15-440
Distributed Systems
Kmeans

November 14, 2019

Zeinab Khalifa

Agenda

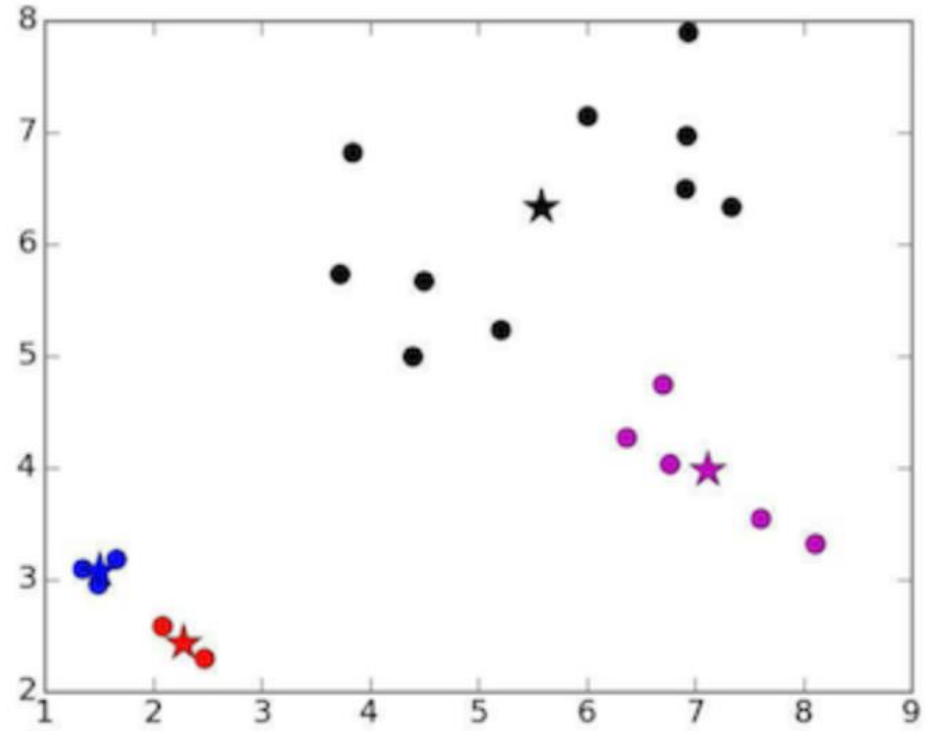
- Sequential Kmeans
- Parallelizing
- DNA clustering

Sequential Kmeans

Initial centroids/means

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9



The blue and red starts are called unlucky centroids (*)
A poor choice of the initial centroids will take longer to converge or may result in bad clustering. You can handle this in:

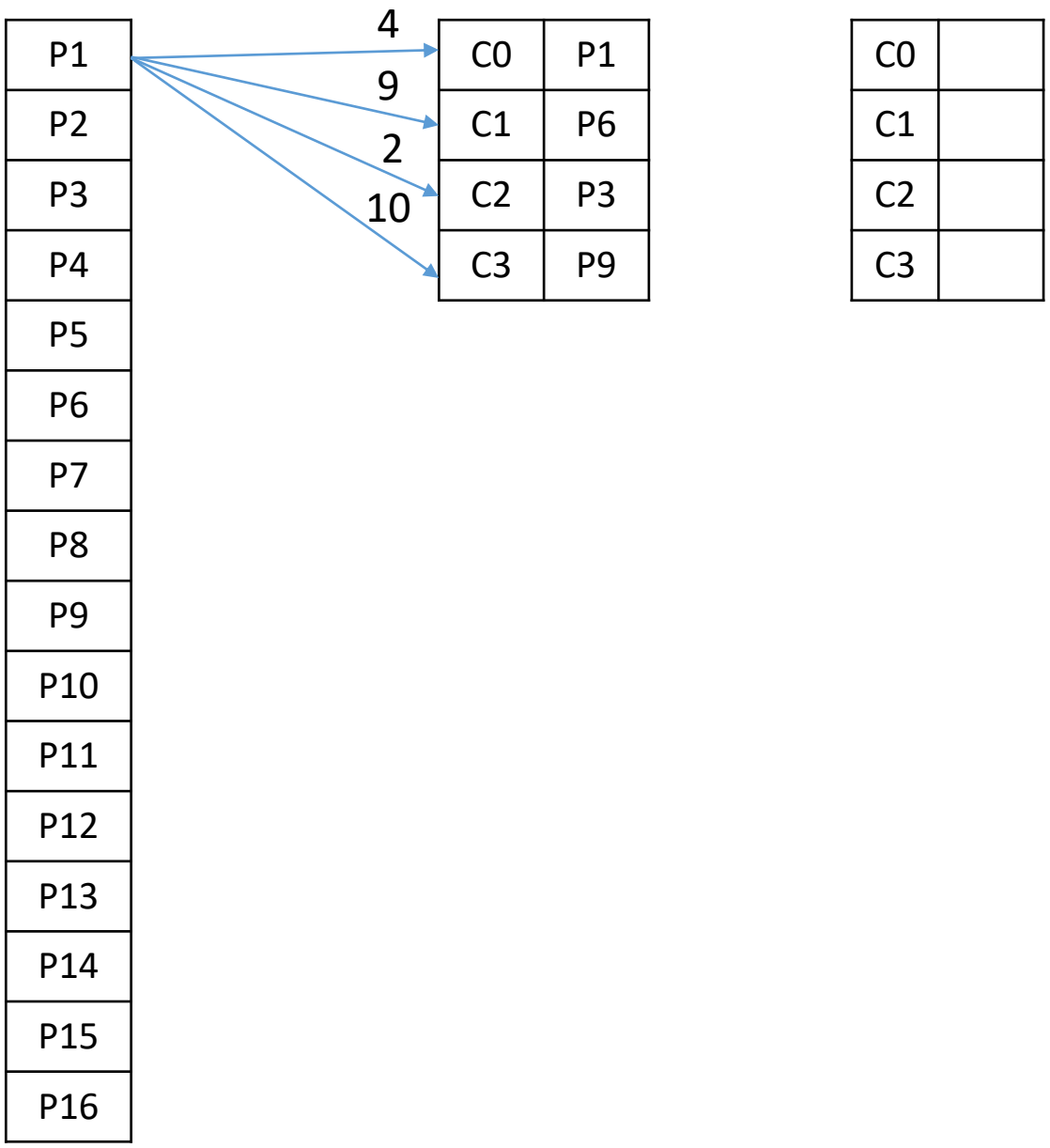
1. Your data generators (generate first k points to be far apart and pick them in your implementation)
2. Try different sets of random centroids, and choose the best set.

