THE RENATURATION RATE DEPENDENCY ON THE LENGTH AND COMPLEXITY OF DNA

DUAL DEGREE PROJECT REPORT

Submitted as final year project for the award of
Dual Degree(Btech+Mtech) in Electrical engineering

Submitted by

ANSHUL SURYAN
EE16B130

Under the guidance of
Dr. Rajamanickam Murugan

Under the co-guidance of

Dr. Bobby George

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY , MADRAS
Chennai, Tamil Nadu - 600036

JULY 2021

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY , MADRAS
Chennai, Tamil Nadu - 600036

CANDIDATE'S DECLARATION

I, **Anshul Suryan**, Roll Number - **EE16B130**, student of Dual Degree in Electrical Engineering, hereby declare that the Project Dissertation titled "**The Renaturation rate dependency on length and complexity of DNA**" which is submitted by me to the **Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai**, Tamilnadu-600036 in partial fulfilment of the requirement for the award of **Dual Degree(Btech+ MTech) in Electrical Engineering,** is original and not copied from any source without proper citation .I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma.

Place: patiala, punjab
Date : 14.07.2021

Anshul Suryan
EE16B130

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY , MADRAS
Chennai, Tamil Nadu - 600036

CERTIFICATE

This is to certify that the thesis titled **The Renaturation rate dependency on the length and complexity of the DNA** submitted by **Anshul Suryan**, Roll number **EE16B130**, to the Indian Institute of Technology, Madras, for the award of the degree of **Dual Degree(Btech+Mtech) in Electrical Engineering** , is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Place: Chennai — Guide: Dr Rajamanickam Murugan

Date: 14.07.2021 — Co - Guide: Dr Boby George

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY , MADRAS
Chennai, Tamil Nadu - 600036

ACKNOWLEDGEMENT

I am grateful to my guide **Dr. Rajamanickam Murugan,** and my coguide **Dr. Bobby George** for their encouragement and guidance throughout my project. They provided me with constant mentorship and guidance and I thank them for giving me the freedom of working in my own way, which was crucial in getting a lot more interested in this area than I was when I began working on the project. Their comments and useful inputs were instrumental in making progress in my research.

I would also like to express my gratitude to the entire department of Electrical engineering of IIT Madras , as the experience and knowledge I gained here is invaluable and without it, this project could not have been possible.

Place: patiala, punjab　　Anshul Suryan

Date : 14.07.2021　　EE16B130

ABSTRACT

DNA renaturation is the process of recombination of two single strands(c-ssDNA) to form a double stranded DNA(dsDNA). Previous studies have shown that renaturation involves a slower process of nucleation of the two reacting c-ssDNA strands followed by a much faster zipping. Hence, The overall rate of renturation is dependent on the nucleation rate and the zippering rate. We model the DNA strands as 3D self avoiding random walks(SARW) confined in a lattice cell and observe the variation of probability of correct contact formation, which is required for the nucleation step, with various parameters. The zipping rate depends on the ratio of length and the complexity of the DNA. experimentally it is known that the overall rate is directly proportional to the square root of length of DNA strands and inversely proportional to its complexity.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

SYMBOLS,ABBREVIATIONS AND THEIR DEFINITION

- DNA: Deoxyribonucleic Acid
- c-ssDNA: complementary single stranded DNA
- dsDNA: double stranded DNA
- L: Length of DNA(number of base pairs)
- l: length of side of lattice box
- c: complexity (length of repeating sequence in DNA)
- ϱ: copy number(L/c)
- SARW: Self avoiding random walk
- RW: random walker
- RV: random variable
- 3D : three dimensional
- k_n: rate of nucleation
- k_z: rate of zipping
- $k_{renaturation}$: overall rate of renaturation
- MSE : Mean squared error

1. INTRODUCTION

DNA ,in its natural state, occurs in a double stranded(dsDNA) helical structure ,which is stabilized by weak hydrogen bonds between nitrogen bases of individual strands of DNA. When heated above the melting temperature, this double stranded structure breaks down yielding two corresponding single strands of DNA(c-ssDNA), this process is called denaturation of the DNA. Once the heat is removed and the DNA starts to cool down again, the c-ssDNA strands reunite back to form the original dsDNA structure, this process is known as the renaturation. In this paper we explore the dependencies of this renaturation rate on various parameters of DNA.

The length and complexity of the interacting c-ssDNA are the main parameters that determine the renaturation rate. The length(L) is the number of the base pairs in the c-ssDNA strands. The complexity(c) is defined as the length of the repeating sequence(if any) in the c-ssDNA strand. In the case where there is no repeating sequence, the complexity will be simply equal to the length of c-ssDNA. The copy number(ρ) is defined as the ratio of total length to complexity of DNA. experimentally it is known that the renaturation rate is directly proportional to square root of Length and inversely proportional to the complexity.

In this paper, we primarily focus on the nucleation and zipping step of the denaturation process. Nucleation occurs when sufficient base pairs(above a certain critical number N) of c-ssDNA are in correct contact to form a nucleus. The Zipping occurs after the formation of nucleus ,resulting in formation on the dsDNA. Since the overall renaturation rate is dependent on the nucleation rate and the zipping rate, we consider these rates to analyse the renaturation rate.

We model the DNA as a self avoiding 3D random walk confined in a lattice box. Using various methods, we analyse the effects that these parameters(length of dna, volume of lattice box, complexity) have on the renaturation rate .

2. MODEL OVERVIEW

We consider a cubical lattice box of side length as *l*. We model the c-ssDNA strands as a self avoiding random walk(SARW) in three dimensions confined in this lattice box. The random walk is a L step walk, where L is the length(number of base pairs) of interacting c-ssDNA strands. We consider a volumetric walk, which means that every step of the random walker is of length √3 (along the longest diagonal of the individual lattice cell). It is important to note that the *L* lies between *0* and l^3 as the length of DNA can not exceed the available lattice points.

Since two c-ssDNA interact to form a dsDNA. We generate two such mutually avoiding c-ssDNA simultaneously in the lattice box.The model accomplishes this by randomly selecting one of the c-ssDNA and incrementing it by one step while looking for the presence of the other strand at every step. By following this process for every step the model generated two self avoiding random walks of length L which are also mutually avoiding.

After obtaining the DNA strands we look for the resulting number of “correct contacts" . A correct contact is defined when the distance between two bases of the same step of the interacting complementary strands(the random walks in this case) is less than √3 in the lattice box. Similarly an incorrect contact is defined when the distance between two bases of different steps of the interacting complementary strands is less than √3 in the lattice box. We define the correct contact probability as the number of correct contacts divided by the number of total(correct and incorrect) contacts. Since the nucleation step involves correct contacts between the base pairs of interacting strands, the nucleation rate(K_n) is directly proportional to the probability of correct contacts.Hence,The relation between nucleation rates and the parameters will be the same as the relation between probability of correct contacts and the parameters . We compute this probability while varying other parameters such as Length of c-ssDNA.

In the case when the DNA has a repeating sequence, an incorrect contact can also result in formation of a partial duplex .For example, let’s consider the case when complexity is 10, which means that a repeating sequence of length 10 forms the entire DNA. Now, a contact between the 2nd base of one c-ssDNA and the 12th base of the other c-ssDNA will lead to formation of a partial duplex. In our model we define the probability of partial contacts as the number of partial contacts divided by the number of total(correct and incorrect) contacts. The variation of this probability with respect to the length of the repeating sequence is also analyzed to observe its dependency on the complexity. The probability of partial contacts scales inversely with the complexity and hence we can conclude that the zipping rate is proportional to repeats(copy number) in the DNA.

