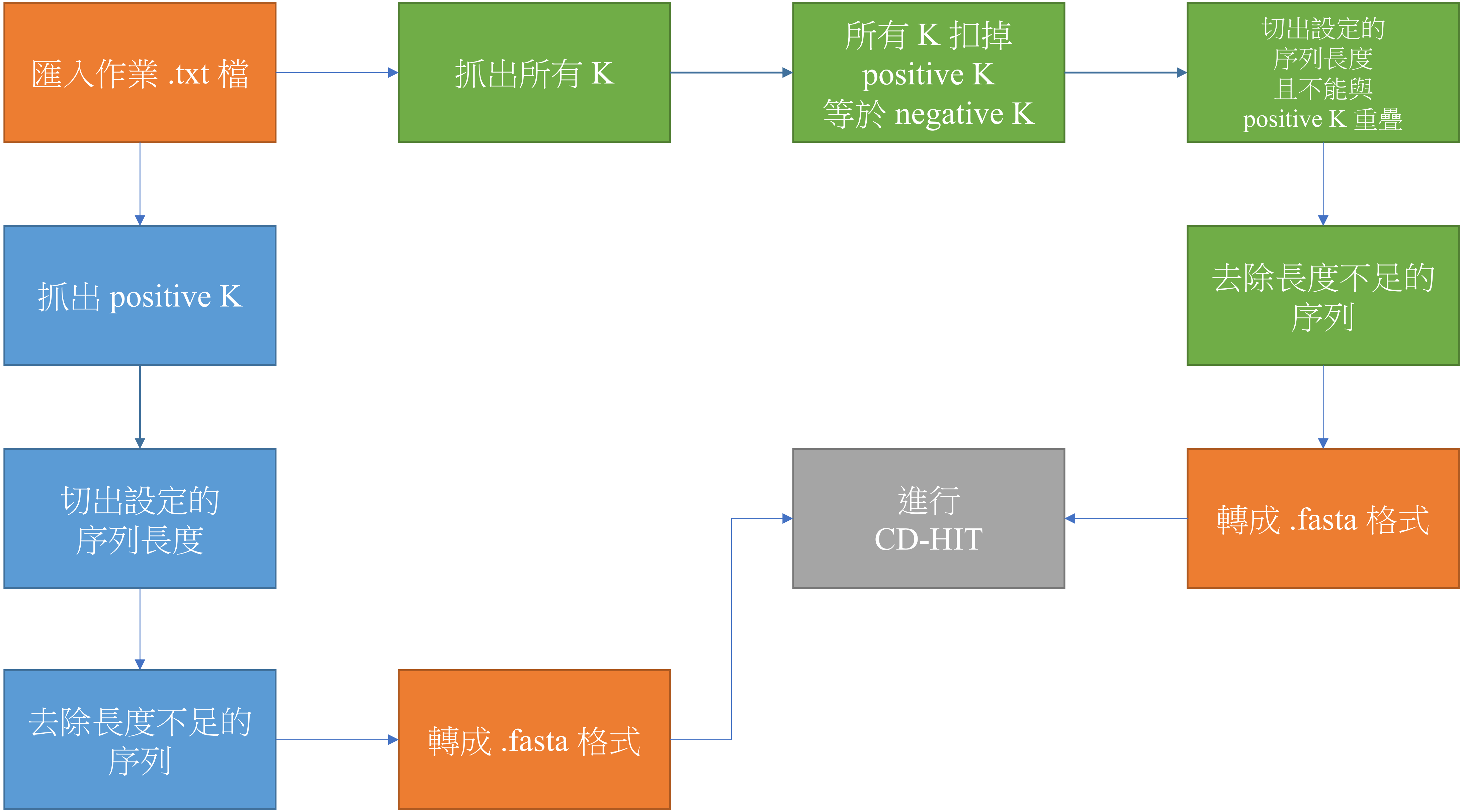
Google Colab

Colab (全名為「Colaboratory」)，是 Google Research 所推出的一項產品。它是一個基於 **Jupyter Notebook** 的雲端開發環境，可以讓你透過瀏覽器編寫及執行 **Python** 程式碼，也可以進行資料分析及機器學習的工具，無須任何設定即可使用。

