Characteristics of genome with extended alphabet

Biopython course project

Kolosov Nikita
Boris Shpak
Katerina Gibitova

2019

Task

- Given the assembled reads of unknown origin with unknown “nucleotides” extract some knowledge about what we are dealing with.

- Characterise sequence from standpoint of string sequence analysis.

After assembly: 3 “chromosomes”

1) 300454
2) 300454

1

3) 402937
4) 402937

2

5) 300001
6) 300001

3

Forward:

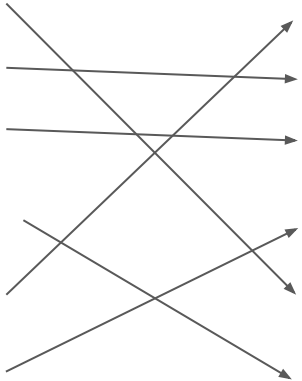
```
{'S': 0.1547,
 'T': 0.1778,
 'O': 0.1524,
 'I': 0.1611,
 'B': 0.1761,
 'N': 0.1778}
```

Reversed:

```
{'S': 0.1761,
 'T': 0.1778,
 'O': 0.1524,
 'I': 0.1778,
 'B': 0.1547,
 'N': 0.1611}
```

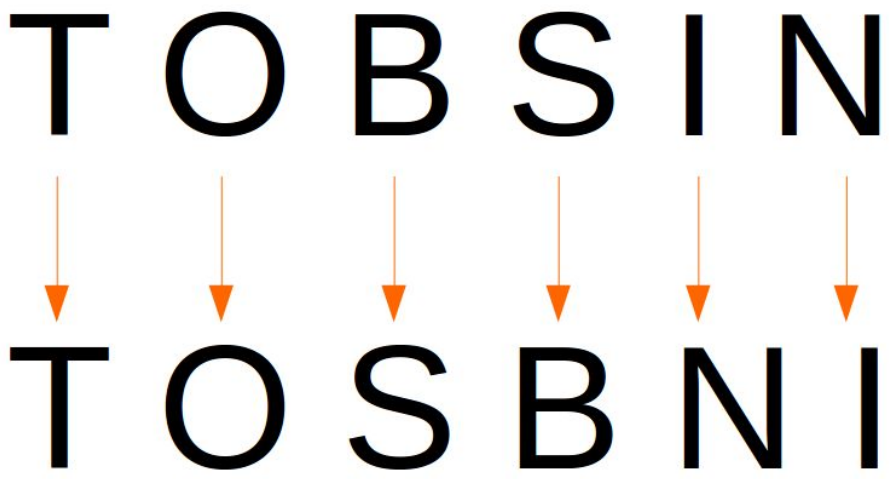
Forward:

```
{'S': 0.1547,
 'T': 0.1778,
 'O': 0.1524,
 'I': 0.1611,
 'B': 0.1761,
 'N': 0.1778}
```



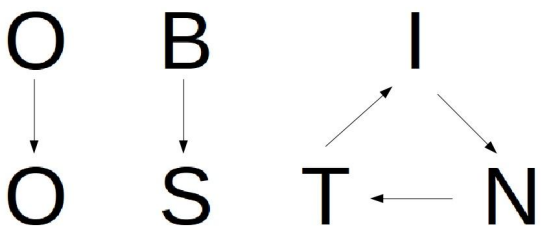
Reversed:

```
{'S': 0.1761,
 'T': 0.1778,
 'O': 0.1524,
 'I': 0.1778,
 'B': 0.1547,
 'N': 0.1611}
```