3. METHODS USED FOR SIMULATION

We begin our simulation by defining a cubical box of length l over a 3D space.Then we randomly choose two different points inside the lattice box. These points denote the starting point of the two c-ssDNA. We check if the chosen points are different before passing them as input parameters to the next functions. Then we take these two points as input in a random walk generating function that returns two lists of length L that represent two self avoiding random walks SARW_1 and SARW_2, that are also mutually avoiding, and every element(a list of length 3) in those lists represents a coordinate point inside the lattice box where the c-ssDNA is present.

3.1 Choosing a SARW to increment

Now, we pass the chosen starting points as input parameters into our random walk generating function. Here,The function chooses a SARW randomly to increment at every step. The function accomplishes this by setting a random variable *r* that can take values either 0 or 1. The function increments SARW_1 if *r* returns a value 0 , and increments SARW_2 if *r* returns value 1. To check the possibilities for the next step, the random walk generating function uses another helper function ,discussed below.

3.2 Incrementing the SARW

Let's say the random walk generating function chose SARW_1 to increment. Let's assume that the SARW_1 is currently at point (X_1,Y_1,Z_1). We define a possibilities function . The function contains the following eight elements:

[1, 1, 1]

[1, 1,-1]

[1,-1, 1]

[-1, 1, 1]

[1,-1,-1]

[-1, 1,-1]

[-1,-1, 1]

[-1,-1,-1]

These elements represent all possible next steps that the walker can take from its current position.The possibilities function randomly chooses an element from the above list depending on the value of a random variable *p*, from the above list and adds it to the current point of SARW_1 . Let's say the possibilities function chooses the element (1,1,-1) based on the value of a random variable ,then the next point added to SARW_1 will be

(X_1+1,Y_1+1,Z_1-1). Now, the walk generating function checks the feasibility of this next point.

The feasibility criteria is:

1. The coordinates of the next point exist within the limits of the lattice box. For our example, this means that,

   $$0 \leq X_1 + 1 \leq l$$

   $$0 \leq Y_1 + 1 \leq l$$

   $$0 \leq Z_1 - 1 \leq l$$

2. The next point is not already a part of either of the SARWs generated until the current step.

If the above conditions are met then the sarw1 is incremented and the process ,starting from randomly choosing a SARW to increment, repeats again. If the conditions are not met then this point is discarded and the possibilities function is called again and another element is chosen randomly from the remaining elements of possibilities list to add to $SARW_1$ and then conditions are verified again .

The process repeats till we generate both the SARWs of required length *L*. This simultaneous generation of mutually avoiding SARWs can be interpreted as 3D-3D diffusion of the c-ssDNA.

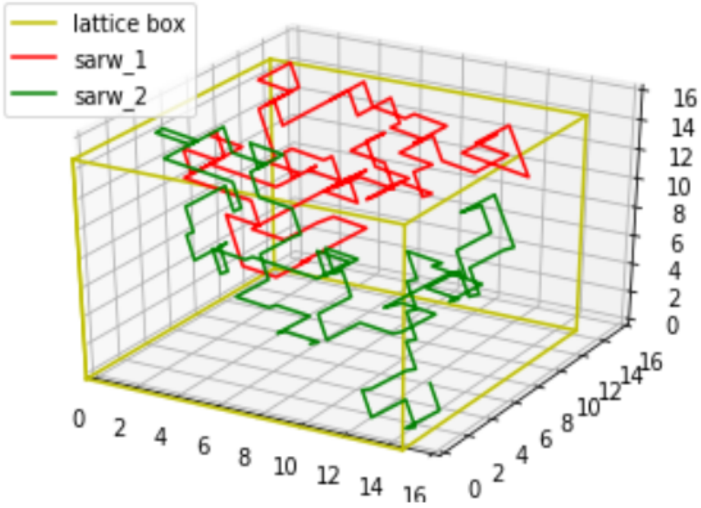
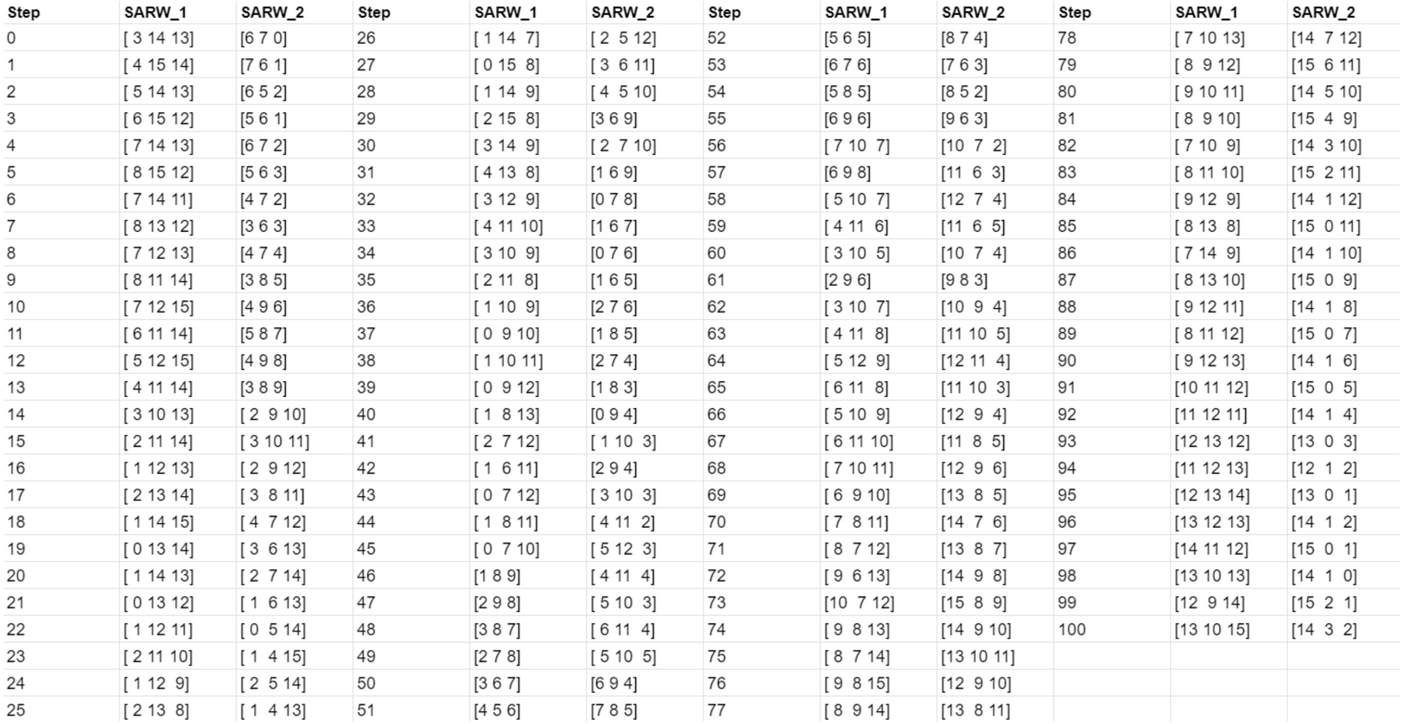


Fig 3.1 Two SARWs of length(*L*) = 100, generated in a cubical lattice box of side length(*l*) = 15