Colab 具有以下優點：

- 不必進行任何設定即可輕鬆上手使用雲端開發環境
- 不用安裝即可使用 Python 編寫和執行代碼
- 輕鬆建立/上傳/共享筆記本(notebooks)

https://medium.com/python4u/google-colab-%E6%95%99%E5%AD%B8-1-python-%E9%9B%B2%E7%AB%AF%E9%96%8B%E7%99%BC%E7%92%B0%E5%A2%83%E5%AE%89%E8%A3%9D%E8%88%87%E5%BF%AB%E9%80%9F%E5%B0%8E%E8%A6%BD-78942200525f



匯入作業 .txt 檔

抓出 positive K

切出設定的
序列長度

去除長度不足的
序列

轉成 .fasta 格式

抓出所有 K

所有 K 扣掉
positive K
等於 negative K

切出設定的
序列長度
且不能與
positive K 重疊

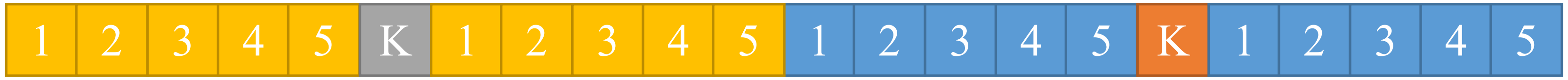
去除長度不足的
序列

轉成 .fasta 格式

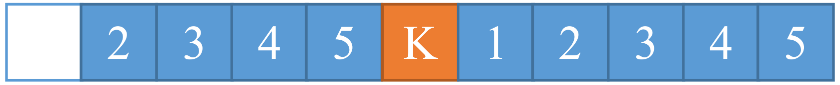
進行
CD-HIT

🟧 K positive K

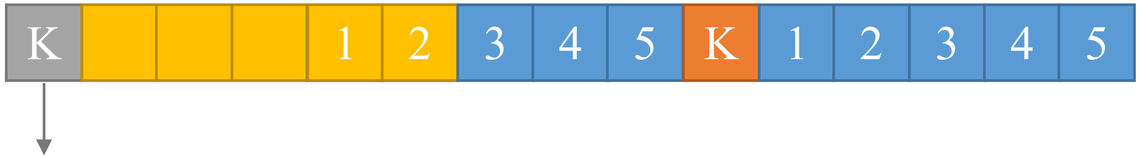
⬜ K negative K



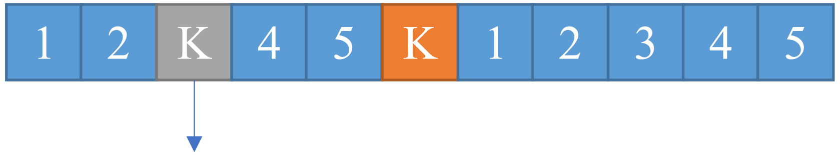
negative K 與 positive K 皆採用



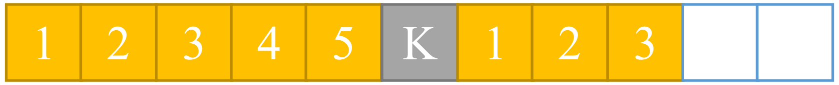
positive K 皆採用
除了長度不足不採用



negative K 與 positive K 範圍有重疊
不採用

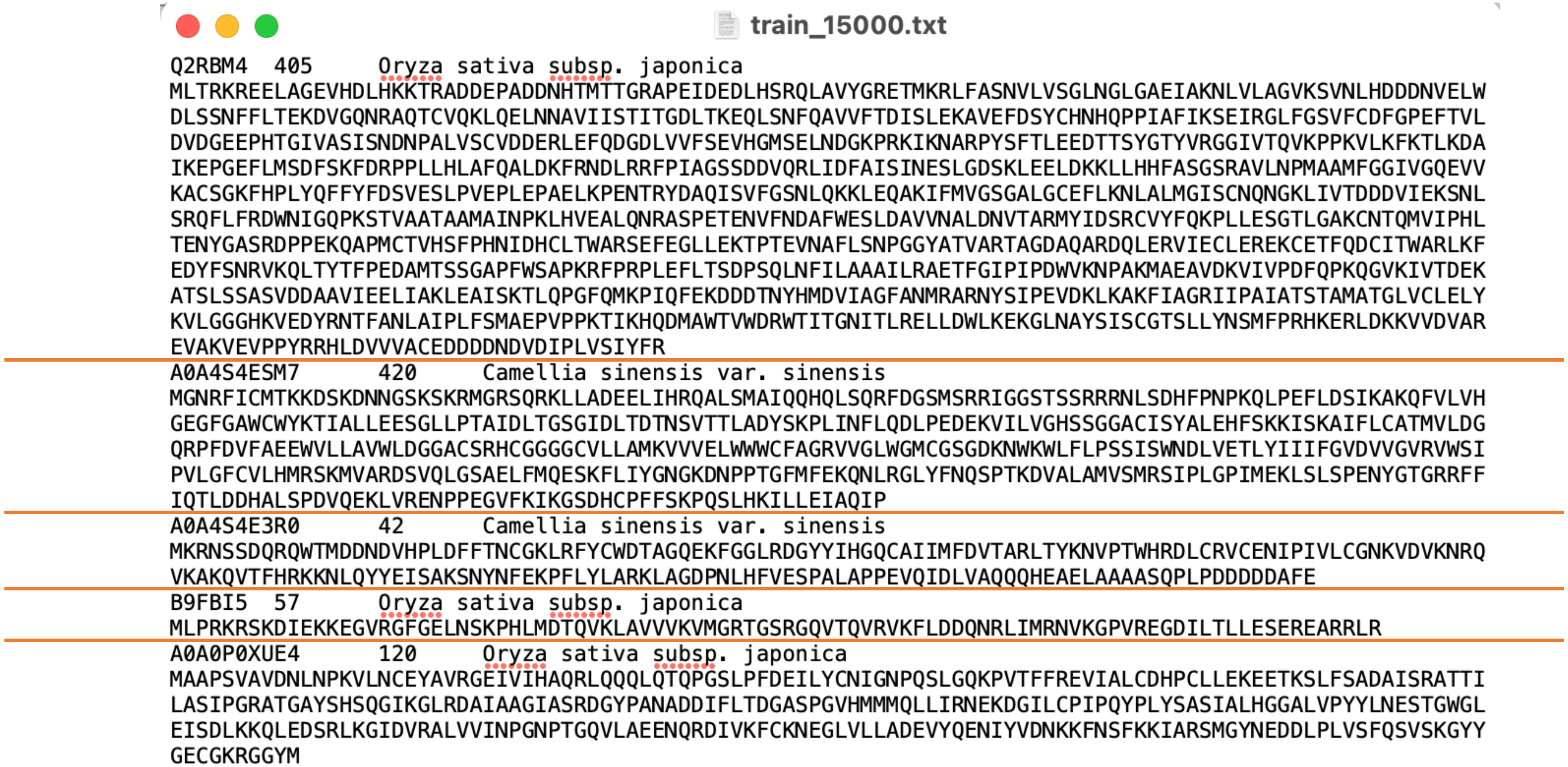