Initial centroids/means

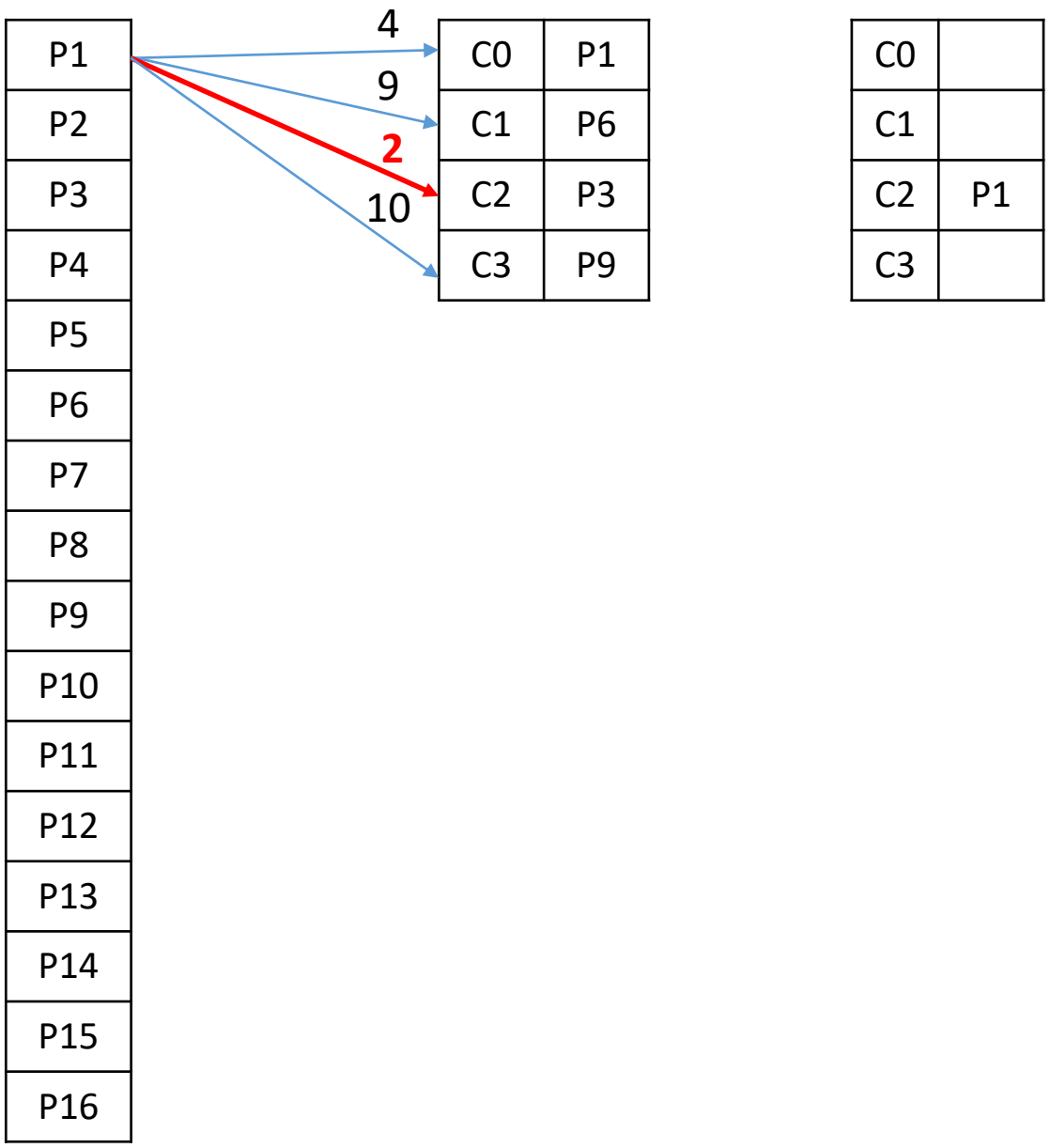
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