Step	SARW_1	SARW_2	Step	SARW_1	SARW_2	Step	SARW_1	SARW_2	Step	SARW_1	SARW_2
0	[3 14 13]	[6 7 0]	26	[1 14 7]	[2 5 12]	52	[5 6 5]	[8 7 4]	78	[7 10 13]	[14 7 12]
1	[4 15 14]	[7 6 1]	27	[0 15 8]	[3 6 11]	53	[6 7 6]	[7 6 3]	79	[8 9 12]	[15 6 11]
2	[5 14 13]	[6 5 2]	28	[1 14 9]	[4 5 10]	54	[5 8 5]	[8 5 2]	80	[9 10 11]	[14 5 10]
3	[6 15 12]	[5 6 1]	29	[2 15 8]	[3 6 9]	55	[6 9 6]	[9 6 3]	81	[8 9 10]	[15 4 9]
4	[7 14 13]	[6 7 2]	30	[3 14 9]	[2 7 10]	56	[7 10 7]	[10 7 2]	82	[7 10 9]	[14 3 10]
5	[8 15 12]	[5 6 3]	31	[4 13 8]	[1 6 9]	57	[6 9 8]	[11 6 3]	83	[8 11 10]	[15 2 11]
6	[7 14 11]	[4 7 2]	32	[3 12 9]	[0 7 8]	58	[5 10 7]	[12 7 4]	84	[9 12 9]	[14 1 12]
7	[8 13 12]	[3 6 3]	33	[4 11 10]	[1 6 7]	59	[4 11 6]	[11 6 5]	85	[8 13 8]	[15 0 11]
8	[7 12 13]	[4 7 4]	34	[3 10 9]	[0 7 6]	60	[3 10 5]	[10 7 4]	86	[7 14 9]	[14 1 10]
9	[8 11 14]	[3 8 5]	35	[2 11 8]	[1 6 5]	61	[2 9 6]	[9 8 3]	87	[8 13 10]	[15 0 9]
10	[7 12 15]	[4 9 6]	36	[1 10 9]	[2 7 6]	62	[3 10 7]	[10 9 4]	88	[9 12 11]	[14 1 8]
11	[6 11 14]	[5 8 7]	37	[0 9 10]	[1 8 5]	63	[4 11 8]	[11 10 5]	89	[8 11 12]	[15 0 7]
12	[5 12 15]	[4 9 8]	38	[1 10 11]	[2 7 4]	64	[5 12 9]	[12 11 4]	90	[9 12 13]	[14 1 6]
13	[4 11 14]	[3 8 9]	39	[0 9 12]	[1 8 3]	65	[6 11 8]	[11 10 3]	91	[10 11 12]	[15 0 5]
14	[3 10 13]	[2 9 10]	40	[1 8 13]	[0 9 4]	66	[5 10 9]	[12 9 4]	92	[11 12 11]	[14 1 4]
15	[2 11 14]	[3 10 11]	41	[2 7 12]	[1 10 3]	67	[6 11 10]	[11 8 5]	93	[12 13 12]	[13 0 3]
16	[1 12 13]	[2 9 12]	42	[1 6 11]	[2 9 4]	68	[7 10 11]	[12 9 6]	94	[11 12 13]	[12 1 2]
17	[2 13 14]	[3 8 11]	43	[0 7 12]	[3 10 3]	69	[6 9 10]	[13 8 5]	95	[12 13 14]	[13 0 1]
18	[1 14 15]	[4 7 12]	44	[1 8 11]	[4 11 2]	70	[7 8 11]	[14 7 6]	96	[13 12 13]	[14 1 2]
19	[0 13 14]	[3 6 13]	45	[0 7 10]	[5 12 3]	71	[8 7 12]	[13 8 7]	97	[14 11 12]	[15 0 1]
20	[1 14 13]	[2 7 14]	46	[1 8 9]	[4 11 4]	72	[9 6 13]	[14 9 8]	98	[13 10 13]	[14 1 0]
21	[0 13 12]	[1 6 13]	47	[2 9 8]	[5 10 3]	73	[10 7 12]	[15 8 9]	99	[12 9 14]	[15 2 1]
22	[1 12 11]	[0 5 14]	48	[3 8 7]	[6 11 4]	74	[9 8 13]	[14 9 10]	100	[13 10 15]	[14 3 2]
23	[2 11 10]	[1 4 15]	49	[2 7 8]	[5 10 5]	75	[8 7 14]	[13 10 11]			
24	[1 12 9]	[2 5 14]	50	[3 6 7]	[6 9 4]	76	[9 8 15]	[12 9 10]			
25	[2 13 8]	[1 4 13]	51	[4 5 6]	[7 8 5]	77	[8 9 14]	[13 8 11]			

Table 3.1 Tabular form of stepwise coordinates of $SARW_1$ and $SARW_2$

3.3 Finding correct and incorrect contacts

In our model we define a Contact checker function that takes the lists SARW1 and SARW2 as input parameters and returns the number of correct contacts, incorrect contacts and partial contacts.

The magnitude of distance vector is taken as euclidean distance between the ith coordinates of the SARWs i.e

$\sqrt{}\{(sarw_1[i][0]\ -\ sarw_2[i][0]\ )^2+(sarw_1[i][1]\ \ -\ sarw_2[i][1]\ )^2+(sarw_1[i][2]\ -\ sarw_2[i][2]\ )^2\}$

A correct contact is defined as a contact where the magnitude of the distance vector between the same step of SARWs is less than √3. The contact checker function returns the total number of correct contacts using the above condition.

An incorrect contact is defined as a contact where the magnitude of distance vector between different steps of SARWs is less than √3.Contact checker function returns the total number of incorrect contacts using the above condition.

The total number of contacts are the sum of correct and incorrect contacts.

3.4 Incorrect contacts resulting in the formation of partial duplex

When there is a repeating sequence in the DNA, an incorrect contact can result in the formation of partial duplex . A partial duplex has overhangs and hence is not stable as a dsDNA. A partial duplex is defined when an i^{th} step of SARW$_1$ is in contact with $(i + nc)^{th}$ step of SARW$_2$, where *n* is a non zero integer conditioned on $0 \leq i + nc \leq L$, and c is the complexity of the DNA. The contact checker function returns the total number of such partial contacts using the above condition.

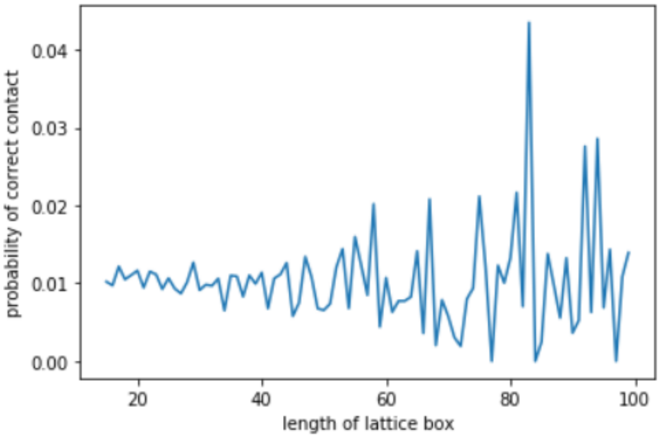
4. RESULTS

4.1 Variation of correct contact probability w.r.t. the volume of the box

When the length of sarw is kept constant and the length of the cubic lattice box is increased, the number of correct as well as incorrect contacts reduces. The reason behind this observation is that when the length of the cube is increased, the volume of the box increases and hence the SARW now has more volume to expand into rather than being compressed into a smaller volume. And since the length of SARW is kept constant, this expansion leads to reduction in contacts between the SARWs.

Now, since the number of correct and total contacts approaches zero very quickly when the volume is increased, their ratio does not provide an accurate idea of variation of probability of correct contacts.

Hence, we look at the number of correct contacts in this case. We compute the number of correct contacts as the length of the box increases and the results show that the number of correct contacts is inversely proportional to the cube of the length of the lattice box.