negative K 在 positive K 範圍內
不採用



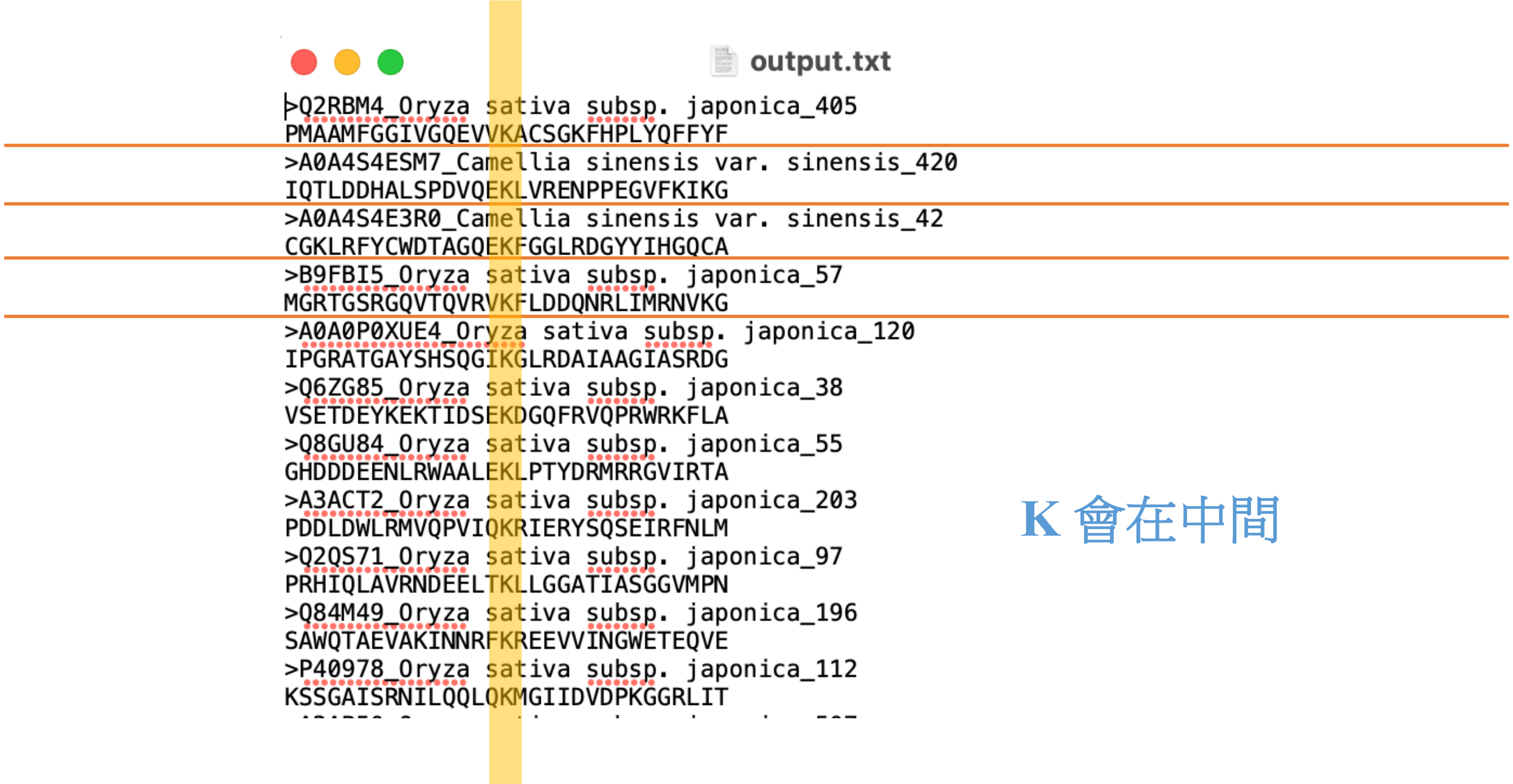
序列長度不足
不採用

input.txt



```
Q2RBM4  405     Oryza sativa subsp. japonica
MLTRKREELAGEVHDLHKKTRADDEPADDNHTMTTGRAPEIDEDLHSRQLAVYGRETMKRLFASNVLVSGLNGLGAEIAKNLVLAGVKSVNLHDDDNVELW
DLSSNFFLTEKDVGQNRAQTCVQKLQELNNAVIISTITGDLTKEQLSNFQAVVFTDISLEKAVEFDSYCHNHQPPIAFIKSEIRGLFGSVFCDFGPEFTVL
DVDGEEPHTGIVASISNDNPALVSCVDDERLEFQDGDLVVFSEVHGMSELNDGKPRKIKNARPYSFTLEEDTTSYGTYVRGGIVTQVKPPKVLKFKTLKDA
IKEPGEFLMSDFSKFDRPPLLHLAFQALDKFRNDLRRFPIAGSSDDVQRLIDFAISINESLGDSKLEELDKKLLHHFASGSRAVLNPMAAMFGGIVGQEVV
KACSGKFHPLYQFFYFDSVESLPVEPLEPAELKPENTRYDAQISVFGSNLQKKLEQAKIFMVGSGALGCEFLKNLALMGISCNQNGKLIVTDDDVIEKSNL
SRQFLFRDWNIGQPKSTVAATAAMAINPKLHVEALQNRASPETENVFNDAFWESLDAVVNALDNVTARMYIDSRCVYFQKPLLESGTLGAKCNTQMVIPHL
TENYGASRDPPEKQAPMCTVHSFPHNIDHCLTWARSEFEGLLEKTPTEVNAFLSNPGGYATVARTAGDAQARDQLERVIECLEREKCETFQDCITWARLKF