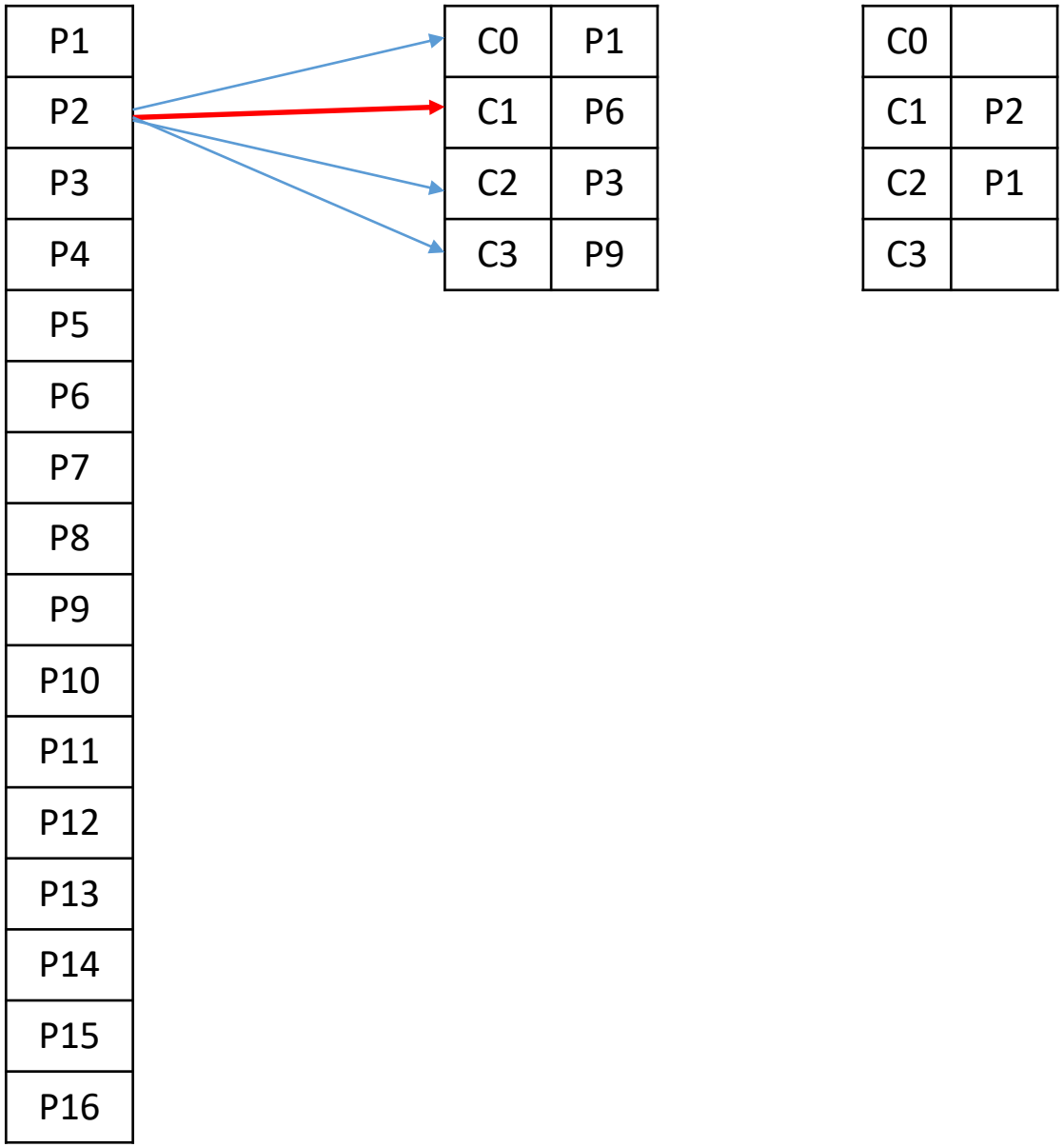
Initial centroids/means



Initial centroids/means



Initial centroids/means



Initial centroids/means

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

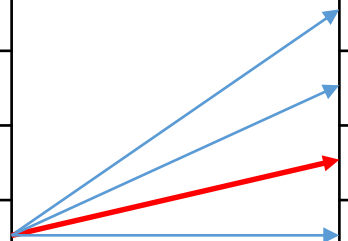
C0	P1
C1	P6
C2	P3
C3	P9

C0	
C1	P2 + P3
C2	P1
C3	

** P1 + P2 = (x1,y1) + (x2,y2) = (x1+x2, y1+y2)*

Initial centroids/means

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

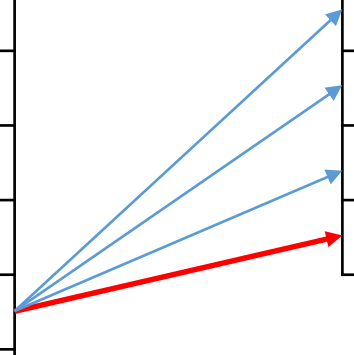


C0	P1
C1	P6
C2	P3
C3	P9

C0	
C1	P2 + P3
C2	P1 + P4
C3	

Initial centroids/means

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16



C0	P1
C1	P6
C2	P3
C3	P9

C0	