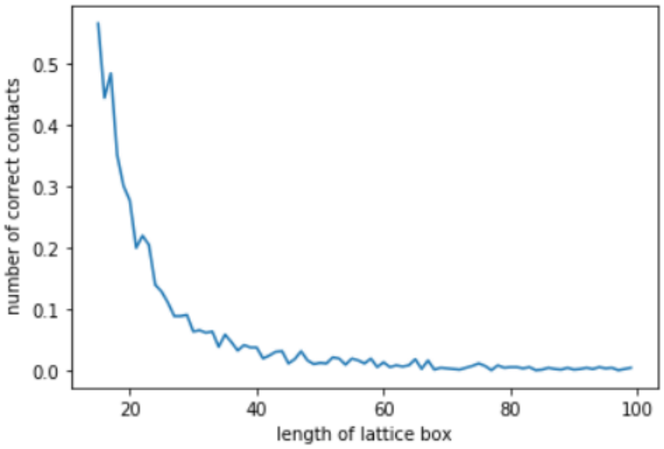


Fig 4.1 Probability of correct contact vs length of cubical lattice box

Fig 4.2 Average number of correct contacts vs length of cubical lattice box

lattice side length	average correct contacts	lattice side length	average correct contacts	lattice side length	average correct contacts	lattice side length	average correct contacts
15	0.697	37	0.046	59	0.015	81	0.006
16	0.477	38	0.036	60	0.004	82	0
17	0.378	39	0.041	61	0.011	83	0.01
18	0.332	40	0.041	62	0.013	84	0.003
19	0.328	41	0.026	63	0.011	85	0.002
20	0.239	42	0.021	64	0.007	86	0.003
21	0.202	43	0.025	65	0.008	87	0.004
22	0.22	44	0.019	66	0.006	88	0.008
23	0.163	45	0.021	67	0.008	89	0.005
24	0.137	46	0.015	68	0.007	90	0.004
25	0.134	47	0.016	69	0.01	91	0.004
26	0.102	48	0.03	70	0.001	92	0.003
27	0.114	49	0.028	71	0	93	0.009
28	0.101	50	0.015	72	0.003	94	0.003
29	0.127	51	0.014	73	0.002	95	0.006
30	0.104	52	0.01	74	0.012	96	0.002
31	0.07	53	0.011	75	0.005	97	0.001
32	0.055	54	0.014	76	0	98	0.001
33	0.078	55	0.01	77	0.006	99	0.004
34	0.089	56	0.016	78	0.005	100	0
35	0.042	57	0.011	79	0.003		
36	0.056	58	0.01	80	0		

Table 4.1 Data showing Average number of correct contacts against length of cubical lattice box

4.2 Variation of correct contact probability w.r.t shape of lattice box

In this segment we explore the variation in the contact probabilities as we vary the shape of the box. The volume of lattice box is kept constant at 1000 and the z-dimension of the box is varied . The other two dimensions of the box are calculated as , $x = y = (\surd 1000/z)$,the values are rounded off to the nearest integer.

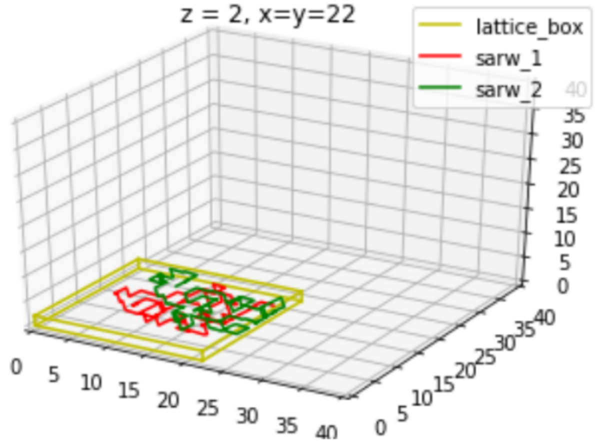


Fig 4.3 SARWs in lattice box with dimensions z = 2, x=y=22

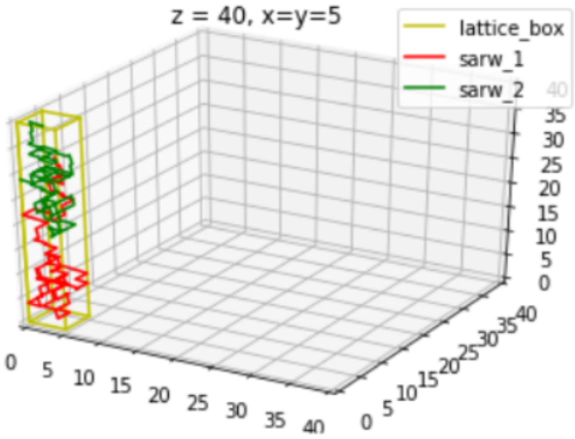


Fig 4.4 SARWs in lattice box with dimensions z = 40, x=y=5

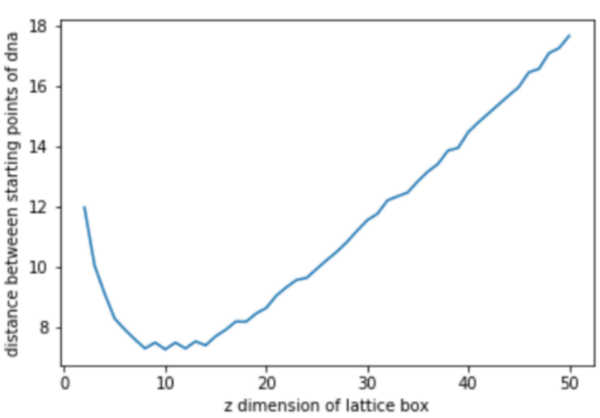


Fig 4.5 Average distance between starting points of SARWSs against z-dimension

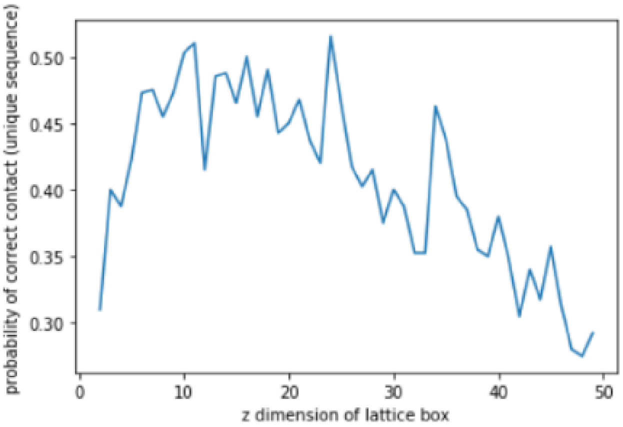


Fig 4.6 Probability of correct contacts against z-dimension

The probability of correct contacts initially increases with increase in z-dimension and then decreases , obtaining a maxima at z = 10. This observation can be explained as the average distance between randomly chosen starting points of the SARWs is minimum when the lattice box is a perfect cube $(x = y = z = 10)$. This is in line with the fact the length of longest diagonal of the box is minimum$(10\sqrt{3})$ when the box is cubical. This length increases when the shape of lattice box deviates from perfect cube.

4.3 <u>Variation of correct contact probability w.r.t the length of SARW</u>

In this segment we observe the variation of probability of correct contact with respect to the length of interacting SARWs. The length of the cube is kept constant at 25 dimensionless units and the length of SARW is varied from 100 to 500 dimensionless units in steps of 5 (100,105,110...495, 500). Under this situation, both correct contacts and total contacts increase .This can be attributed to the fact that since the volume of the cube is constant, more and more number of base pairs are being interacting in a confined space and hence contacts increase. But the rate of increase of the total contacts surpasses that of the correct contacts, hence the ratio, which gives us the probability of correct contacts, decreases as *L* increases.

