Spring
2019

Computer Simulation of PCR

Introduction to Bioinformatics – Project 1

BINI ELSA PAUL

Abstract

Polymerase chain reaction (PCR) is a method widely used in molecular biology to make many copies of a specific DNA segment[1]. In this project "Computer Simulation of PCR" the three steps of the PCR (Denaturation, Annealing and Extension) is simulated with Python and for 30 cycles. We have also considered few limitations like the number of primers available, the length of the initial DNA, number of cycles and age of the taq polymerase. The results are shown in graph and tabular form to get a clear idea of fragment length and the numbers after the PCR. Even though the length of the initial DNA has not much effect in the PCR cycle, the number of cycles, the amount of the primers present and the age of the taq polymerase can affect the number of fragments and its length after the PCR cycle.

1.0 Introduction

The purpose of a PCR (Polymerase Chain Reaction) is to make a huge number of copies of a gene[2]. There are three major steps in PCR as shown in the figure below,

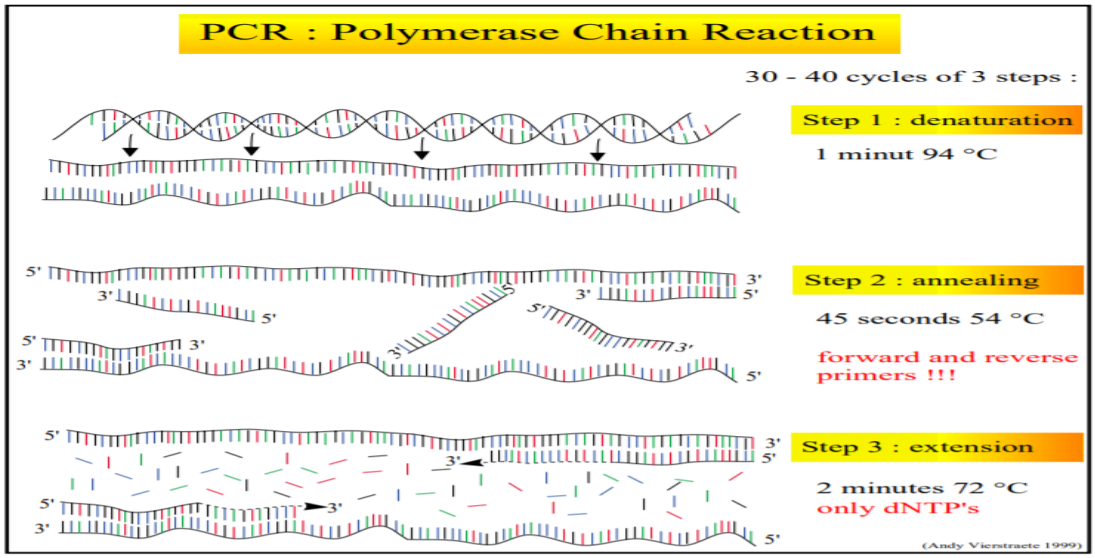

Figure 1: PCR steps

a) Denaturation

During the denaturation, the double strand melts open to single stranded DNA. In our program, in this step, we are generating the random DNA strand of length 2000. Its complimentary strand also stored.

To find the primers, we have spotted a random location in the initial DNA segment (3' -5') and selected 20 bases from that index and complemented it. Our primer has the length of 20 and we have verified that the primer is not present anywhere else in the segment. To get the backward primer (5'-3') we have select the primer in the complementary strand in such a way that the new segment after the extension will align with the first primer. So we complemented the 20 bases from 180 index after the start of the forward primer on the other strand of the DNA. The below figure shows the two strands of the DNA and its primers.