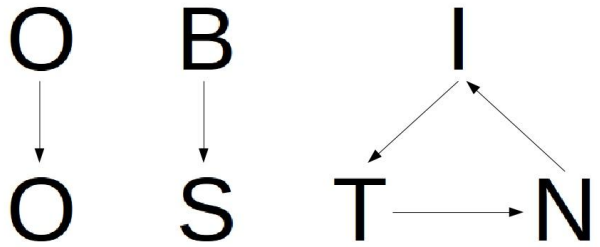

TOBSIN
TOSBNI

OR

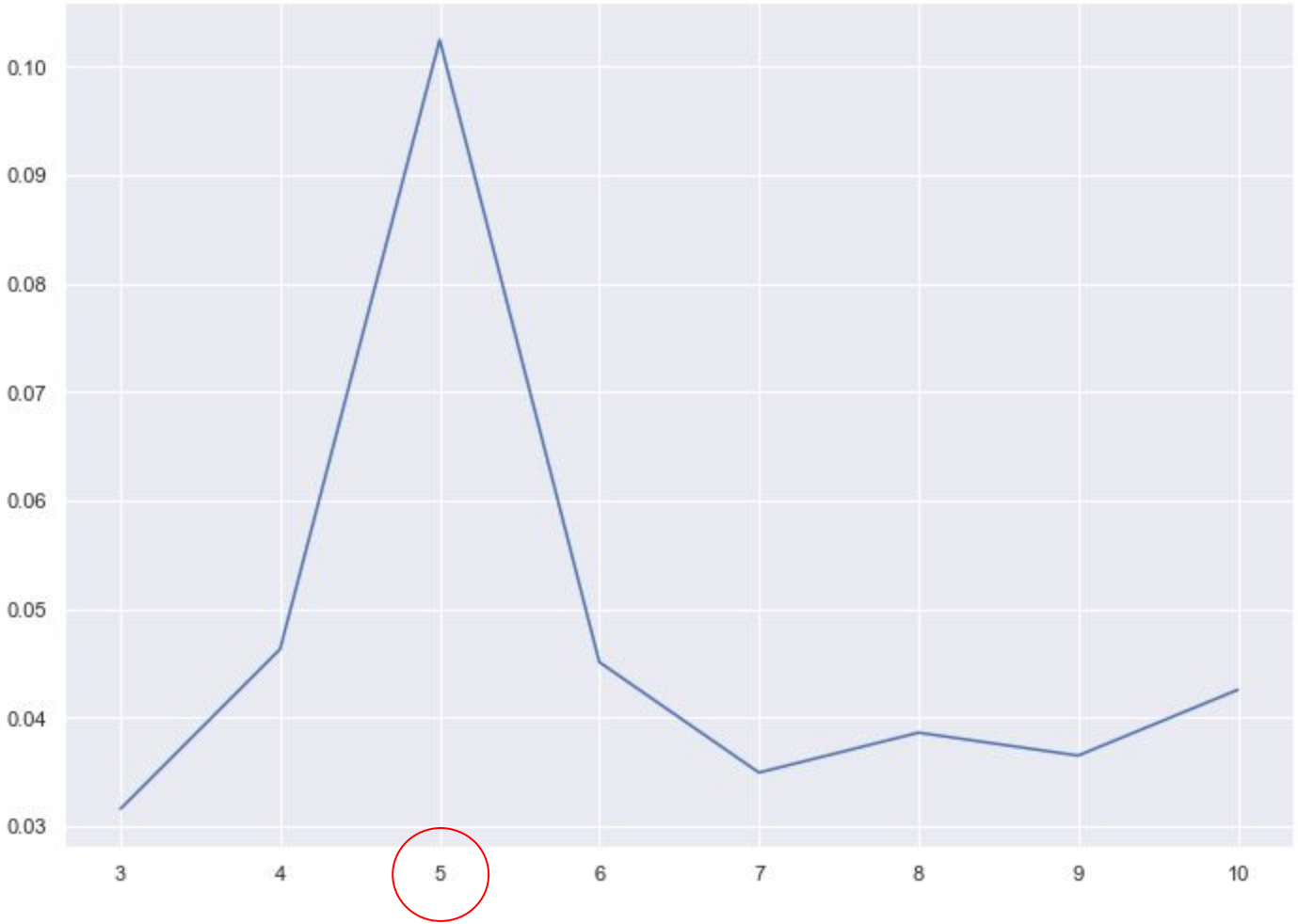
Forward



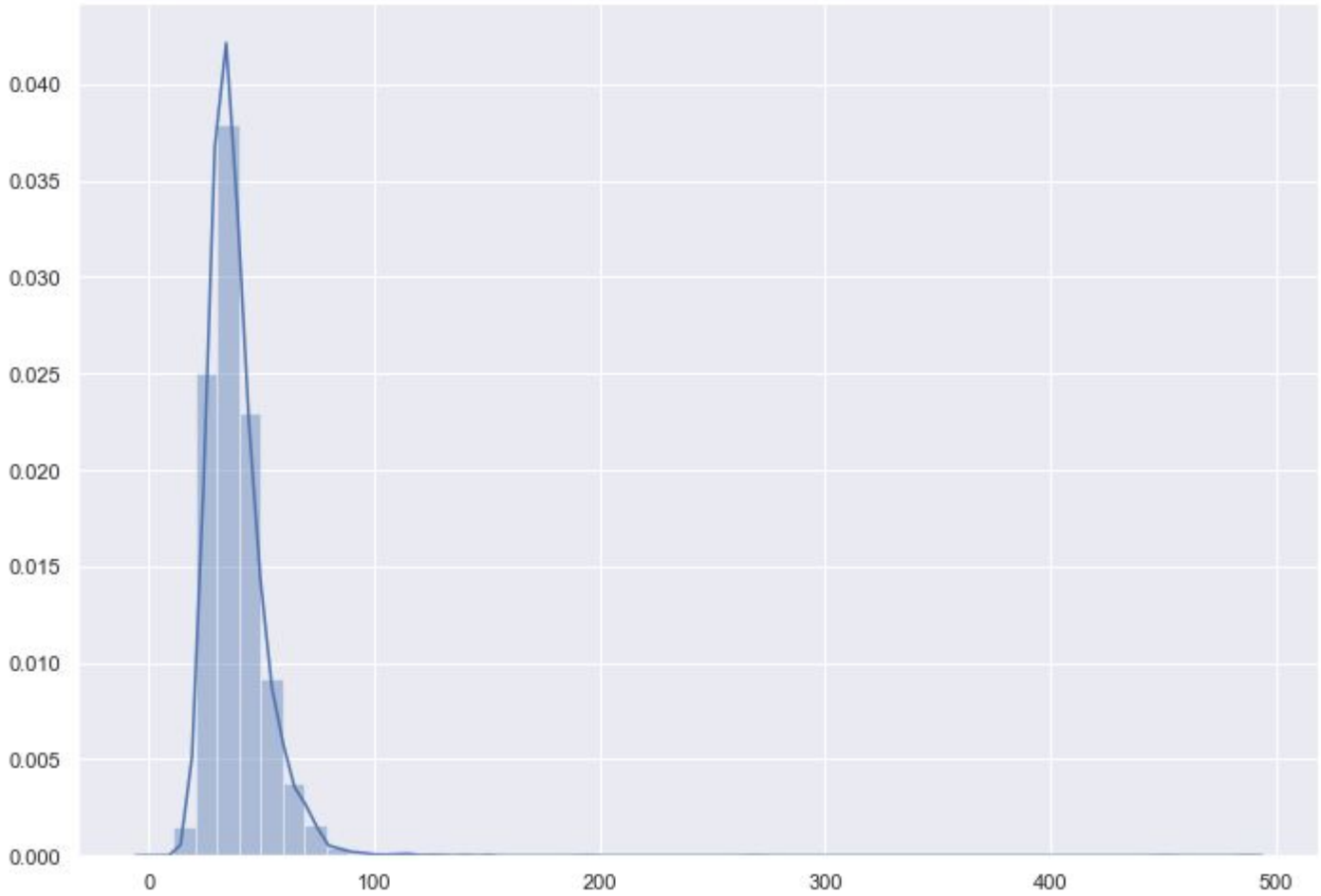
Reversed

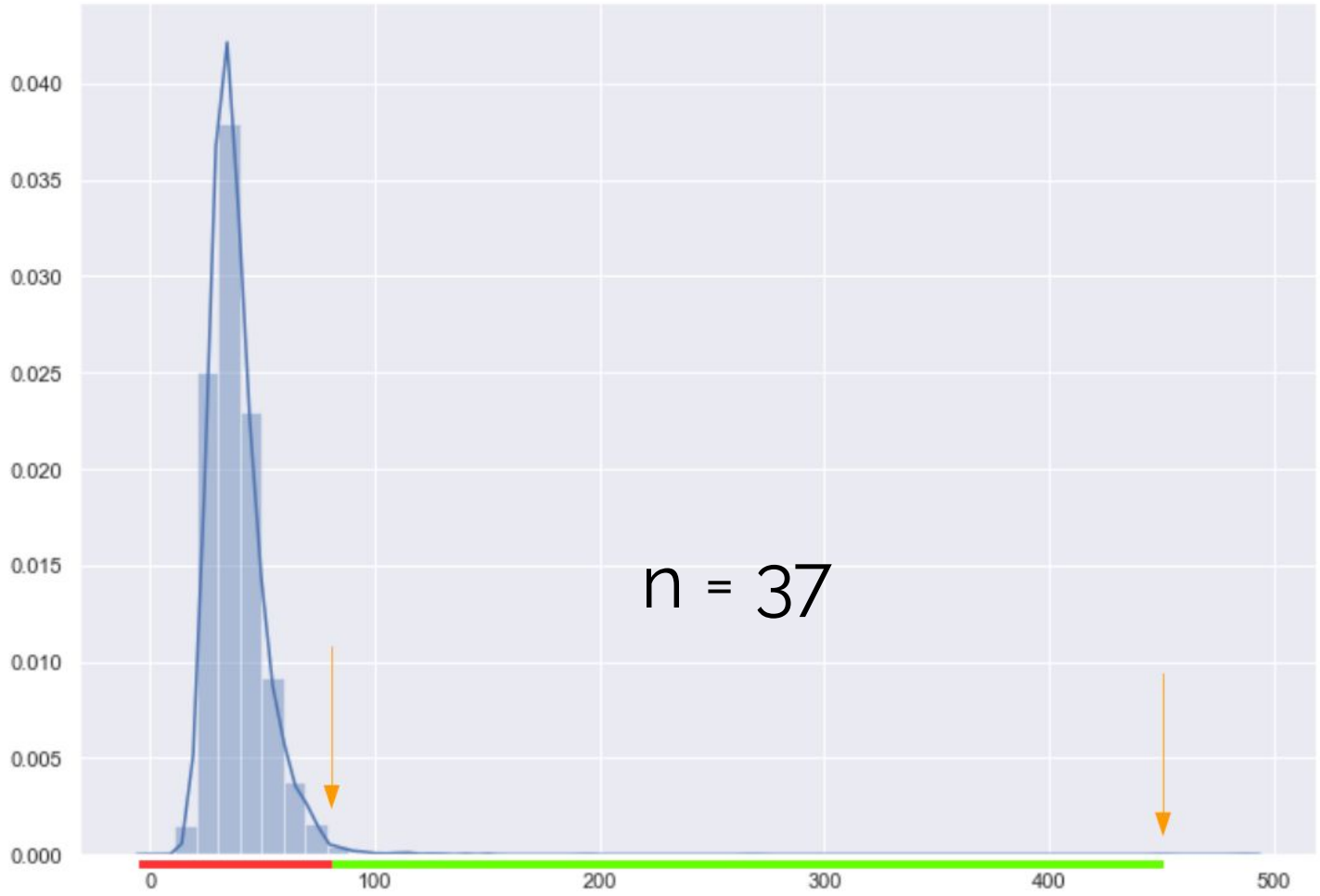


Chromosome **1**



k = 5

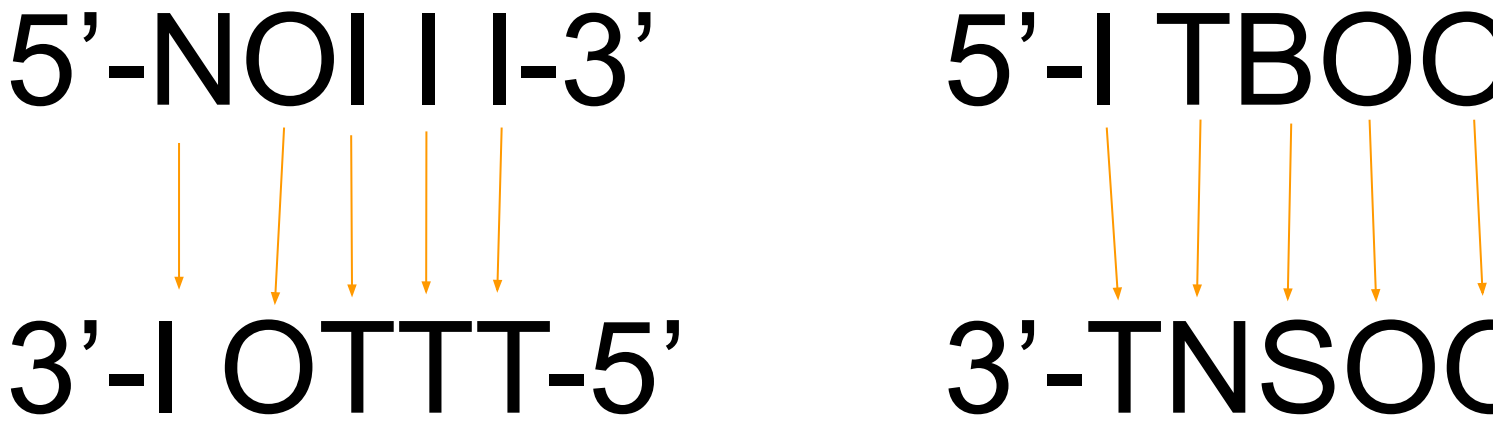

0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
500

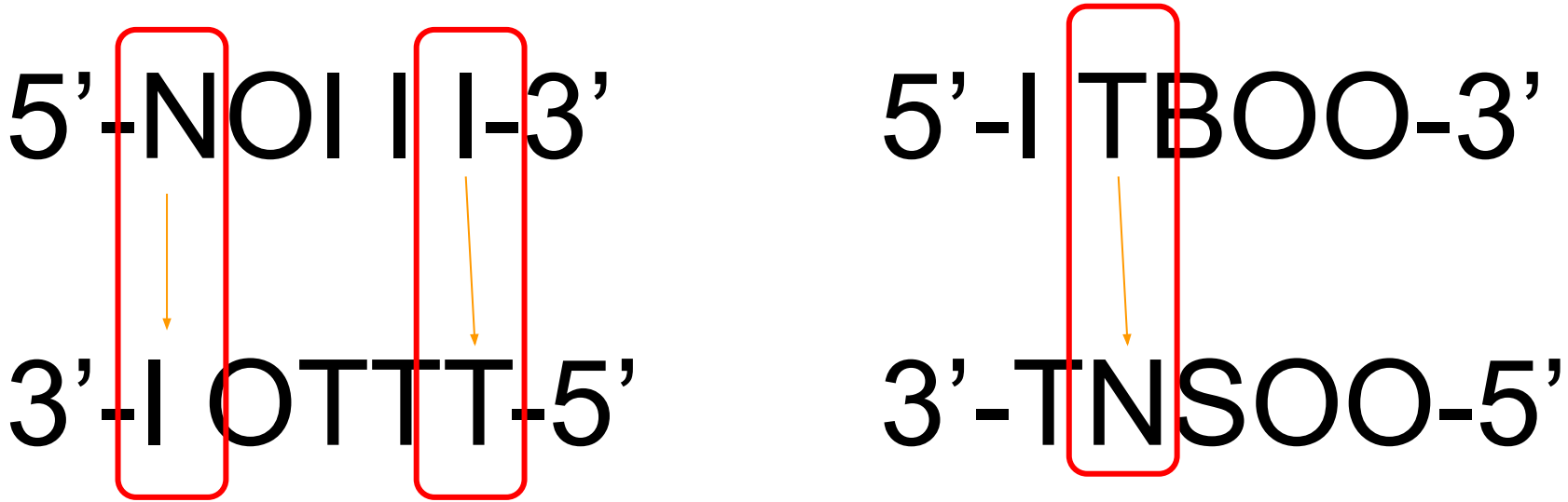



```
[('IIIOO', 88),
 ('SISII', 88),
 ('TOOSN', 88),
 ('SBNIN', 89),
 ('IIISI', 90),
 ('NBNTT', 91),
 ('OSNTT', 91),
 ('NINIO', 93),
 ('ISIII', 94),
 ('TBTTT', 95),
 ('OTTTO', 95),
 ('IIIIS', 95),
 ('ITTTO', 96),
 ('TTOIS', 99),
 ('IIIII', 99),
 ('OSNTN', 102),
 ('BNTTT', 106),
 ('STTTO', 108),
 ('NINIS', 108),
 ('NINIT', 112),
 ('TTOIT', 112),
 ('TTOIB', 113),
 ('IBNIN', 115),
 ('TTTTO', 115),
 ('TTOIN', 116),
 ('NINII', 121),
 ('NININ', 126),
 ('NINIB', 128),
 ('BTTTO', 129),
 ('BBNIN', 139),
 ('NBNIN', 140),
 ('NTTTO', 149),
 ('TBNIN', 151),
 ('TTOTS', 194),
 ('OOSNT', 269),
 ('TTTOI', 451)]
```