EDYFSNRVKQLTYTFPEDAMTSSGAPFWSAPKRFPRPLEFLTSDPSQLNFILAAAILRAETFGIPIPDWVKNPAKMAEAVDKVIVPDFQPKQGVKIVTDEK
ATSLSSASVDDAAVIEELIAKLEAISKTLQPGFQMKPIQFEKDDDTNYHMDVIAGFANMRARNYSIPEVDKLKAKFIAGRIIPAIATSTAMATGLVCLELY
KVLGGGHKVEDYRNTFANLAIPLFSMAEPVPPKTIKHQDMAWTVWDRWTITGNITLRELLDWLKEKGLNAYSISCGTSLLYNSMFPRHKERLDKKVVDVAR
EVAKVEVPPYRRHLDVVVACEDDDDNDVDIPLVSIYFR
A0A4S4ESM7      420     Camellia sinensis var. sinensis
MGNRFICMTKKDSKDNNGSKSKRMGRSQRKLLADEELIHRQALSMAIQQHQLSQRFDGSMSRRIGGSTSSRRRNLSDHFPNPKQLPEFLDSIKAKQFVLVH
GEGFGAWCWYKTIALLEESGLLPTAIDLTGSGIDLTDTNSVTTLADYSKPLINFLQDLPEDEKVILVGHSSGGACISYALEHFSKKISKAIFLCATMVLDG
QRPFDVFAEEWVLLAVWLDGGACSRHCGGGGCVLLAMKVVVELWWWCFAGRVVGLWGMCGSGDKNWKWLFLPSSISWNDLVETLYIIIFGVDVVGVRVWSI
PVLGFCVLHMRSKMVARDSVQLGSAELFMQESKFLIYGNGKDNPPTGFMFEKQNLRGLYFNQSPTKDVALAMVSMRSIPLGPIMEKLSLSPENYGTGRRFF