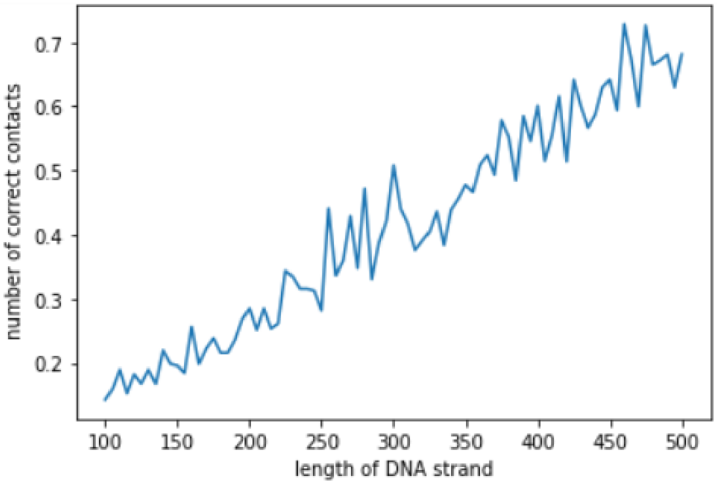
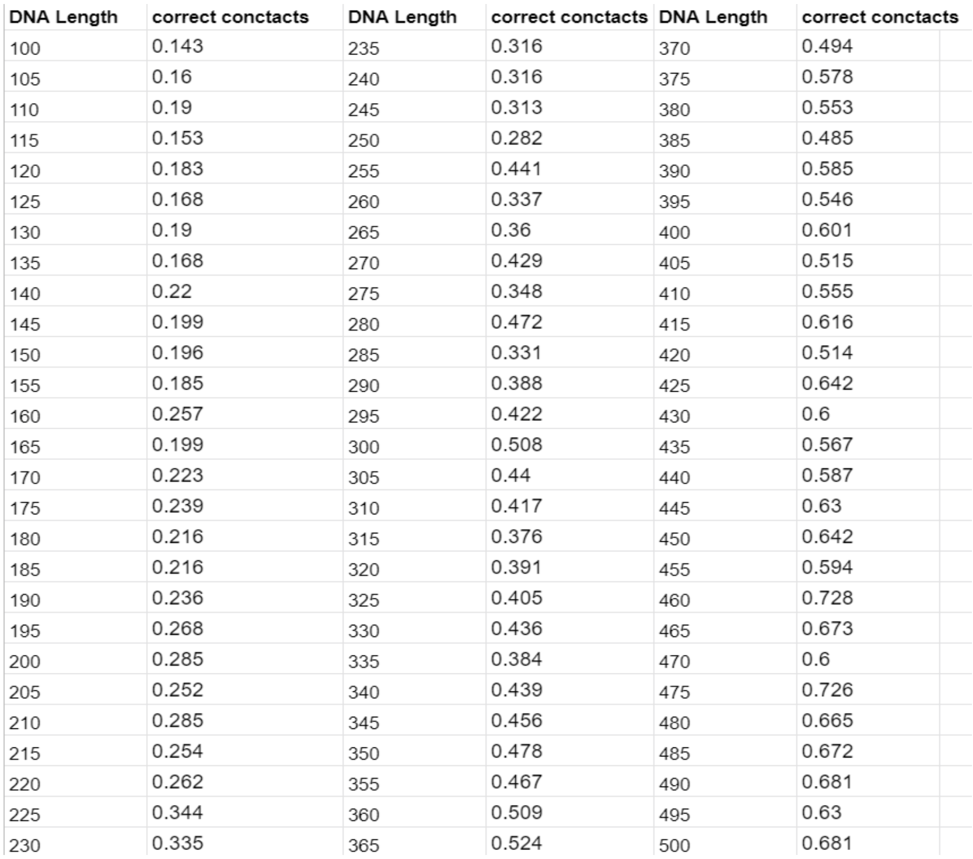


Fig 4.7 Average number of correct contact against the length of DNA strand(*L*)



DNA Length	correct conctacts	DNA Length	correct conctacts	DNA Length	correct conctacts
100	0.143	235	0.316	370	0.494
105	0.16	240	0.316	375	0.578
110	0.19	245	0.313	380	0.553
115	0.153	250	0.282	385	0.485
120	0.183	255	0.441	390	0.585
125	0.168	260	0.337	395	0.546
130	0.19	265	0.36	400	0.601
135	0.168	270	0.429	405	0.515
140	0.22	275	0.348	410	0.555
145	0.199	280	0.472	415	0.616
150	0.196	285	0.331	420	0.514
155	0.185	290	0.388	425	0.642
160	0.257	295	0.422	430	0.6
165	0.199	300	0.508	435	0.567
170	0.223	305	0.44	440	0.587
175	0.239	310	0.417	445	0.63
180	0.216	315	0.376	450	0.642
185	0.216	320	0.391	455	0.594
190	0.236	325	0.405	460	0.728
195	0.268	330	0.436	465	0.673
200	0.285	335	0.384	470	0.6
205	0.252	340	0.439	475	0.726
210	0.285	345	0.456	480	0.665
215	0.254	350	0.478	485	0.672
220	0.262	355	0.467	490	0.681
225	0.344	360	0.509	495	0.63
230	0.335	365	0.524	500	0.681

Table 4.2 Data showing the average number of correct contact against the length of DNA strand

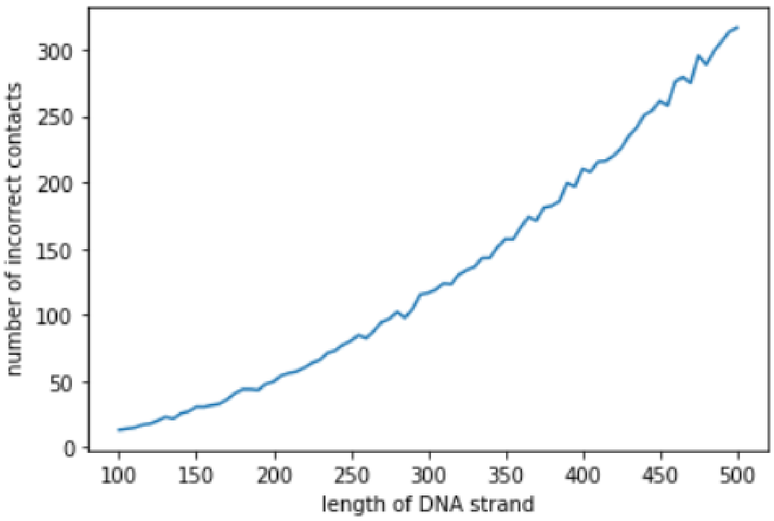
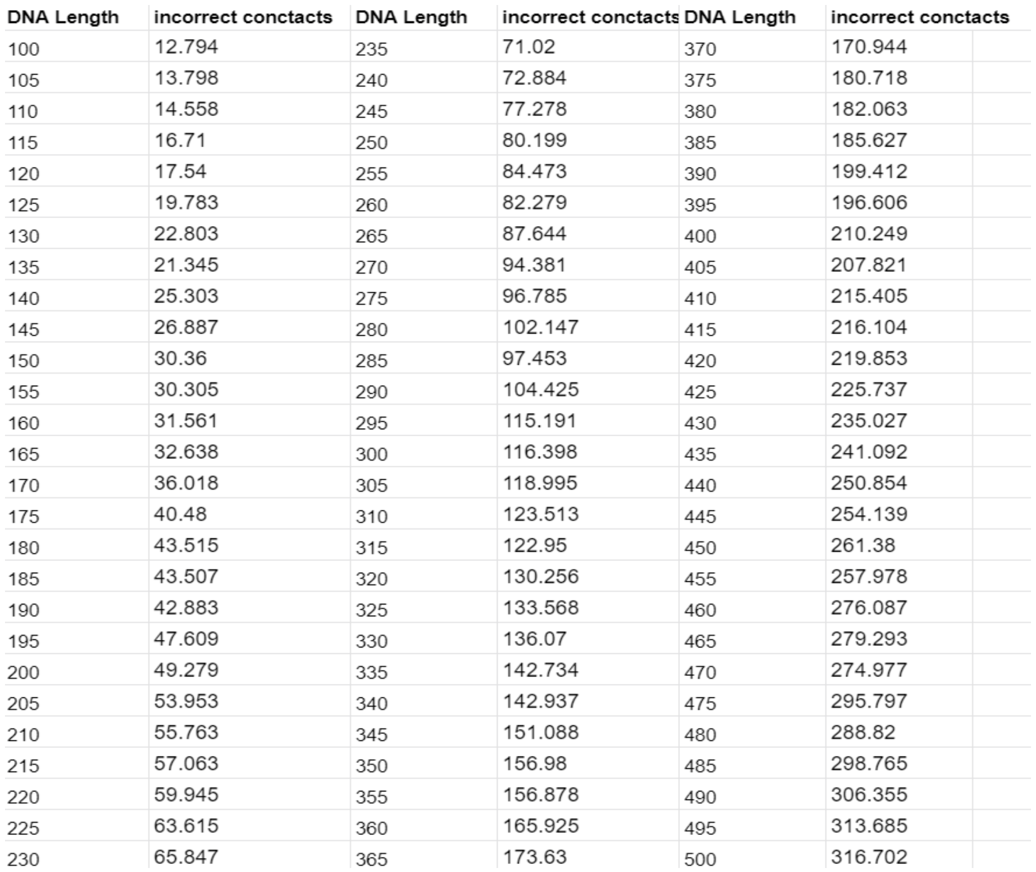