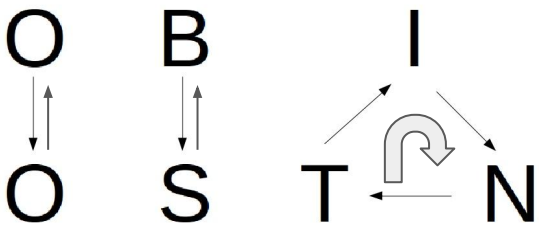
```
[('TBOOI', 88),
 ('OONNN', 88),
 ('NNBNB', 88),
 ('TNTSB', 89),
 ('NBNNN', 90),
 ('IITBO', 91),
 ('IITST', 91),
 ('ONTNT', 93),
 ('NNNBN', 94),
 ('OIIIO', 95),
 ('IIISI', 95),
 ('BNNNN', 95),
 ('OIIIN', 96),
 ('NNNNN', 99),
 ('BNOII', 99),
 ('TITBO', 102),
 ('IIITS', 106),
 ('BNTNT', 108),
 ('OIIIB', 108),
 ('INOII', 112),
 ('INTNT', 112),
 ('SNOII', 113),
 ('OIIII', 115),
 ('TNTSN', 115),
 ('TNOII', 116),
 ('NNTNT', 121),
 ('TNTNT', 126),
 ('SNTNT', 128),
 ('OIIIS', 129),
 ('TNTSS', 139),
 ('TNTST', 140),
 ('OIIIT', 149),
 ('TNTSI', 151),
 ('BIOII', 194),
 ('ITBOO', 269),
 ('NOIII', 451)]
```

```
[('IIIOO', 88),
 ('SISII', 88),
 ('TOOSN', 88),
 ('SBNIN', 89),
 ('IIISI', 90),
 ('NBNTT', 91),
 ('OSNTT', 91),
 ('NINIO', 93),
 ('ISIII', 94),
 ('TBTTT', 95),
 ('OTTTO', 95),
 ('IIIIS', 95),
 ('ITTTO', 96),
 ('TTOIS', 99),
 ('IIIII', 99),
 ('OSNTN', 102),
 ('BNTTT', 106),
 ('STTTO', 108),
 ('NINIS', 108),
 ('NINIT', 112),
 ('TTOIT', 112),
 ('TTOIB', 113),
 ('IBNIN', 115),
 ('TTTTO', 115),
 ('TTOIN', 116),
 ('NINII', 121),
 ('NININ', 126),
 ('NINIB', 128),
 ('BTTTO', 129),
 ('BBNIN', 139),
 ('NBNIN', 140),
 ('NTTTO', 149),
 ('TBNIN', 151),
 ('TTOTS', 194),
 ('OOSNT', 269),
 ('TTTOI', 451)]
```

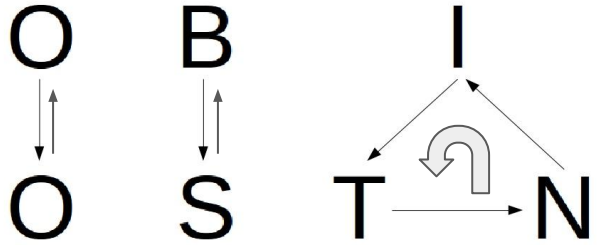

5’-NOI I I-3’
3’-I OTTT-5’
5’-I TBOO-3’
3’-TNSOO-5’


5’-NOI I I-3’
3’-I OTTT-5’
5’-I TBOO-3’
3’-TNSOO-5’