IQTLDDHALSPDVQEKLVRENPPEGVFKIKGSDHCPFFSKPQSLHKILLEIAQIP
A0A4S4E3R0      42      Camellia sinensis var. sinensis
MKRNSSDQRQWTMDDNDVHPLDFFTNCGKLRFYCWDTAGQEKFGGLRDGYYIHGQCAIIMFDVTARLTYKNVPTWHRDLCRVCENIPIVLCGNKVDVKNRQ
VKAKQVTFHRKKNLQYYEISAKSNYNFEKPFLYLARKLAGDPNLHFVESPALAPPEVQIDLVAQQQHEAELAAAASQPLPDDDDDAFE
B9FBI5  57      Oryza sativa subsp. japonica
MLPRKRSKDIEKKEGVRGFGELNSKPHLMDTQVKLAVVVKVMGRTGSRGQVTQVRVKFLDDQNRLIMRNVKGPVREGDILTLLESEREARRLR
A0A0P0XUE4      120     Oryza sativa subsp. japonica
MAAPSVAVDNLNPKVLNCEYAVRGEIVIHAQRLQQQLQTQPGSLPFDEILYCNIGNPQSLGQKPVTFFREVIALCDHPCLLEKEETKSLFSADAISRATTI
LASIPGRATGAYSHSQGIKGLRDAIAAGIASRDGYPANADDIFLTDGASPGVHMMMQLLIRNEKDGILCPIPQYPLYSASIALHGGALVPYYLNESTGWGL
EISDLKKQLEDSRLKGIDVRALVVINPGNPTGQVLAEENQRDIVKFCKNEGLVLLADEVYQENIYVDNKKFNSFKKIARSMGYNEDDLPLVSFQSVSKGYY
GECGKRGGYM
```