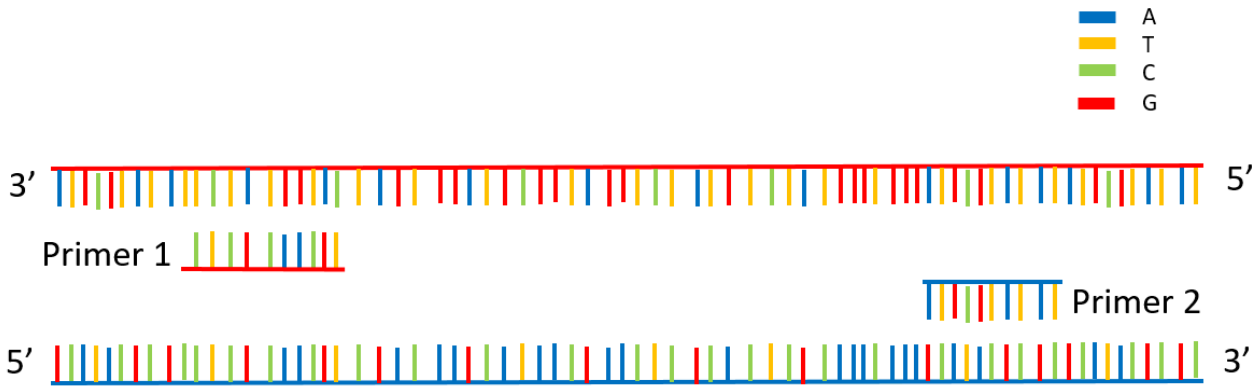

Figure 2: Two strands of the DNA and the primers

b) Annealing

In this step, the primers are binding to the DNA fragments. The primers will be binding to the complementary bases in the DNA segment. If the whole primer (20 nucleotides) is not attaching, then we have set a threshold of 10 nucleotides. That means at least 10 of the bases in the primer is attaching to the DNA segment, then the Extension step is done, and otherwise we are not using this segment for making the copy.

The below figure shows the DNA and its segments after the first cycle.

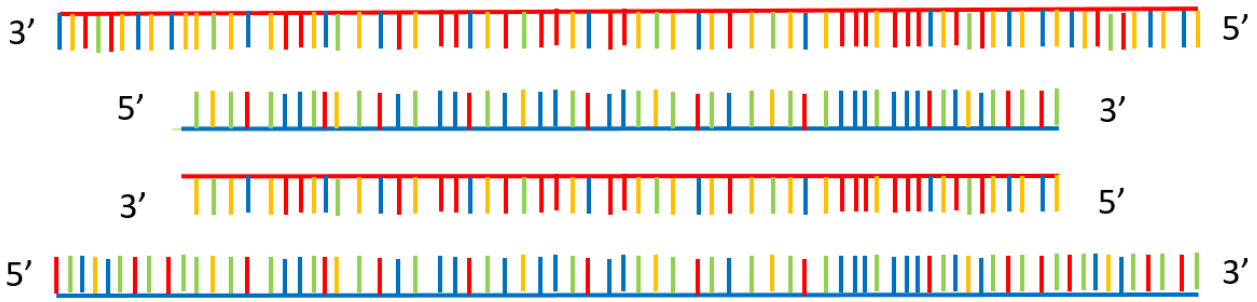

Figure3: DNA and its segments after the first cycle

c) Extension

In this step, first calculates the value for taq polymerase (“fall-off rate”) which is d+r, where d is a fixed constant, and r is a random number between [-e, e], (assumed d = 200, e=50). The complement of the bases starting from the location where the primer is attached till d+r is the new fragment which is stored in the appropriate list (there are two lists, 3’-5’ list and 5’-3’ list). The new fragment from the 3’-5’ segment will have the direction 5’-3’ and vice versa.

These steps are repeated for 30 cycles. The below figure shows the Annealing step in the cycle 2. Primer 1 will attach to its complementary bases in the 3’-5’ DNA. Primer 2 will attach to its complementary bases in the 5’-3’ DNA.

Figure 5 shows the extension step in cycle 2. The new segment to 3’-5’ will have direction 5’-3’ and vice versa.