Forward

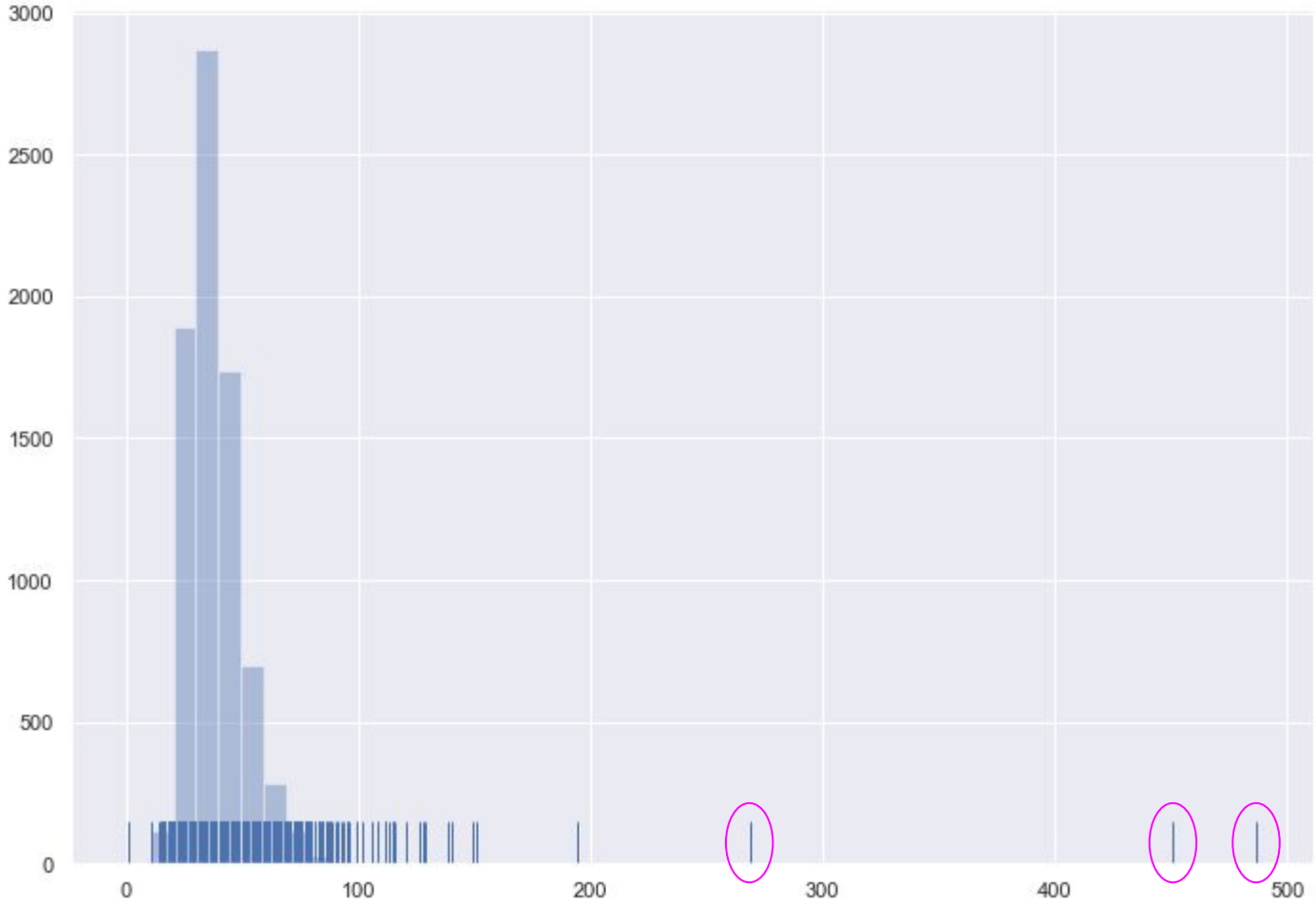


Reversed



Example: TTTOI BSNNITS

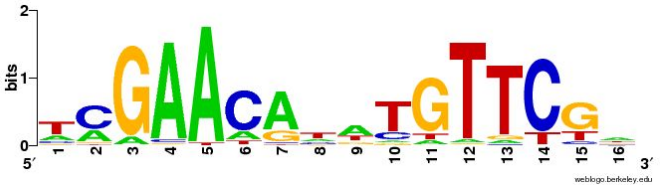
I I IONSBTTNIB

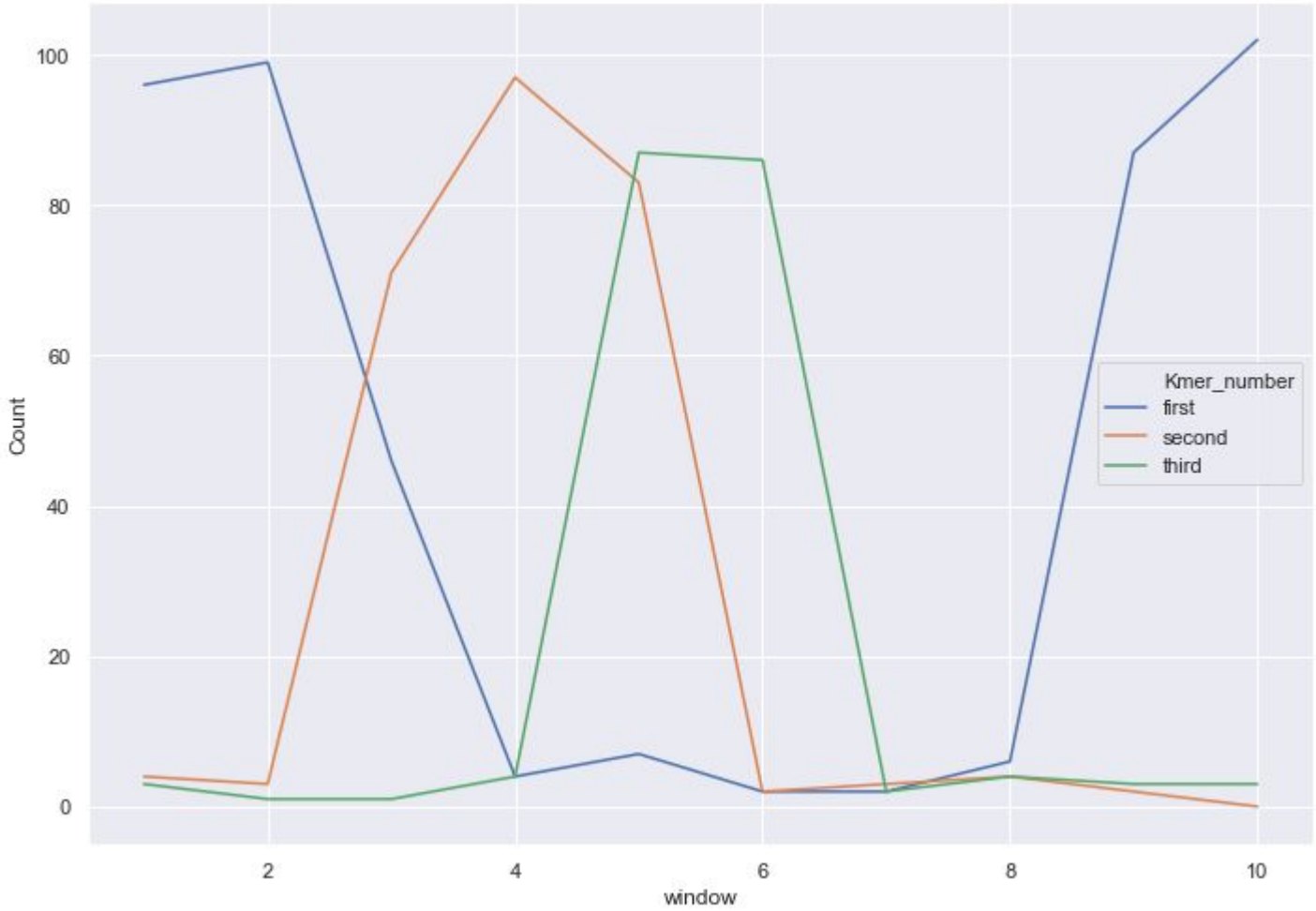



```
[('IIIOO', 88),
 ('SISII', 88),
 ('TOOSN', 88),
 ('SBNIN', 89),
 ('IIISI', 90),
 ('NBNTT', 91),
 ('OSNTT', 91),
 ('NINIO', 93),
 ('ISIII', 94),
 ('TBTTT', 95),
 ('OTTTO', 95),
 ('IIIIS', 95),
 ('ITTTO', 96),
 ('TTOIS', 99),
 ('IIIII', 99),
 ('OSNTN', 102),
 ('BNTTT', 106),
 ('STTTO', 108),
 ('NINIS', 108),
 ('NINIT', 112),
 ('TTOIT', 112),
 ('TTOIB', 113),
 ('IBNIN', 115),
 ('TTTTO', 115),
 ('TTOIN', 116),
 ('NINII', 121),
 ('NININ', 126),
 ('NINIB', 128),
 ('BTTTO', 129),
 ('BBNIN', 139),
 ('NBNIN', 140),
 ('NTTTO', 149),
 ('TBNIN', 151),
 ('TTOTS', 194),
 ('OOSNT', 269),
 ('TTTOI', 451)]
```

```
[('IIIOO', 9.710847946399747e-05),
 ('SISII', 0.00010006588585499018),
 ('IIISI', 0.00010420802794588949),
 ('IIIIS', 0.00010420802794588949),
 ('ISIII', 0.0001042080279458895),
 ('IIIII', 0.00010852163047962208),
 ('OOSNT', 0.0001136123317424768),
 ('TOOSN', 0.00011361233174247681),
 ('TTOIS', 0.0001200741112034278),
 ('NINIO', 0.00012510534214676386),
 ('NINIS', 0.0001269960754560493),
 ('OTTTO', 0.00013054908316174825),
 ('NINII', 0.00013225296980151914),
 ('STTTO', 0.00013252208843710274),
 ('TTOTS', 0.00013252208843710274),
 ('OSNTT', 0.0001325543370827858),
 ('OSNTN', 0.00013258659357602969),
 ('TTOIB', 0.0001367212190475519),
 ('ITTTO', 0.00013800772738187435),
 ('TTOIT', 0.00013800772738187435),
 ('TTTOI', 0.00013800772738187435),
 ('TTOIN', 0.0001380413109327703),
 ('SBNIN', 0.0001388550594462431),
 ('NINIB', 0.00014460284633037763),
 ('IBNIN', 0.00014460284633037766),
 ('NINIT', 0.00014596351856740675),
 ('NININ', 0.00014599903812379504),
 ('BTTTO', 0.00015089498727291785),
 ('TTTTO', 0.00015231486679188688),
 ('NTTTO', 0.00015235193191989204),
 ('BBNIN', 0.0001581059631267851),
 ('TBNIN', 0.00015959369590656626),
 ('NBNIN', 0.00015963253230443205),
 ('TBTTT', 0.00017605293985528564),
 ('BNTTT', 0.00017609578153508274),
 ('NBNTT', 0.00017613863364020712)]
```

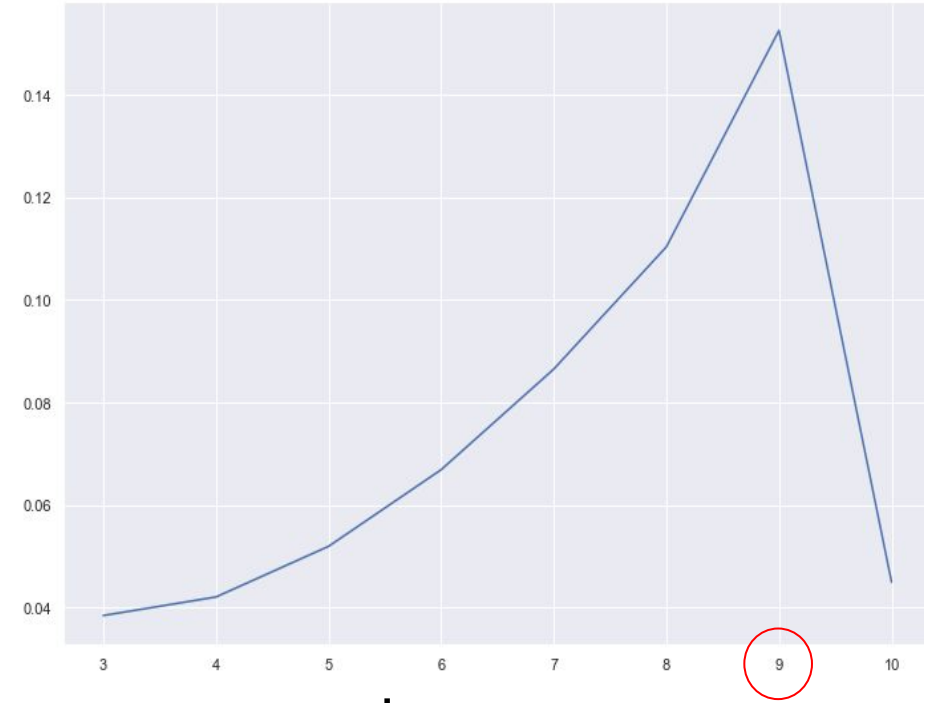
pLOGO for 'TTTOI':