Fig 4.8 Average number of incorrect contact against the length of DNA strand(L)



DNA Length	incorrect conctacts	DNA Length	incorrect conctacts	DNA Length	incorrect conctacts
100	12.794	235	71.02	370	170.944
105	13.798	240	72.884	375	180.718
110	14.558	245	77.278	380	182.063
115	16.71	250	80.199	385	185.627
120	17.54	255	84.473	390	199.412
125	19.783	260	82.279	395	196.606
130	22.803	265	87.644	400	210.249
135	21.345	270	94.381	405	207.821
140	25.303	275	96.785	410	215.405
145	26.887	280	102.147	415	216.104
150	30.36	285	97.453	420	219.853
155	30.305	290	104.425	425	225.737
160	31.561	295	115.191	430	235.027
165	32.638	300	116.398	435	241.092
170	36.018	305	118.995	440	250.854
175	40.48	310	123.513	445	254.139
180	43.515	315	122.95	450	261.38
185	43.507	320	130.256	455	257.978
190	42.883	325	133.568	460	276.087
195	47.609	330	136.07	465	279.293
200	49.279	335	142.734	470	274.977
205	53.953	340	142.937	475	295.797
210	55.763	345	151.088	480	288.82
215	57.063	350	156.98	485	298.765
220	59.945	355	156.878	490	306.355
225	63.615	360	165.925	495	313.685
230	65.847	365	173.63	500	316.702

Table 4.3 Data showing the number of incorrect contact against the length of DNA strand

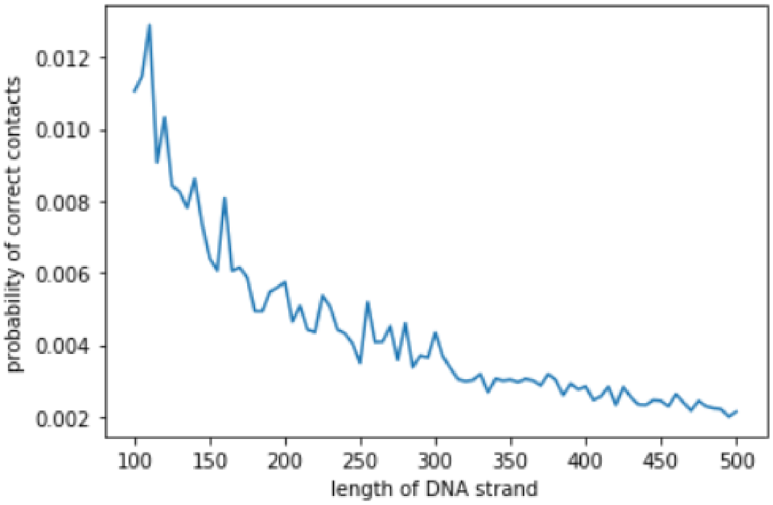
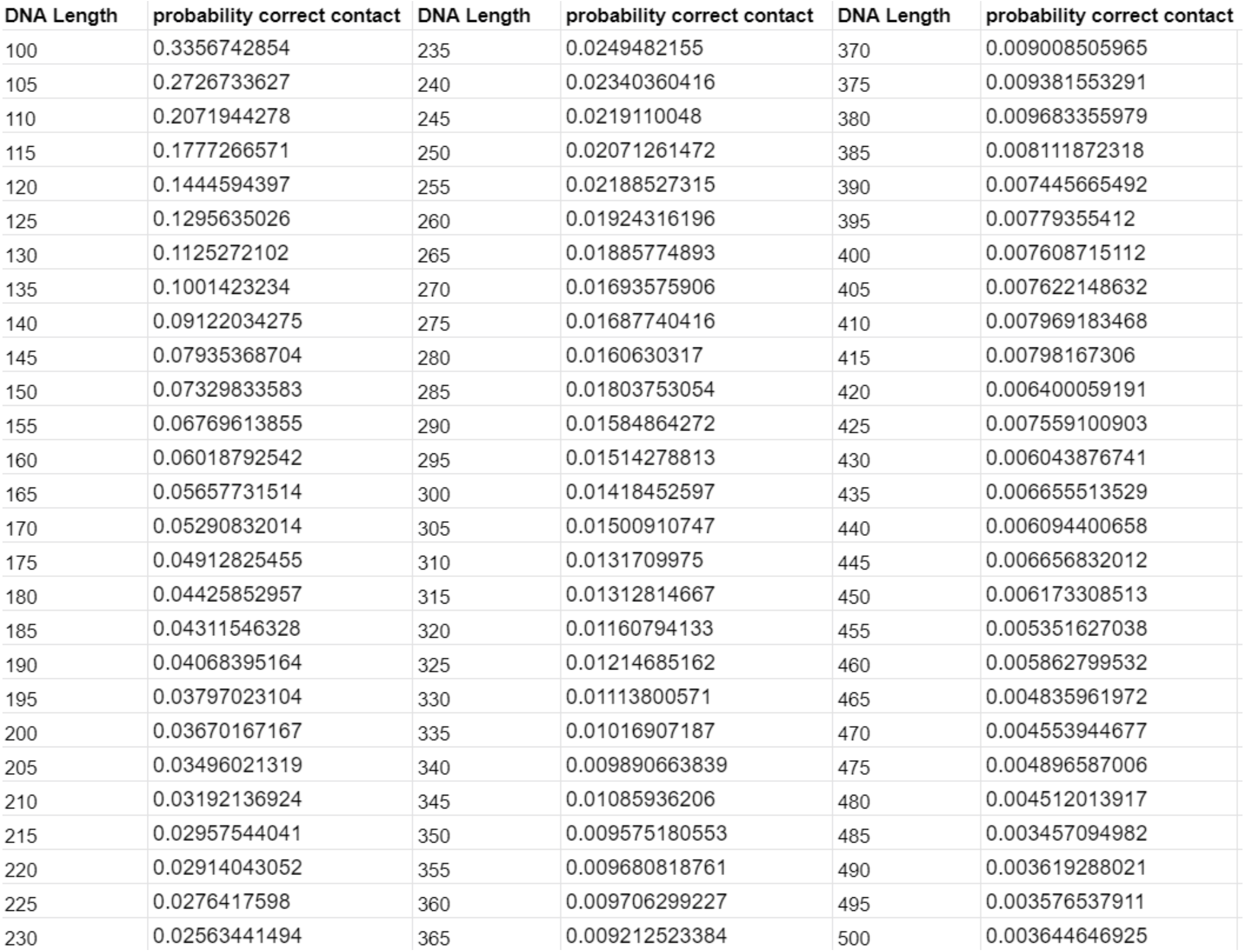