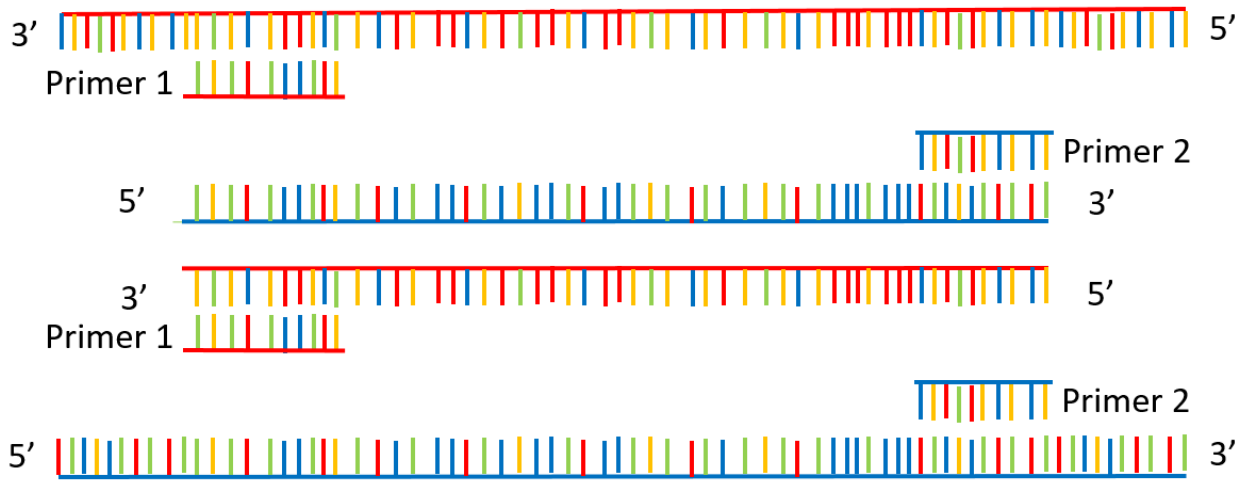

Figure 4: Annealing in the cycle 2

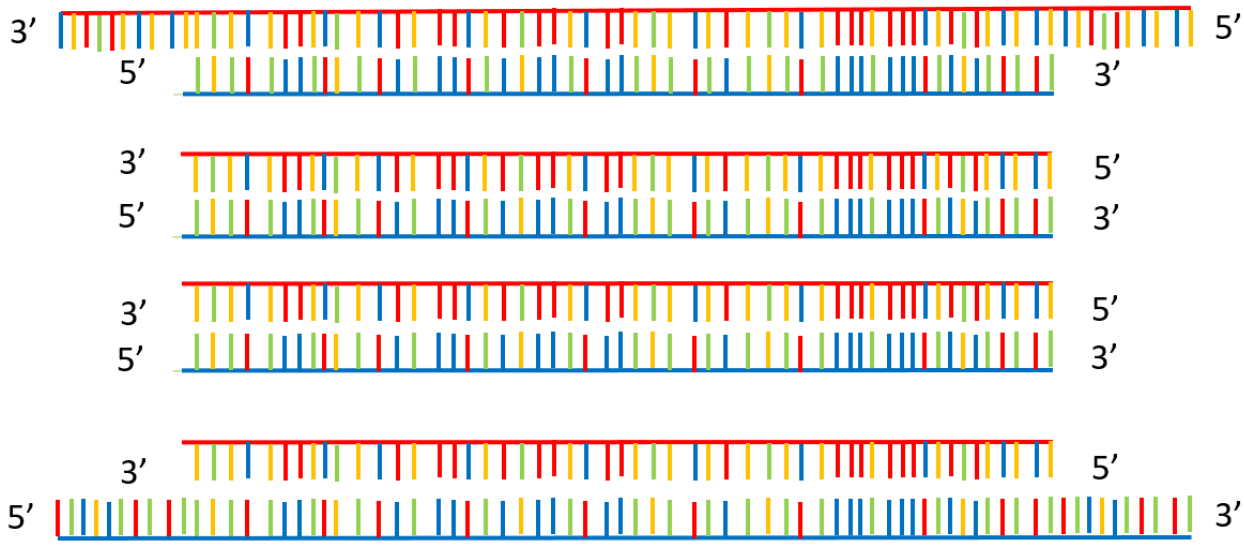

Figure 5: Extension step in the cycle 2

2.0 Implementation and Algorithm

i) Data Structures, Language and Packages used

- We have used two Lists() to store the new segments. One list to store all the 3'-5' segment and other for 5'-3' segment.
- Used Python, PyCharm (IDE) and Github (for simultaneous/ shared coding)
- Random generators to generate random numbers and random DNA.
- Prettytable and matplotlib.pyplot for showing the results.

ii) Algorithm

List1 – stores the 3'-5' segments

List2 – stores the 5'-3' segments

(1) Denaturation

(a) Generate random DNA of length 2000 (say the direction 3'-5')

(b) Take its complement to get 5'-3'(A ⇔ T, G ⇔ C)

(c) Get the primer for 3'-5' DNA segment (say forward primer)

(i) Select the random position in the 3'-5' DNA segment

(ii) Select 20 bases from that position and complement it which is the forward primer
(iii) Check whether the primer is present anywhere else in the segment, if yes, repeat (i), (ii),(iii) till the primers is unique.

(d) Get the primer for 5'-3' DNA segment (say backward primer)
(i) Starting pos = staring index of forward primer + 180
(ii) Select 20 bases from staring pos in the 5'-3' strand and complement it which is the backward primer

(2) Annealing
(a) Iterate through List1 and check whether the forward primer is binding (the complement of the primer should be an exact match to the DNA segment).
(b) If yes, that segment goes to the Extension phase.
(c) Otherwise discard the starting base and repeat (a),(b)and (c) 10 times
(d) If the primer is not binding, this segment will not consider for Extension.

(3) Extension
(a) Find the taq polymerase fall – off rate, (d+r, d= 200 and r is random number between -50,50)
(b) Get the new segment which is the complement of the substring of the DNA segment starting from primer binding position and length d+r
(c) The new segment from 3'-5' will have direction from 5'-3' and will go to the List2 and vice versa.

(4) Repeat (2) and (3) for 30 cycles.

3.0 Result

The result without any constraints and 30 iterations are as shown. The total number is less than the possible number of DNA fragments since the taq falls off before copying is done.
Figure 6 shows the average length of the DNA fragment in each cycle. Figure 8 shows the distribution of number of fragments. The length of the original DNA is not considered to find the distribution and average length.