100
80
60
40
20
0
Count
2
4
6
8
10
window
Kmer_number
first
second
third

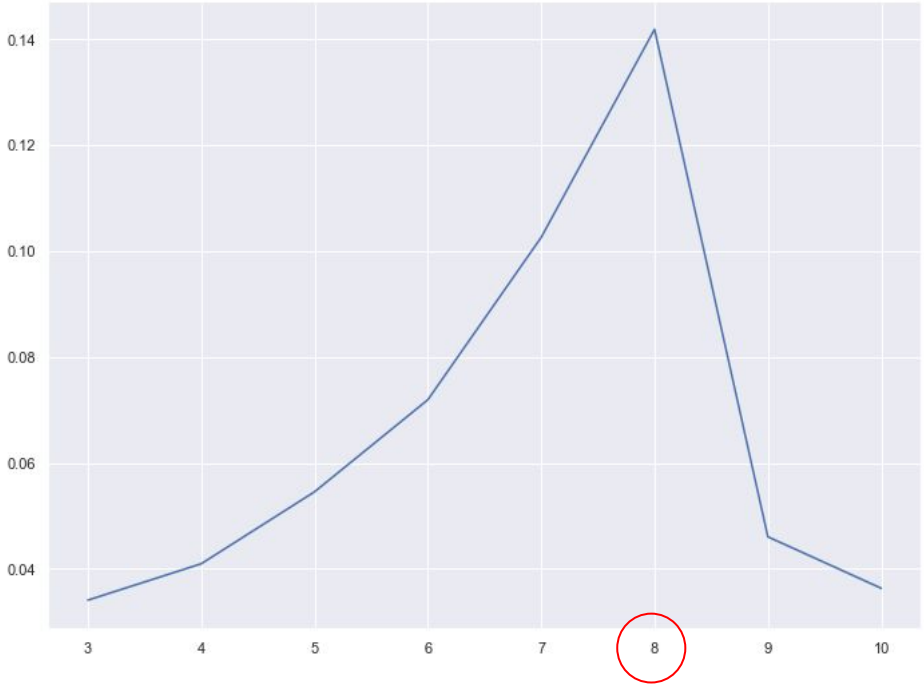
Chromosome **2** and **3**

2-nd chromosome



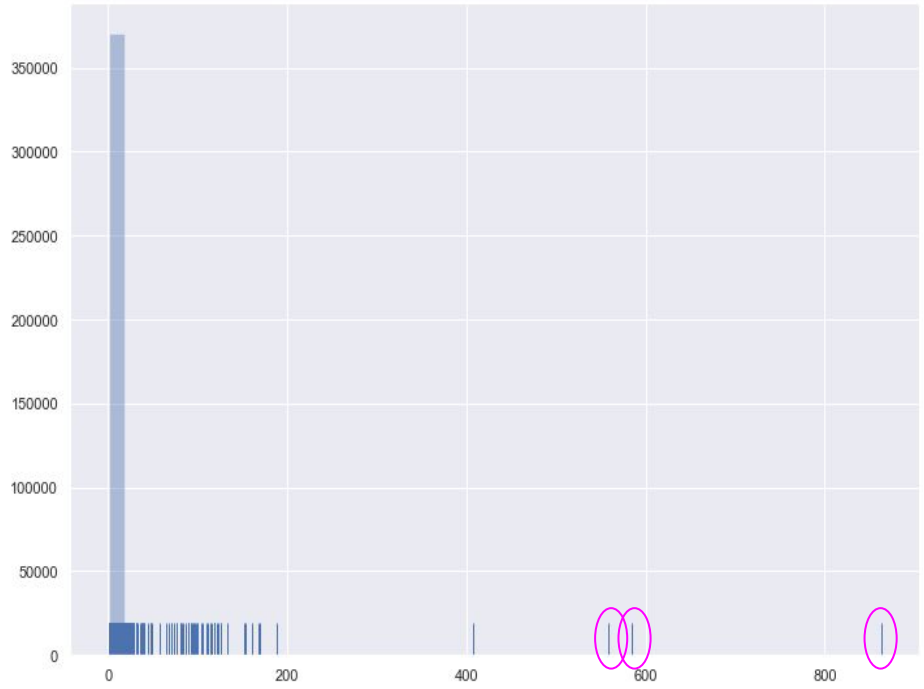
k = 9

3-d chromosome

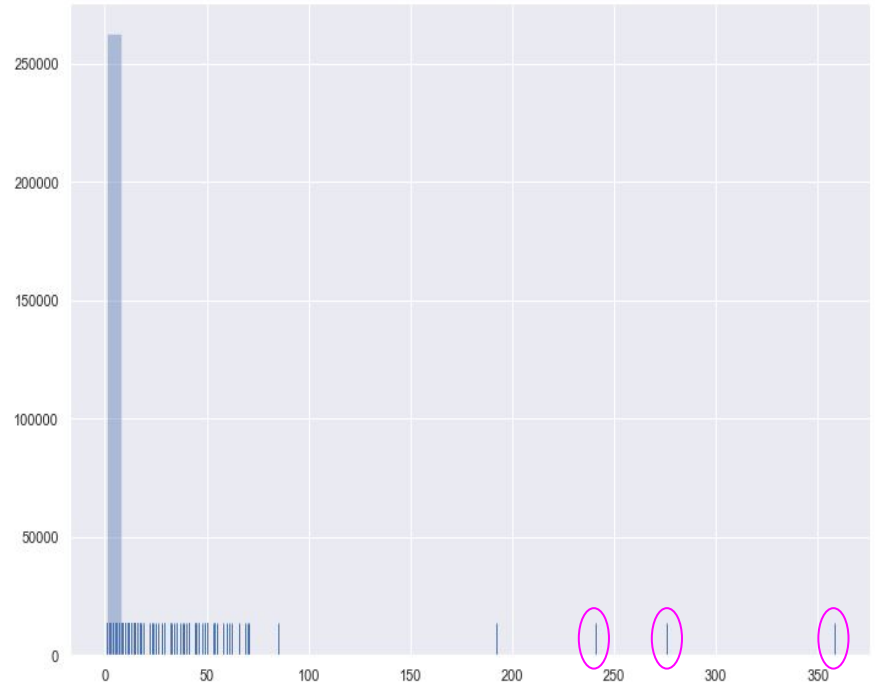


k = 8

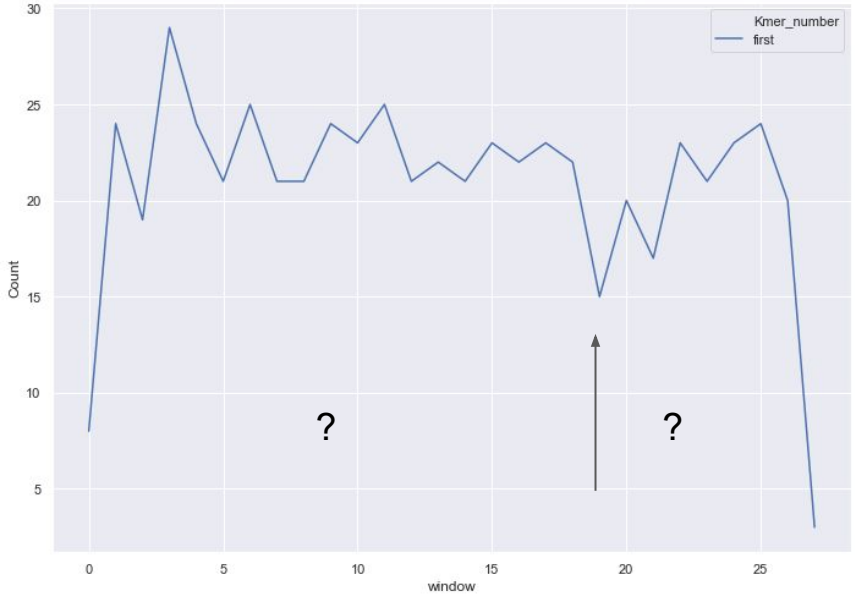
