

Homework 4: miR regulatory network

1

- <hw4.tsv> is a tab-separated values (tsv) file that contains the predicted targets of human microRNAs (adapted from TargetScan 6.2)

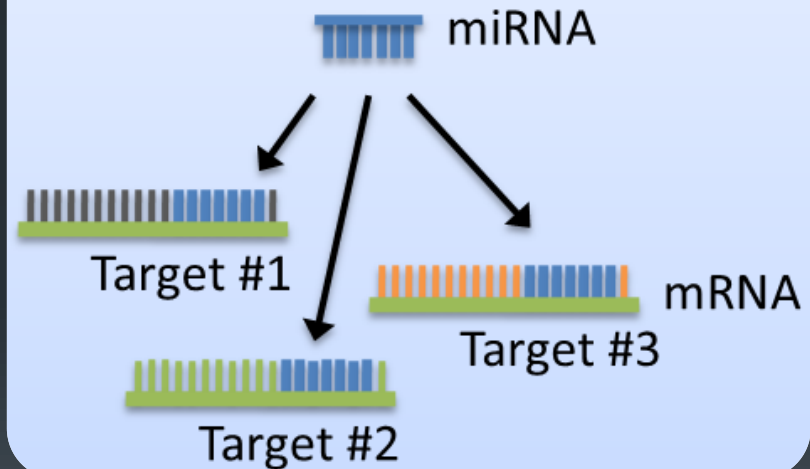
Preview in Excel

miR Family	Gene Symbol
let-7/98/4458/4500	A1CF
let-7/98/4458/4500	AAK1
let-7/98/4458/4500	AASS
let-7/98/4458/4500	ABCB9
let-7/98/4458/4500	ABCC10
let-7/98/4458/4500	ABCC5



The escape sequence for “tab” is ‘\t’

miR regulatory network



- <hw4.tsv> has been sorted and uniquified.

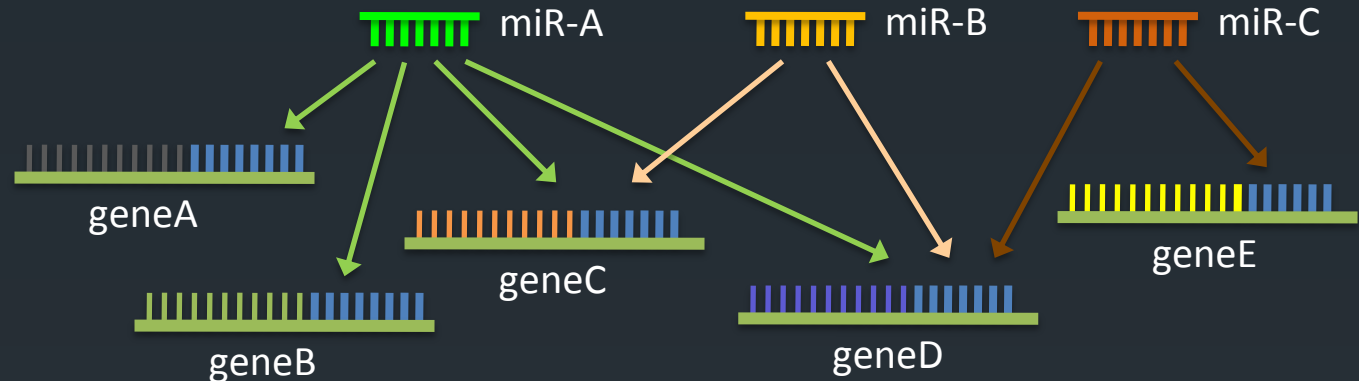
(Optional) For more information about the miR regulatory network, you may visit

<http://www.biomedcentral.com/1471-2105/12/S1/S41>

<input file>

miR Family	Gene Symbol
miR-A	geneA
miR-A	geneB
miR-A	geneC
miR-A	geneD
miR-B	geneC
miR-B	geneD
miR-C	geneD
miR-C	geneE

A toy example



The network view

Properties

	Out-degree
miR-A	4
miR-B	2
miR-C	2

	In-degree
geneA	1
geneB	1
geneC	2
geneD	3
geneE	1

Family pairs	Number of shared targets
miR-A & miR-B	2
miR-B & miR-C	1
miR-A & miR-C	1

Homework 4: miR regulatory network

3

- Write a Python program to construct the network
 - Print out the following information:
 - The name of the miR family that has the max out-degree
 - The number of the max out-degree
 - Average out-degree of all miR families (round to 2 decimal point)
 - The symbol of the gene that has the max in-degree
 - The number of the max in-degree
 - Average in-degree of all genes (round to 2 decimal point)
 - The names of the 2 distinct miR families that have most shared targets
 - The number of the most shared targets

```
let-7/98
876
54.32
TP53
21
3.54
miR-123 miR-456
101
```

輸出範例



- You may use **dicts** and **sets** to implement the network.
- When output, use single space to separate strings.