output.fasta

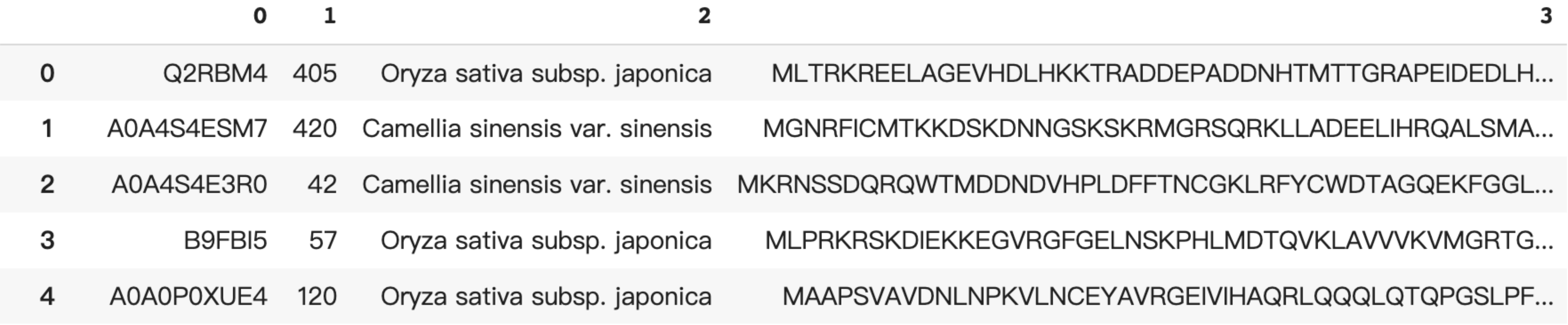


```
>Q2RBM4_Oryza sativa subsp. japonica_405
PMAAMFGGIVGQEVVKACSGKFHPLYQFFYF
>A0A4S4ESM7_Camellia sinensis var. sinensis_420
IQTLDDHALSPDVQEKLVRENPPEGVFKIKG
>A0A4S4E3R0_Camellia sinensis var. sinensis_42
CGKLRFYCWDTAGQEKFGGLRDGYYIHGQCA
>B9FBI5_Oryza sativa subsp. japonica_57
MGRTGSRGQVTQVRVKFLDDQNRLIMRNVKG
>A0A0P0XUE4_Oryza sativa subsp. japonica_120
IPGRATGAYSHSQGIKGLRDAIAAGIASRDG
>Q6ZG85_Oryza sativa subsp. japonica_38
VSETDEYKEKTIDSEKDGQFRVQPRWRKFLA
>Q8GU84_Oryza sativa subsp. japonica_55
GHDDDEENLRWAALEKLPTYDRMRRGVIRTA
>A3ACT2_Oryza sativa subsp. japonica_203
PDDLDWLRMVQPVIQKRIERYSQSEIRFNLM
>Q2QS71_Oryza sativa subsp. japonica_97
PRHIQLAVRNDEELTKLLGGATIASGGVMPN
>Q84M49_Oryza sativa subsp. japonica_196
SAWQTAEVAKINNRFKREEVVINGWETEQVE
>P40978_Oryza sativa subsp. japonica_112
KSSGAISRNILQQLQKMGIIDVDPKGGRLIT
```

K 會在中間

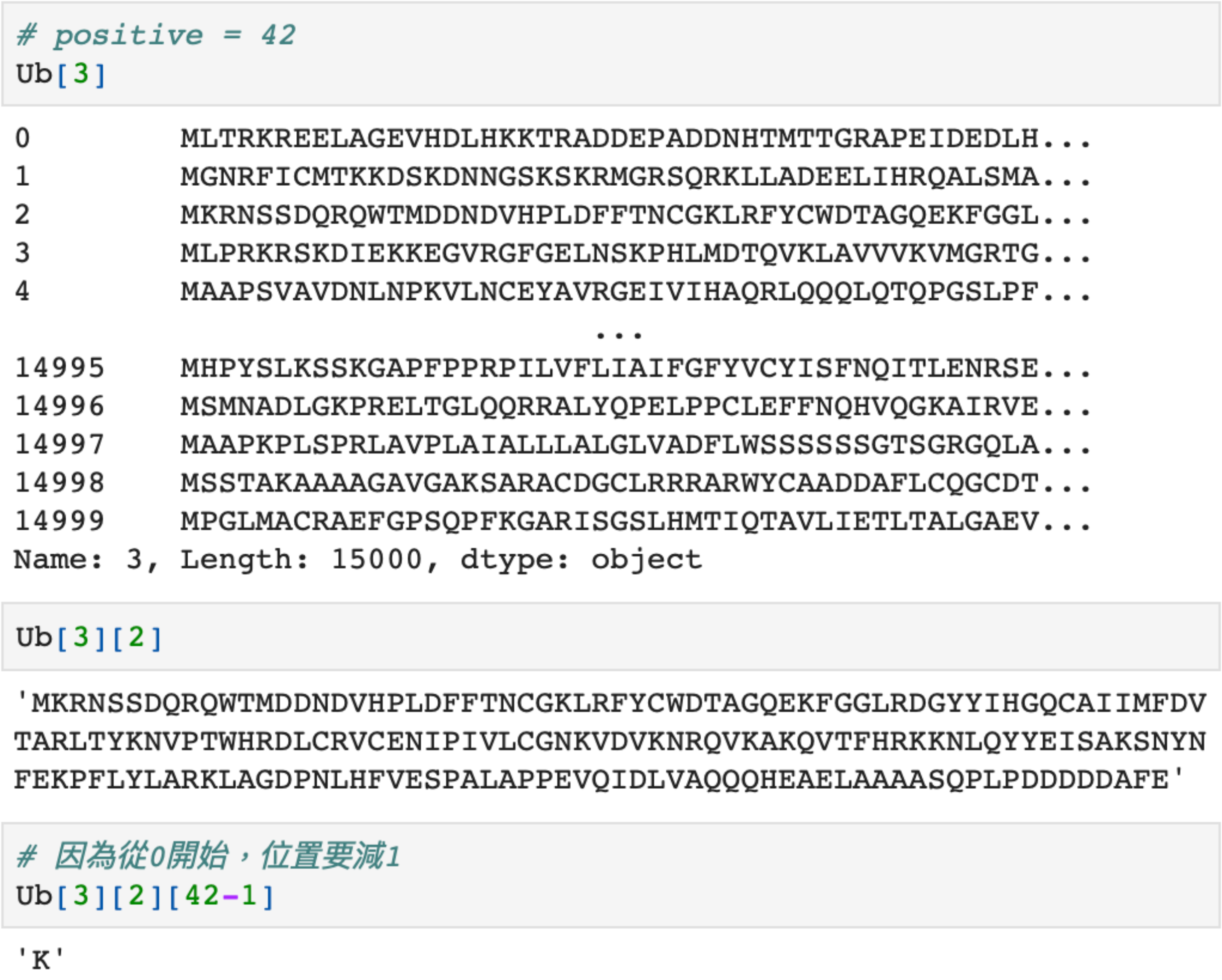
讀檔

```
import pandas as pd
Ub = pd.read_csv(‘檔案', sep='\t', header=None)
```