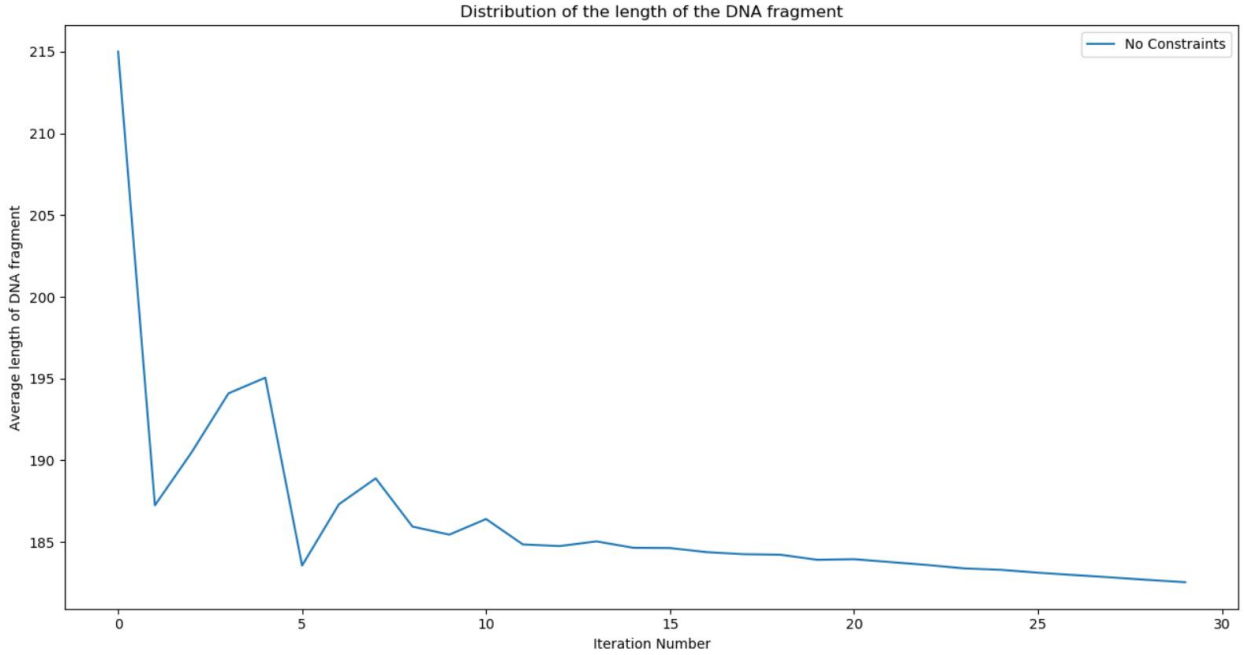

Figure 6: Graph with Distribution of the length for 30 iterations

```
Number of DNA Fragments =  10385325

Average Length of DNA Fragments = 182.77317320353478
```

Iteration	Number of Fragments	Avg Length
0	4	215.0
1	8	187.25
2	14	190.5
3	24	194.1
4	42	195.05555555555554
5	74	183.5625
6	122	187.3125
7	201	188.8987341772152
8	337	185.9485294117647
9	557	185.45454545454547
10	918	186.40997229916897
11	1518	184.85666666666665
12	2485	184.75594622543952
13	4045	185.0423076923077
14	6590	184.65186640471512
15	10699	184.63470430761743
16	17398	184.38408717719062
17	28248	184.2589861751152
18	45924	184.2266349852908
19	74828	183.91655134237476
20	121926	183.95231220009342
21	199014	183.7758016811955
22	325276	183.59797088593558
23	532102	183.38948198002186
24	870965	183.30211914549537
25	1427334	183.1269031883516
26	2341404	182.98135263163653
27	3844029	182.8382796772315
28	6316223	182.684873031809
29	10385325	182.54750802511217