2-nd chromosome



3-d chromosome

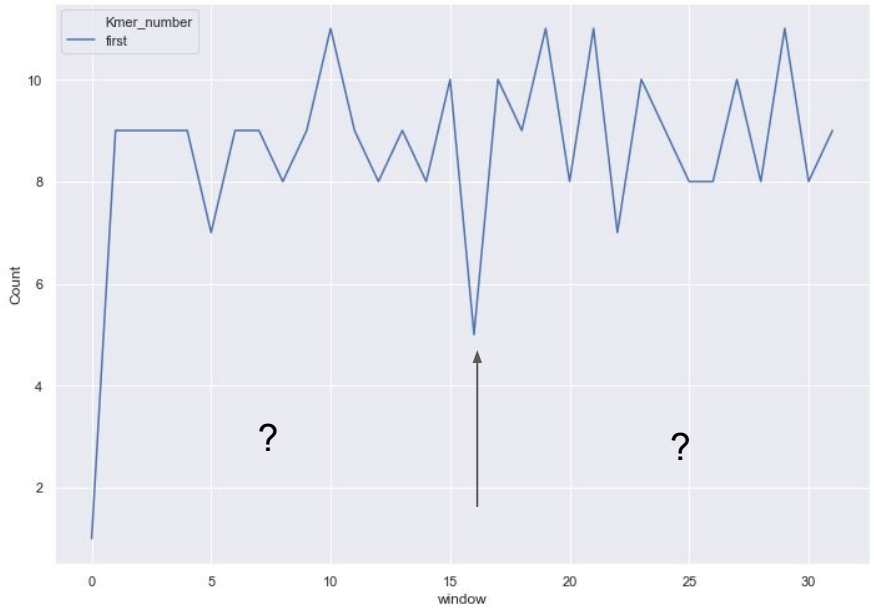


2-nd chromosome



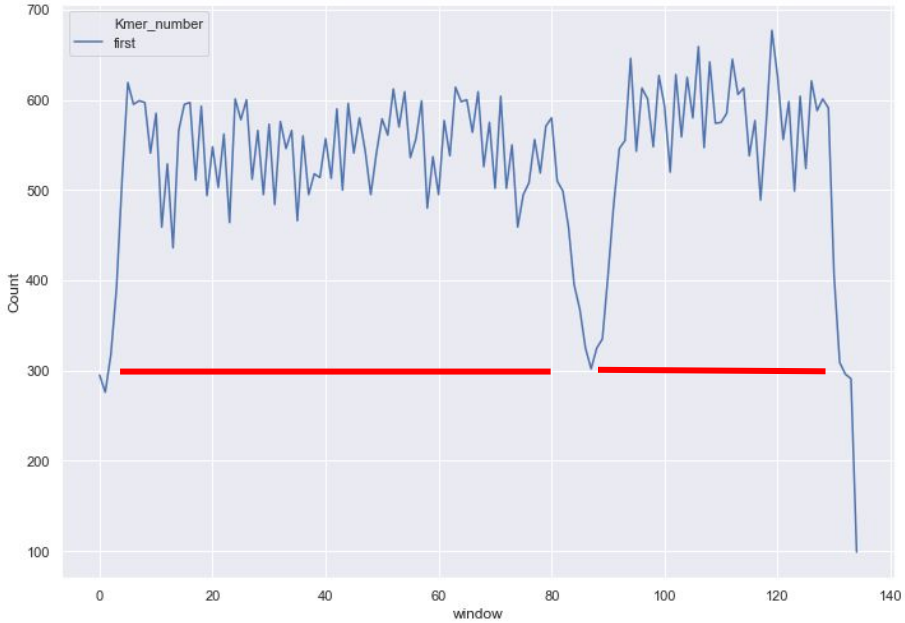
The most frequent kmer: 'OOOTIOOOI'

3-d chromosome



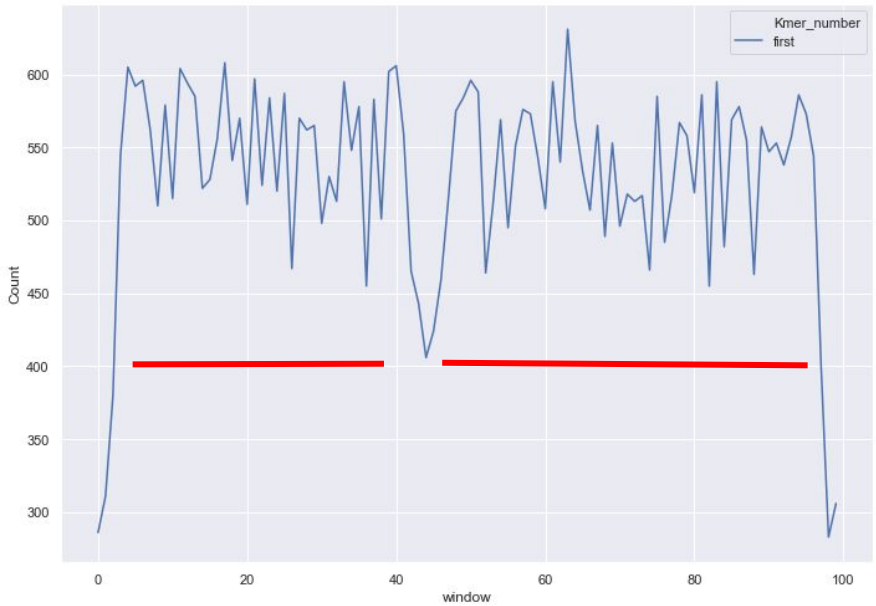
The most frequent kmer: 'OBITNNST'

2-nd chromosome



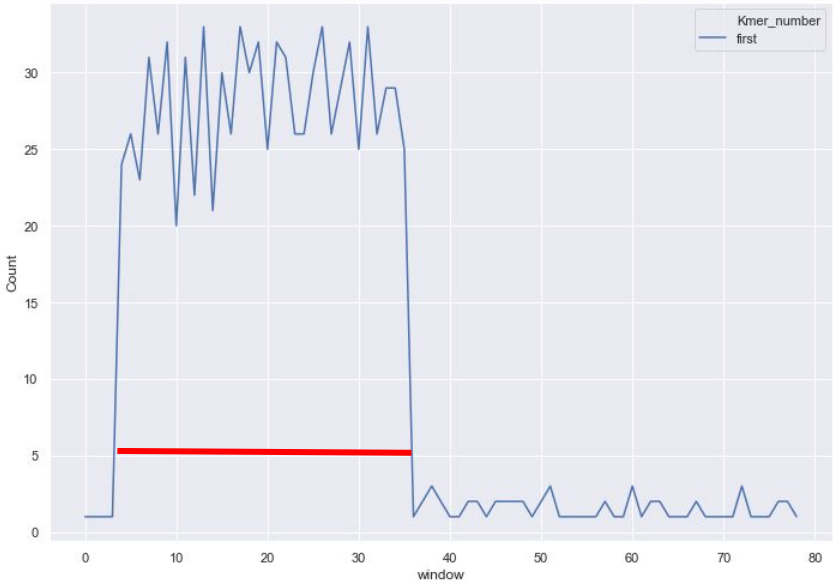
'TI'