	0	1	2	3
0	Q2RBM4	405	Oryza sativa subsp. japonica	MLTRKREELAGEVHDLHKKTRADDEPADDNHTMTTGRAPEIDEDLH...
1	A0A4S4ESM7	420	Camellia sinensis var. sinensis	MGNRFICMTKKDSKDNNGSKSKRMGRSQRKLLADEELIHRQALSMA...
2	A0A4S4E3R0	42	Camellia sinensis var. sinensis	MKRNSSDQRQWTMDDNDVHPLDFFTNCGKLRFYCWDTAGQEKFGGL...
3	B9FBI5	57	Oryza sativa subsp. japonica	MLPRKRSKDIEKKEGVRGFGELNSKPHLMDTQVKLAVVVKVMGRTG...
4	A0A0P0XUE4	120	Oryza sativa subsp. japonica	MAAPSVAVDNLNPKVLNCEYAVRGEIVIHAQRLQQQLQTQPGSLPF...

抓 positive 位置 K



```
# positive = 42
Ub[3]
```

```
0        MLTRKREELAGEVHDLHKKTRADDEPADDNHTMTTGRAPEIDEDLH...
1        MGNRFICMTKKDSKDNNGSKSKRMGRSQRKLLADEELIHRQALSMA...
2        MKRNSSDQRQWTMDDNDVHPLDFFTNCGKLRFYCWDTAGQEKFGGL...
3        MLPRKRSKDIEKKEGVRGFGELNSKPHLMDTQVKLAVVVKVMGRTG...
4        MAAPSVAVDNLNPKVLNCEYAVRGEIVIHAQRLQQQLQTQPGSLPF...
                                  ...
14995    MHPYSLKSSKGAPFPPRPILVFLIAIFGFYVCYISFNQITLENRSE...
14996    MSMNADLGKPRELTGLQQRRALYQPELPPCLEFFNQHVQGKAIRVE...
14997    MAAPKPLSPRLAVPLAIALLLALGLVADFLWSSSSSSGTSGRGQLA...
14998    MSSTAKAAAAGAVGAKSARACDGCLRRRARWYCAADDAFLCQGCDT...
14999    MPGLMACRAEFGPSQPFKGARISGSLHMTIQTAVLIETLTALGAEV...
Name: 3, Length: 15000, dtype: object
```