Figure 7: Table with statistics for 30 iterations

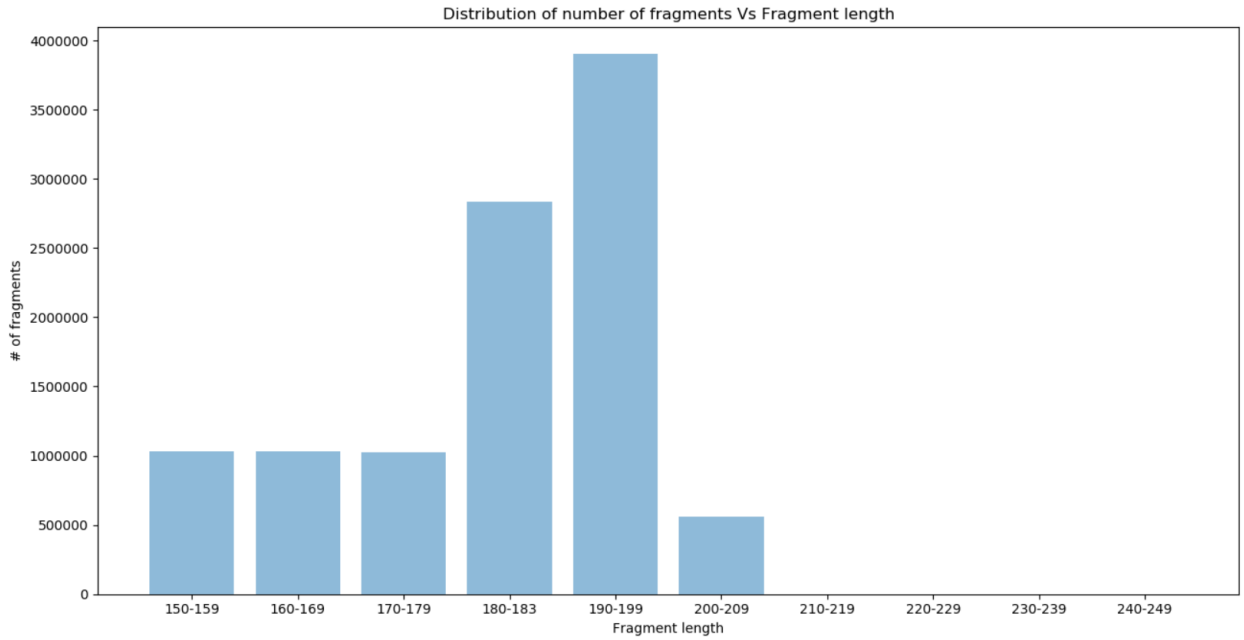

Figure 8: Bar graph with Distribution of number of fragments for 30 iterations

Results with constraints are as shown below. The default iteration is 20 and 15 iterations are used for iteration constraint. 2^10 primers are used for primer constraint and age of the taq polymerase is calculated as Age = (iteration number//3)*2

Number of DNA Fragments = 37138

Average Length of DNA Fragments = 178.955409553557

Iteration	Number of Fragments	Avg Length
0	4	231.0
1	8	187.0
2	15	202.0
3	27	192.75
4	48	187.71428571428572
5	81	187.12121212121212
6	130	185.10204081632654
7	209	181.15189873417722
8	327	179.89830508474577
9	508	183.35911602209944
10	789	182.48042704626334
11	1233	181.25225225225225
12	1916	180.70278184480233
13	2950	181.36557059961316
14	4535	179.9261829652997
15	6944	179.82399335823993
16	10579	180.07620357634113
17	16124	178.91487826871054
18	24489	178.4222355050807
19	37138	178.11281524231165

No Constraint, iteration = 20

Number of DNA Fragments = 2637

Average Length of DNA Fragments = 180.31551004929844

Iteration	Number of Fragments	Avg Length
0	4	225.5
1	8	180.75
2	14	191.66666666666666
3	24	185.5
4	38	183.85714285714286
5	58	180.75
6	86	184.64285714285714
7	134	180.5
8	203	186.53623188405797
9	316	180.28318584070797
10	488	179.7906976744186
11	750	181.49618320610688
12	1149	179.87468671679198
13	1742	180.55817875210792
14	2637	179.58659217877096

Iteration Constraint, iteration = 15

Figure 9: Table with statistics for results with constraints

Number of DNA Fragments = 650

Average Length of DNA Fragments = 182.64461538461538

Iteration	Number of Fragments	Avg Length
0	4	195.5
1	7	181.0
2	11	196.0
3	15	209.5
4	21	196.83333333333334
5	29	194.25
6	42	190.92307692307693
7	65	188.3913043478261
8	103	183.39473684210526
9	164	187.1639344262295
10	260	182.80208333333334
11	413	180.9607843137255
12	650	181.39662447257385

