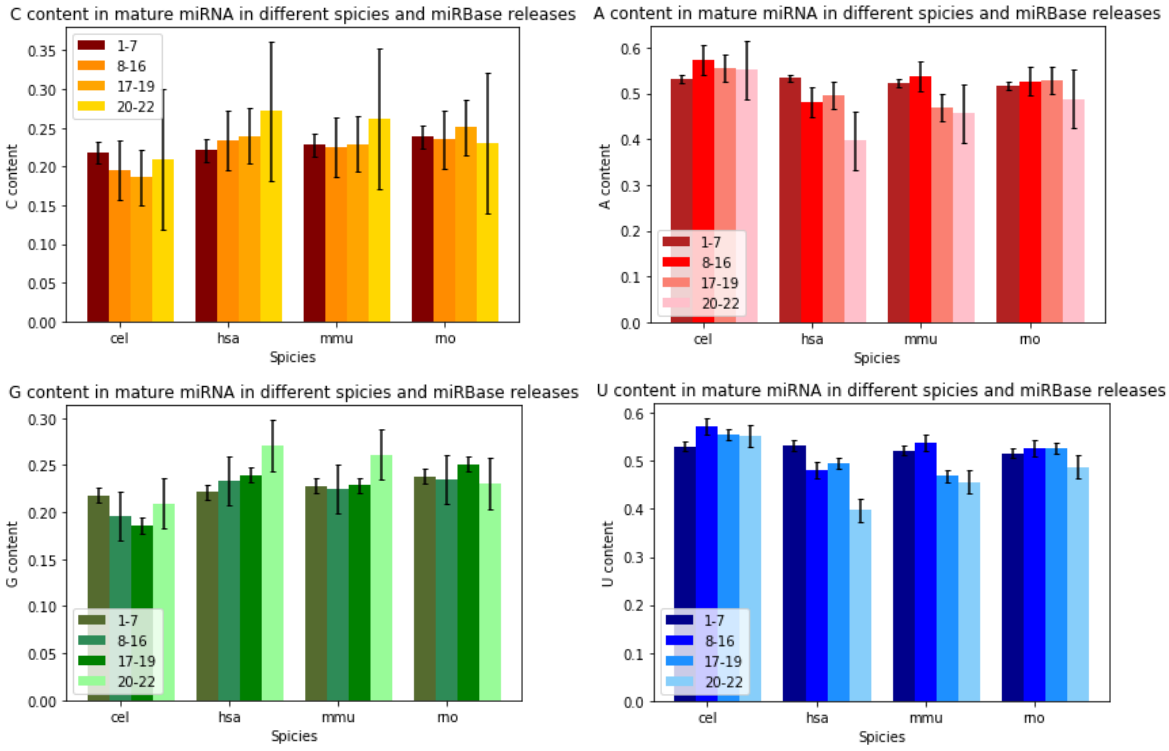


Fig. 5 C, A, G, U content in the precursor of miRNA in different releases and different species. Data presented as Mean ± Standard Deviation. The lighter the color, the later the release.

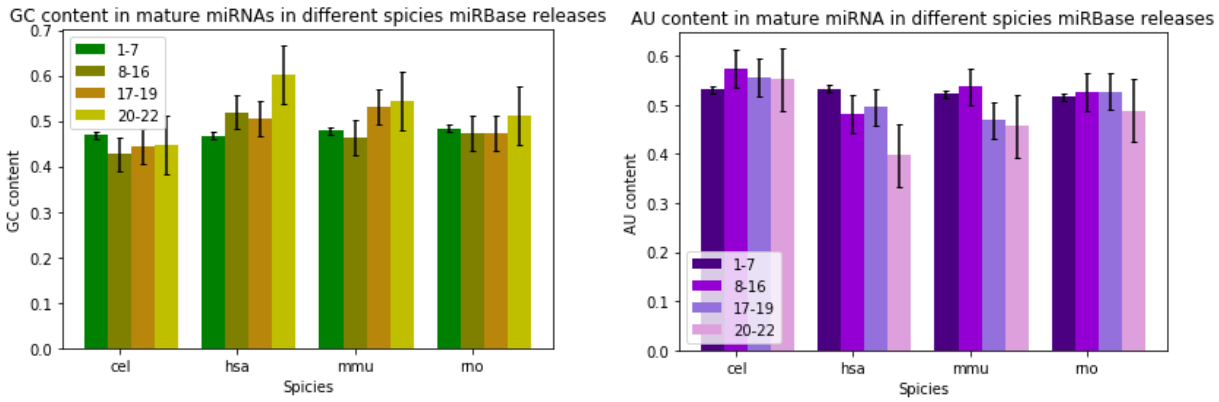


Fig. 6. GC and AU content in the precursor of miRNA in different releases and different species. Data presented as Mean ± Standard Deviation. The lighter the color, the later the release.

Ex. 5.

The main observations were explained in ex. 4.

Overall, the results are:

1. The biggest number of miRNAs was described for human, almost 2000 (without "dead" ones, that were proven to be false positives). Almost the half of them was described in latest virsions.
2. In first releases there were less miRNAs, however, they seem to be more definite. In early releases the mature/precursor ratio was close to 2. That means, that for every precursor 2 mature miRNAs were described. Also, the GC content was lower than in the next versions.
3. Nucleotide content varies depending on a species. For C. elegans the GC content was generally lower, than for H. sapience, M. musculus and R. norvegicus.

We can conclude, that many of recently described miRNA precursors may be false positives. In order to be sure, we need to find according matute miRNAs and look at the GC content, when evaluating free energy of a possible precursor.

P.S. Thank you for the interesting homework! I feel like a Mathplotlib goddess now ☺