C1	P2 + P3
C2	P1 + P4
C3	P5

Initial centroids/means

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

C0	P6 + P8 + P10 + P13
C1	P2 + P3 + P7 + P11
C2	P1 + P4 + P12 + P15 + P16
C3	P5 + P9 + P14

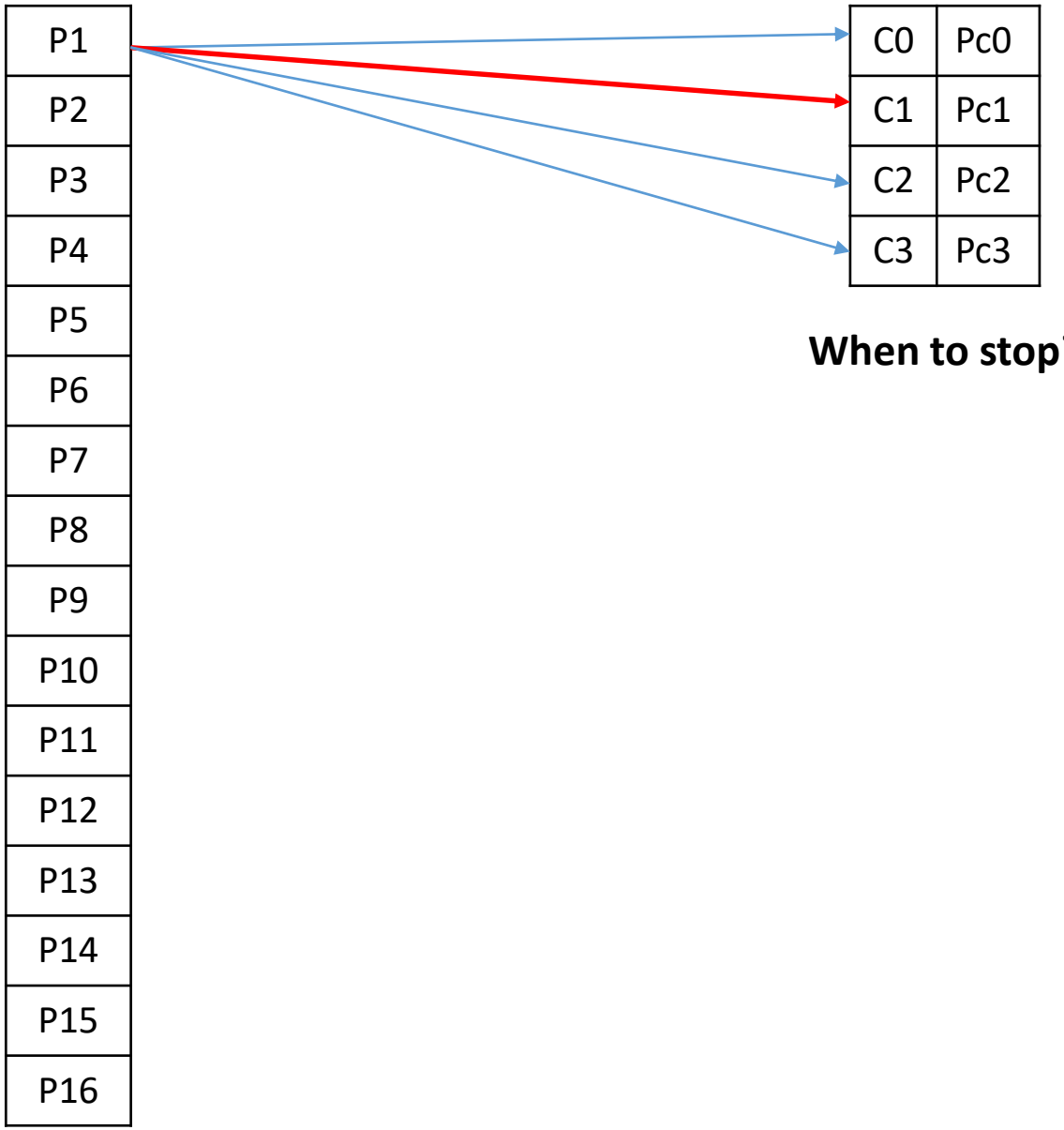
Centroids after iteration 1

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

C0	**(**P6 + P8 + P10 + P13**)/4**
C1	**(**P2 + P3 + P7 + P11**)/4**
C2	**(**P1 + P4 + P12 + P15 + P16**)/5**
C3	**(**P5 + P9 + P14**)/3**

** P/N = (x/N,y/N)*


P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16
C0
Pc0
C1
Pc1
C2
Pc2
C3
Pc3

When to stop?

Parallel K-Means

How can we parallelize?

P1
P2
P3
P4

C0	P1
C1	P6
C2	P3
C3	P9

C0	P2 + P3
C1	0
C2	P1
C3	P4

P5
P6
P7
P8

C0	P1
C1	P6
C2	P3
C3	P9

C0	P5
C1	P7
C2	P8
C3	P6

P9
P10
P11
P12

C0	P1
C1	P6
C2	P3
C3	P9

C0	0
C1	P12
C2	P10 + P11
C3	P9

P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

C0	P13 + P14 + P16
C1	0
C2	0
C3	P15

How can we parallelize?

P1
P2
P3
P4

C0	P1
C1	P6
C2	P3
C3	P9

C0	P2 + P3
C1	0
C2	P1
C3	P4

C0	P5
C1	P7
C2	P8
C3	P6

C0	0
C1	P12
C2	P10 + P11
C3	P9

C0	P13 + P14 + P16
C1	0
C2	0
C3	P15

P5
P6
P7
P8

C0	P1
C1	P6
C2	P3
C3	P9

P9
P10
P11
P12

C0	P1
C1	P6
C2	P3
C3	P9

P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

How can we parallelize?