# of primer Constraint, iteration = 15, # primer = 2^10

Number of DNA Fragments = 2289

Average Length of DNA Fragments = 180.911751856706

Iteration	Number of Fragments	Avg Length
0	4	203.5
1	7	173.66666666666666
2	10	213.0
3	15	212.6
4	24	181.44444444444446
5	38	192.07142857142858
6	62	185.58333333333334
7	100	187.28947368421052
8	161	180.9672131147541
9	252	183.5164835164835
10	393	182.45390070921985
11	616	182.74887892376682
12	961	181.2521739130435
13	1481	181.8576923076923
14	2289	178.54950495049505

Age of taq Constraint, iteration = 15,
Age = (iteration number//3)*2

Figure 10: Table with statistics for results with constraints

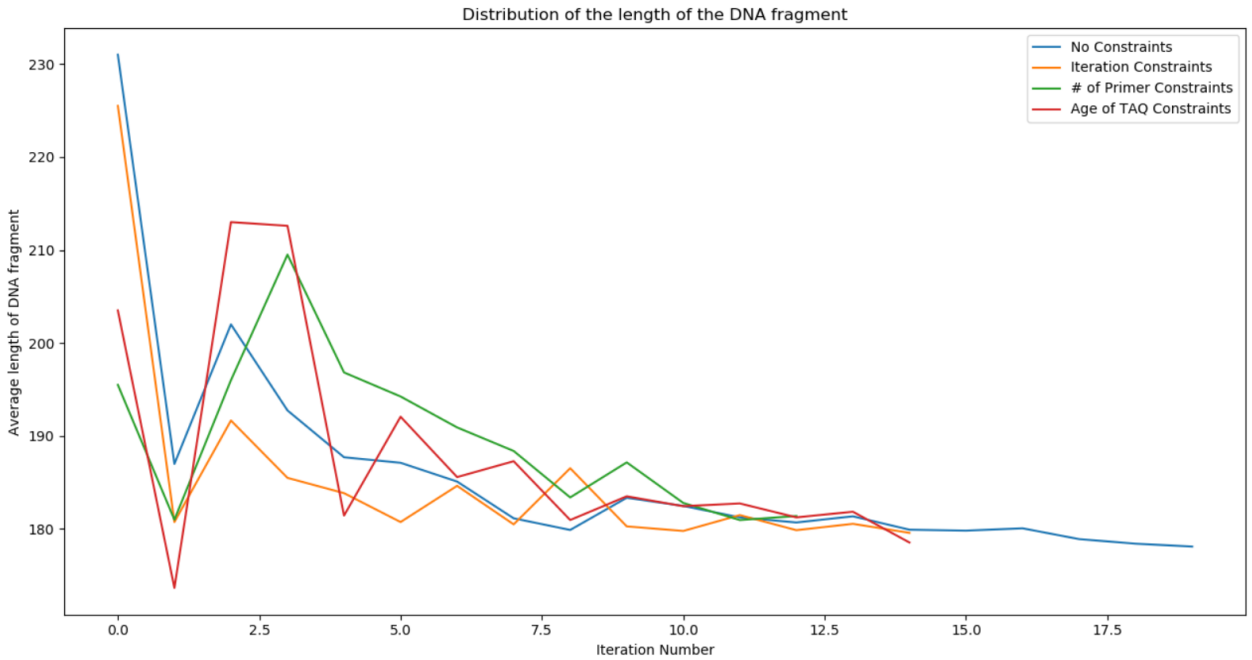

Figure 11: Graph with Distribution of the length for results with constraints

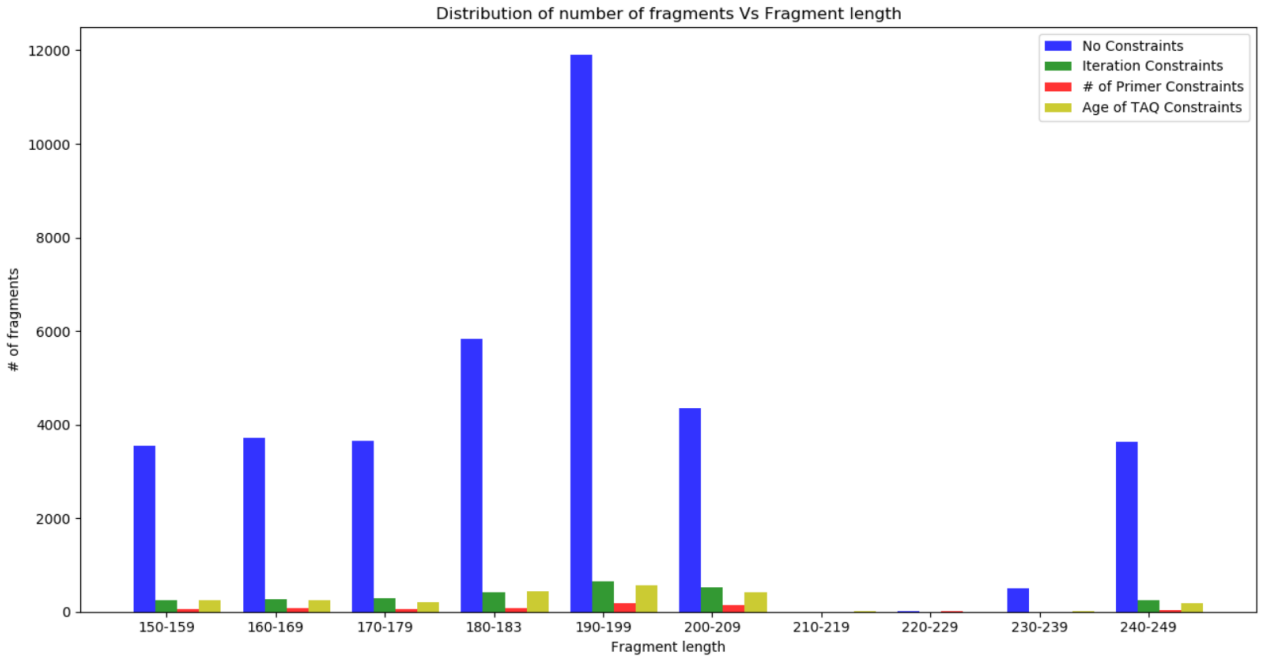

Figure 12: Bar graph with Distribution of number of fragments for results with constraints

4.0 Discussion

Result with no constraints

i) The total number of DNA fragments are less than the possible numbers since taq polymerase has different fall off rate for each segment and those segments which are short to bind at least 10 bases of the primers are discarded in the extension step.

ii) The length of the DNA fragments and the average length except for the original ones will lies between 150 – 250 since the taq falloff rate is 200 +[-50,50]

Result with constraints and its comparison

i) In figure 11, the blue line goes till 20 iterations since there was no constraints used.

ii) The orange line goes up to 14 since 15 iterations are used (iteration counts from 0 – 14)

iii) The green line goes up to 12.5 (13.5) iterations. The amount of primer given was 2^10 and in the ideal case it should go up to 10 iterations. But here some of the fragments are discarded due to its small size in the extension stage, so the primers are not binding to them. These primers are used in the next iteration.

iv) Red line shows the constraint with age of the taq polymerase, and as the ahe / iteration increases, the average length decreases.

v) Figure 12 shows the distribution of the length of the fragments. Most of the fragment lies in the 190-199 region.

5.0 Conclusion

We have simulated the process of PCR and analyzed the results. We have also added few constraints like the number of cycles, amount of primers present and age of taq polymerase and compared the results. Even though the length of the initial DNA has not much effect in the PCR cycle, the number of cycles, the amount of the primers presemt and the age of the taq polymerase can affect the number of fragments and its length after the PCR cycle.

6.0 References

[1] “Polymerase chain reaction,” *Wikipedia*. 10-Feb-2019.

[2] “Principle of the PCR.” [Online]. Available: https://users.ugent.be/~avierstr/principles/pcr.html. [Accessed: 18-Feb-2019].