Fig 4.9 Probability of correct contact against the length of DNA strand(L)



DNA Length	probability correct contact	DNA Length	probability correct contact	DNA Length	probability correct contact
100	0.3356742854	235	0.0249482155	370	0.009008505965
105	0.2726733627	240	0.02340360416	375	0.009381553291
110	0.2071944278	245	0.0219110048	380	0.009683355979
115	0.1777266571	250	0.02071261472	385	0.008111872318
120	0.1444594397	255	0.02188527315	390	0.007445665492
125	0.1295635026	260	0.01924316196	395	0.00779355412
130	0.1125272102	265	0.01885774893	400	0.007608715112
135	0.1001423234	270	0.01693575906	405	0.007622148632
140	0.09122034275	275	0.01687740416	410	0.007969183468
145	0.07935368704	280	0.0160630317	415	0.00798167306
150	0.07329833583	285	0.01803753054	420	0.006400059191
155	0.06769613855	290	0.01584864272	425	0.007559100903
160	0.06018792542	295	0.01514278813	430	0.006043876741
165	0.05657731514	300	0.01418452597	435	0.006655513529
170	0.05290832014	305	0.01500910747	440	0.006094400658
175	0.04912825455	310	0.0131709975	445	0.006656832012
180	0.04425852957	315	0.01312814667	450	0.006173308513
185	0.04311546328	320	0.01160794133	455	0.005351627038
190	0.04068395164	325	0.01214685162	460	0.005862799532
195	0.03797023104	330	0.01113800571	465	0.004835961972
200	0.03670167167	335	0.01016907187	470	0.004553944677
205	0.03496021319	340	0.009890663839	475	0.004896587006
210	0.03192136924	345	0.01085936206	480	0.004512013917
215	0.02957544041	350	0.009575180553	485	0.003457094982
220	0.02914043052	355	0.009680818761	490	0.003619288021
225	0.0276417598	360	0.009706299227	495	0.003576537911
230	0.02563441494	365	0.009212523384	500	0.003644646925

Table 4.4 Data showing the probability of contact against the length of DNA strand

4.4 Variation of partial contact probability w.r.t complexity of the DNA

Finally, we observe the variation of probability of partial contact with respect to change in the complexity. It is important to note that the probability of correct contact is independent of the complexity since a correct contact can only happen between the same step base of interacting c-ssDNA and complexity does not have any impact on this.

The findings show an inverse relation between probability of partial contact and the complexity. The explanation of this is that when the complexity increases, keeping the length of DNA constant, the copy number decreases which means that the number of repeats in the DNA decreases , and this reduces the probability of partial contact

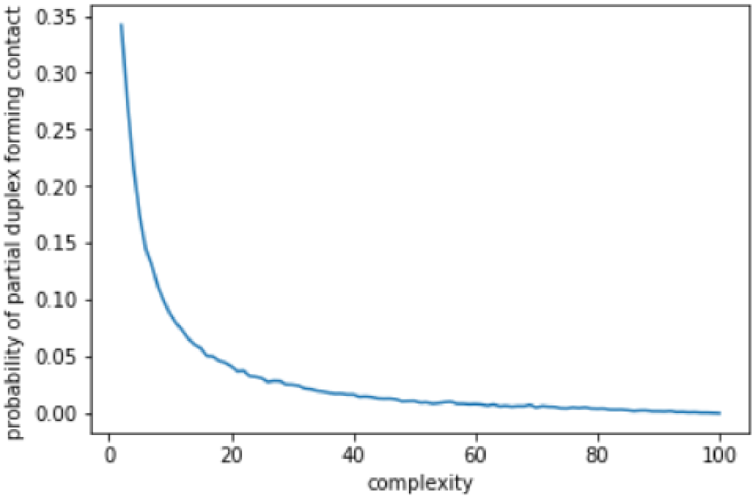
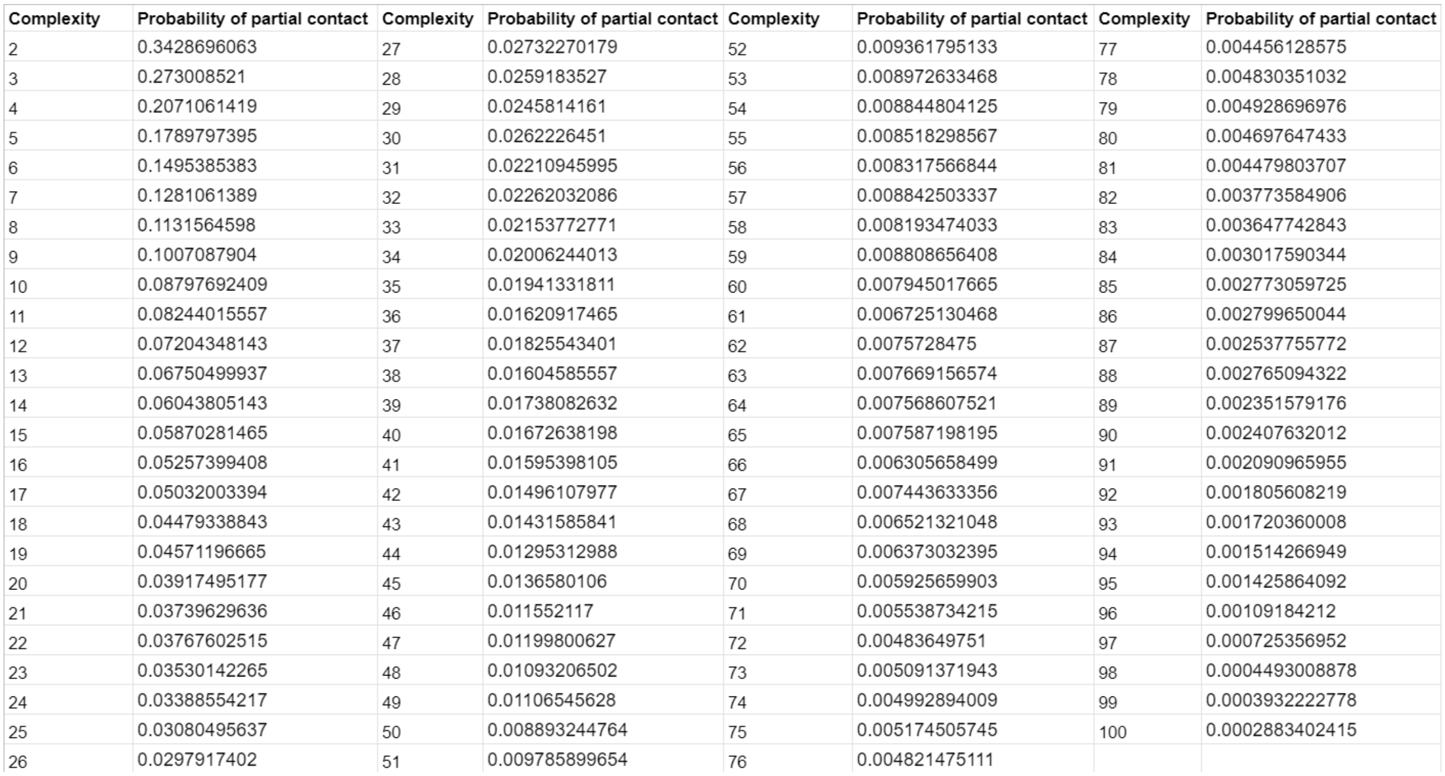