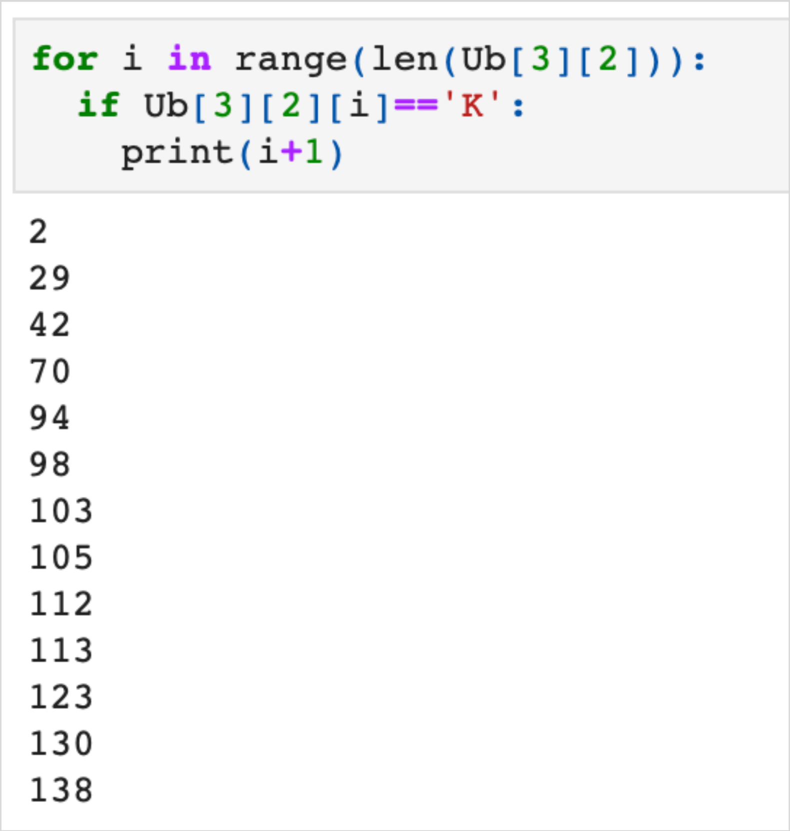
```
Ub[3][2]
```

```
'MKRNSSDQRQWTMDDNDVHPLDFFTNCGKLRFYCWDTAGQEKFGGLRDGYYIHGQCAIIMFDV
TARLTYKNVPTWHRDLCRVCENIPIVLCGNKVDVKNRQVKAKQVTFHRKKNLQYYEISAKSNYN
FEKPFLYLARKLAGDPNLHFVESPALAPPEVQIDLVAQQQHEAELAAAASQPLPDDDDDAFE'
```

```
# 因為從0開始，位置要減1
Ub[3][2][42-1]
```

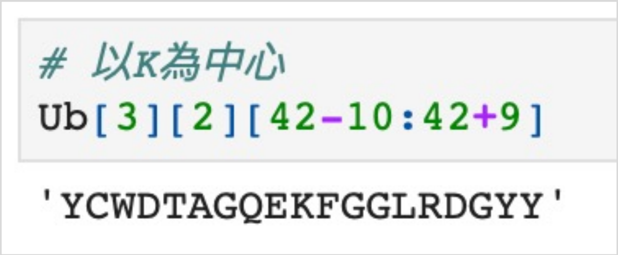
```
'K'
```

抓全部K位置 & 切出設定序列長度



```
for i in range(len(Ub[3][2])):
    if Ub[3][2][i]=='K':
        print(i+1)
```

```
2
29
42
70
94
98
103
105
112
113
123
130
138
```



```
# 以K為中心
Ub[3][2][42-10:42+9]
```

```
'YCWDTAGQEKFGGLRDGYY'
```

CD-HIT 安裝

```
%%bash

wget https://github.com/weizhongli/cdhit/releases/download/V4.8.1/cd-hit-v4.8.1-2019-0228.tar.gz

tar zxvf /content/cd-hit-v4.8.1-2019-0228.tar.gz

cd /content/cd-hit-v4.8.1-2019-0228

make
```

CD-HIT 指令

```
!./cd-hit -i 檔案 -o 存檔 -c 0.4 -n 2
```

-i：輸入檔案名稱。

-o：輸出名稱。

-c：相似度設定，例如0.9 = 90%，即相似度高於90%的序列會被當成同一群集。

-n：框架長度，不同的相似度有其建議的n值設定，例如90%時建議n=5。

```
Word size selection：
    -n 5 for threshold 0.7~1.0
    -n 4 for threshold 0.6~0.7
    -n 3 for threshold 0.5~0.6
    -n 2 for threshold 0.4~0.5
```

https://lsl.sinica.edu.tw/Blog/2017/07/cd-hit/