3-d chromosome



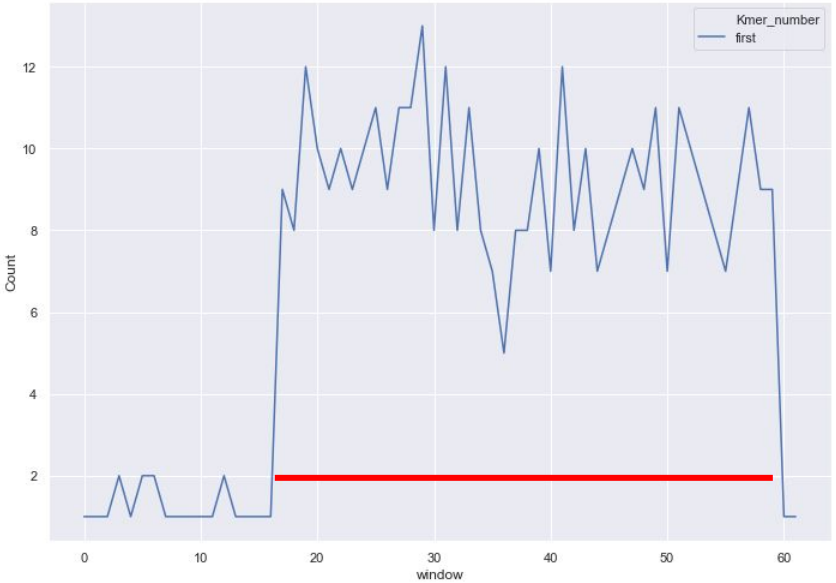
'T'

2-nd chromosome



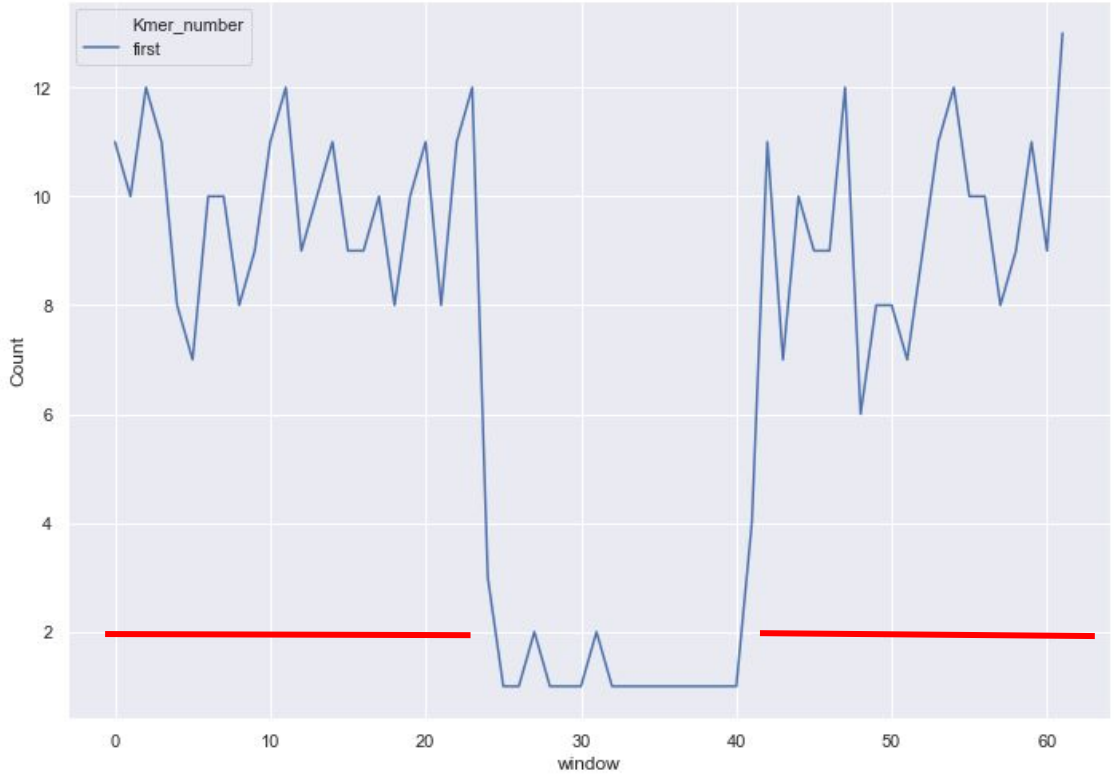
'TNNTN'

3-d chromosome



'NSTNS'

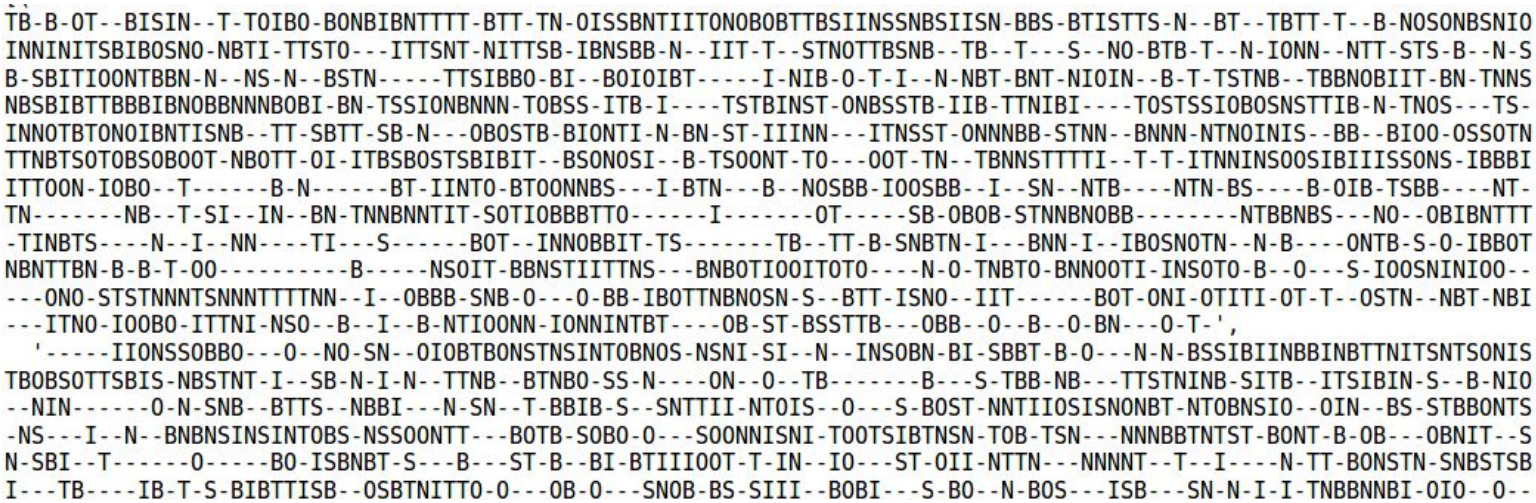
1-st chromosome



'TTTOI'

alignment?

- This fragments align better than random one # (527, 564)
- But I do not have stats...



diqqətinizə görə
təşəkkür edirik !