P1
P2
P3
P4

C0	P1
C1	P6
C2	P3
C3	P9

C0	P2 + P3 + P5 + P13 + P14 + P16
C1	P7 + P12
C2	P8 + P10 + P11
C3	P6 + P9 + P15

/6
/2
/3
/3

P5
P6
P7
P8

C0	P1
C1	P6
C2	P3
C3	P9

P9
P10
P11
P12

C0	P1
C1	P6
C2	P3
C3	P9

P13
P14
P15
P16

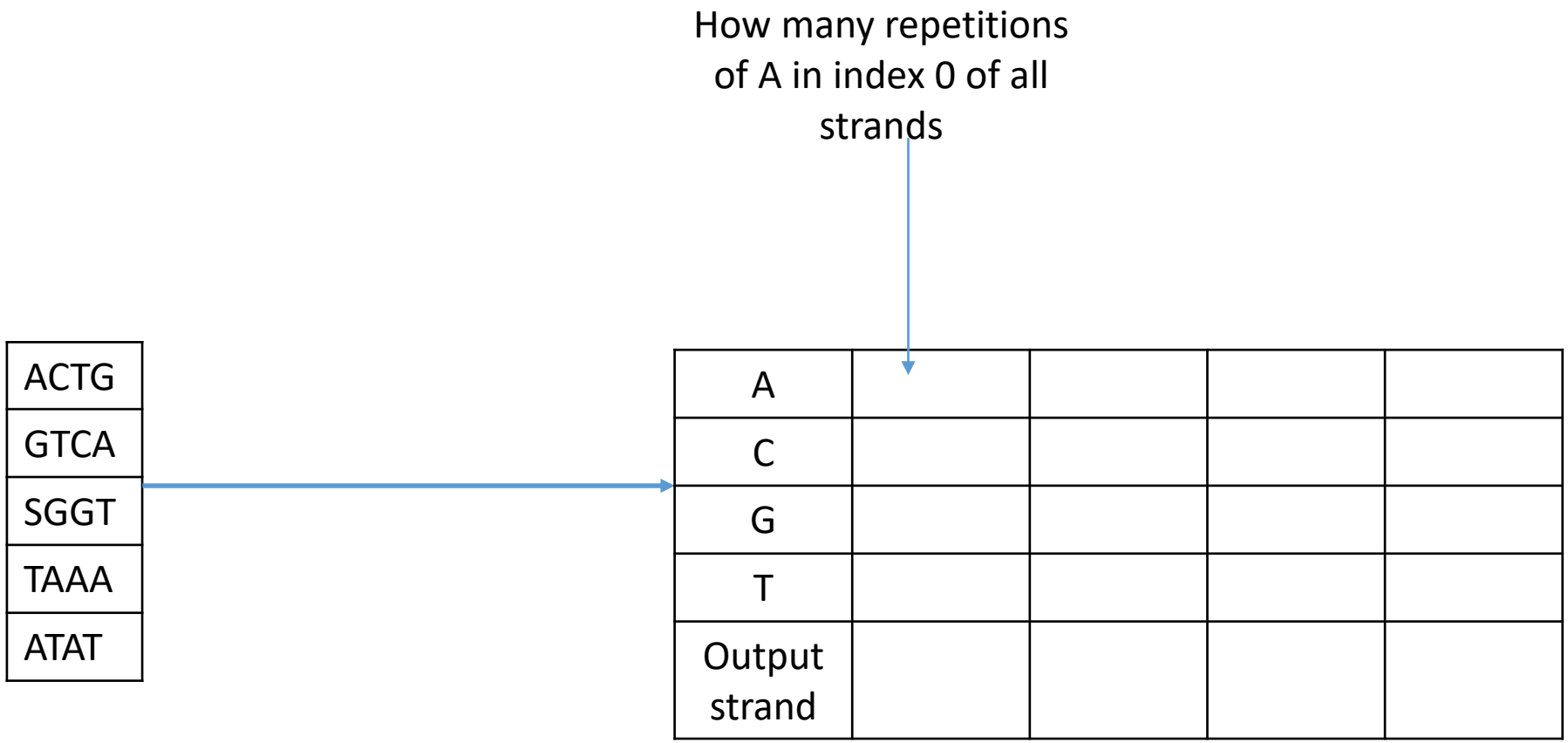
C0	P1
C1	P6
C2	P3
C3	P9

DNA stranding

ACTG
GTCA
SGGT
TAAA
ATAT

ACTG
GTCA
SGGT
TAAA
ATAT

How to get the centroid of these DNA strands?


How many repetitions of A in index 0 of all strands
ACTG
GTCA
SGGT
TAAA
ATAT
A
C
G
T
Output strand

ACTG
GTCA
SGGT
TAAA
ATAT

A	**2**			
C				
G				
T				
Output strand				

ACTG
GTCA
SGGT
TAAA
ATAT

A	2	1	2	1
C	0	1	1	0
G	1	1	1	1
T	1	2	1	2
Output strand				

ACTG
GTCA
SGGT
TAAA
ATAT

A	2	1	2	1
C	0	1	1	0
G	1	1	1	1
T	1	2	1	2
Output strand				

Get the mean or the median (sort the values and select the middle one)

ACTG
GTCA
SGGT
TAAA
ATAT

A	2	1	2	1
C	0	1	1	0
G	1	1	1	1
T	1	2	1	2
Output strand	**T**	**G**	**C**	**A**

ACTG
GTCA
SGGT
TAAA
ATAT

A	2	1	2	1
C	0	1	1	0
G	1	1	1	1
T	1	2	1	2
Output strand	**T**	**G**	**C**	**A**