Homework 4: Hints

- File reading
 - 參考作業三或是本次上課內容
- 建立出網路的資料結構
 - 在紙上先畫畫圖可能有幫助
- 在 dictionary 中找出最大值以及擁有最大值的 key(s)
 - 有最大值的 key(s) 不一定只有一個
 - 有多種解法，例如 sorting, 或是先找出最大值再回頭找 key(s)
- 計算 in-degrees for target genes 其實只是反過來
 - 同樣地，改變資料結構或許是個好方法
 - From top-down view to bottom-up view
- What are shared targets?
 - 指的就是兩個 miR families 各自標的基因群的重覆部份
- 本次作業理論上僅憑課堂上教過的基本語法與資料結構便可完成。有興趣的同學，亦可參考使用其他容器或迭代法：
 - <http://docs.python.org/library/collections.html>
 - <http://docs.python.org/library/itertools.html>

More hints for homework 4

5

- 將任何容器傳入 `len` 函數中，會得到目前該容器中元素的個數

- 如何創造出空的容器：

- `[]` or `list()` 可以得到空的 list
- `set()` 可以得到空的 set
- `{}` or `dict()` 可以得到空的 dict

- 如何創造只含一個元素的容器：

e.g., 以下容器都會只含有 1 這個數字

- `[1]` 可以得到只包含 1 的 list, 等同 `list((1,))` 或
- `(1,)` 可以得到只包含 1 的 tuple, 等同 `tuple([1])` 或 `tuple((1,))`
- `{1}` 可以得到只包含 1 的 set, 等同 `set([1])` 或 `set((1,))`
- `{1: None}` 可以得到只包含 (1, None) 的 dict, 等同 `dict([(1, None)])`

```
>>> a = [4, 5, 6]
>>> len(a)
3
>>> b = set([7, 8, 9])
>>> len(b)
3
>>> c = {'x': 1, 'y': 2, 'z': 3}
>>> len(c)
3
```

Traps!

6

Preview in Excel

miR Family	Gene Symbol
let-7/98	A2BP1
let-7/98	AAK1
let-7/98	ABCB9
let-7/98	ABCC5
let-7/98	ABL2
let-7/98	ACSL6
let-7/98	ACTR10
let-7/98	ACVR1B
let-7/98	ACVR1C

```
>>> D = {}  
>>> D['let-7/98'] = 'A2BP1'  
>>> D  
{'let-7/98': 'A2BP1'}  
>>> D['let-7/98'] = 'AAK1'  
>>> D  
{'let-7/98': 'AAK1'}
```



字典裡的值不一定只能用數字或字串，
也可以是某**容器(Collection)**

Bonus

- From our site, download `<Predicted_Targets_Info.txt.zip>*`
 - Unzip the file to get the text file `<Predicted_Targets_Info.txt>`
 - Change your file source from `<hw4.tsv>` to this file
- Do the processes in homework 4
 - This time you have to parse the table on your own
- Let users decide the species
 - Users will enter a Tax ID (e.g., Tax ID for human, mouse, and rat: 9606, 10090, and 10116)
 - Print out the network information of that species



A terminal window with a dashed border and a cursor icon in the top-left corner. It displays a list of values in green text: 9606, let-7/98, 876, 54.32, TP53, 21, 3.54, miR-123 miR-456, and 101. A large orange bracket on the right side groups all these values together. To the right of the bracket, there are two stacked orange callout boxes: the top one contains the word "human" and the bottom one contains the text "輸出範例" (Output Example).

9606	let-7/98	876	54.32	TP53	21	3.54	miR-123 miR-456	101
------	----------	-----	-------	------	----	------	-----------------	-----

human
輸出範例



A terminal window with a dashed border and a cursor icon in the top-left corner. It displays a list of values in green text: 10090, miR-124, 456, 12.32, NFIX, 13, 3.14, miR-789 miR-456, and 598. A large orange bracket on the right side groups all these values together. To the right of the bracket, there are two stacked orange callout boxes: the top one contains the word "mouse" and the bottom one contains the text "輸出範例" (Output Example).

10090	miR-124	456	12.32	NFIX	13	3.14	miR-789 miR-456	598
-------	---------	-----	-------	------	----	------	-----------------	-----

mouse
輸出範例