Fig 4.10 Probability of partial contact against the complexity of DNA (*c*)



Complexity	Probability of partial contact	Complexity	Probability of partial contact	Complexity	Probability of partial contact	Complexity	Probability of partial contact
2	0.3428696063	27	0.02732270179	52	0.009361795133	77	0.004456128575
3	0.273008521	28	0.0259183527	53	0.008972633468	78	0.004830351032
4	0.2071061419	29	0.0245814161	54	0.008844804125	79	0.004928696976
5	0.1789797395	30	0.0262226451	55	0.008518298567	80	0.004697647433
6	0.1495385383	31	0.02210945995	56	0.008317566844	81	0.004479803707
7	0.1281061389	32	0.02262032086	57	0.008842503337	82	0.003773584906
8	0.1131564598	33	0.02153772771	58	0.008193474033	83	0.003647742843
9	0.1007087904	34	0.02006244013	59	0.008808656408	84	0.003017590344
10	0.08797692409	35	0.01941331811	60	0.007945017665	85	0.002773059725
11	0.08244015557	36	0.01620917465	61	0.006725130468	86	0.002799650044
12	0.07204348143	37	0.01825543401	62	0.0075728475	87	0.002537755772
13	0.06750499937	38	0.01604585557	63	0.007669156574	88	0.002765094322
14	0.06043805143	39	0.01738082632	64	0.007568607521	89	0.002351579176
15	0.05870281465	40	0.01672638198	65	0.007587198195	90	0.002407632012
16	0.05257399408	41	0.01595398105	66	0.006305658499	91	0.002090965955
17	0.05032003394	42	0.01496107977	67	0.007443633356	92	0.001805608219
18	0.04479338843	43	0.01431585841	68	0.006521321048	93	0.001720360008
19	0.04571196665	44	0.01295312988	69	0.006373032395	94	0.001514266949
20	0.03917495177	45	0.0136580106	70	0.005925659903	95	0.001425864092
21	0.03739629636	46	0.011552117	71	0.005538734215	96	0.00109184212
22	0.03767602515	47	0.01199800627	72	0.00483649751	97	0.000725356952
23	0.03530142265	48	0.01093206502	73	0.005091371943	98	0.0004493008878
24	0.03388554217	49	0.01106545628	74	0.004992894009	99	0.0003932222778
25	0.03080495637	50	0.008893244764	75	0.005174505745	100	0.0002883402415
26	0.0297917402	51	0.009785899654	76	0.004821475111		

Table 4.5 Data showing the probability of partial contact against the complexity of DNA strand

5. ANALYSIS

The analysis of the results is performed using Least square fitting to obtain the functional dependence between the probabilities and the various parameters.

The inbuilt function of python $curve\_fit$ is used to generate the fitting curve. In the following plots, the data is represented by a bold blue line and the generated fit is represented by a dashed green line. The error analysis is performed using the Mean squared error.

We analyse the following results and find the their functional dependencies using the $curve\_fit$.

- Variation of number of correct contacts w.r.t. volume of the box
- Variation of number of correct contacts w.r.t the length of SARW
- Variation of number of total contacts w.r.t the length of SARW
- Variation of correct contact probability w.r.t the length of SARW
- Variation of partial contact probability w.r.t complexity of the DNA

5.1 Variation of number of correct contacts w.r.t. volume of the box

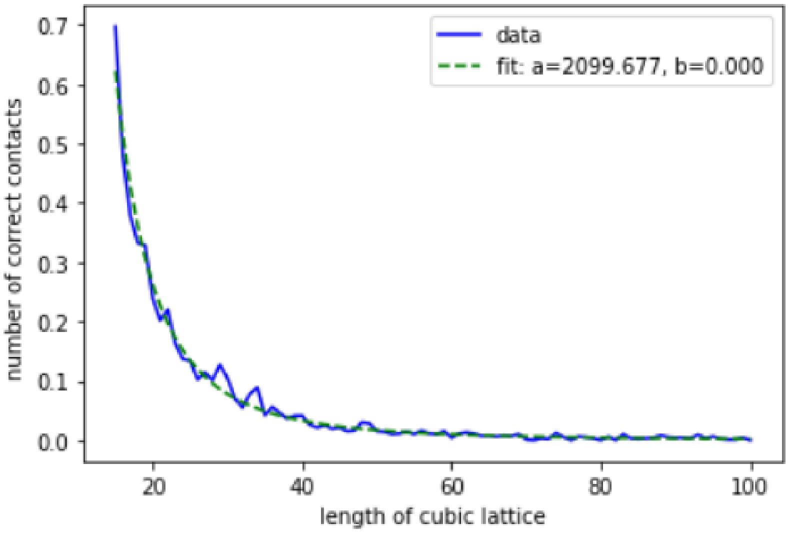


Fig 5.1 Fitting number of correct contacts against the length of cubic lattice

Since the number of correct contacts depends on the number of lattice points available , or in other other words, the volume of the lattice box, the functional dependency of number of correct contacts(y) varies inversely w.r.t the cube of length of cubic lattice box(x) .Thus, we take the function, $y = a/(x^3 + b)$, the values of a and b obtained by the $curve\_fit$ function are $a = 2099.677$, $b = 0$

The Mean Square error between the data and the fit ,

$MSE = 0.00021203504888841875$

5.2 Variation of number of the contacts w.r.t. Length of the SARW

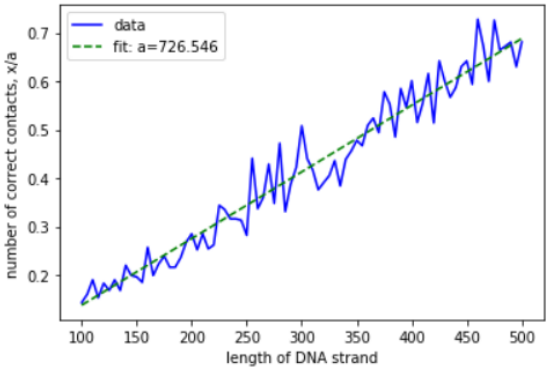


Fig 5.2 Fitting number of correct contacts

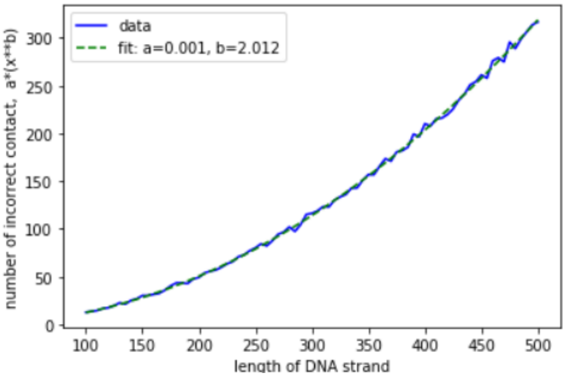


Fig 5.3 Fitting number of incorrect contacts

Both the number of correct contacts and incorrect contacts increase with the length of DNA .the number of correct contacts(n_{CC}) varies linearly w.r.t the length of DNA(L) .Thus, we take the function, $n_{CC} = L/a + b,$ the constant term(b) of this function will be 0 as there can be no contact when the length of DNA(L) is zero. the value of a obtained by the $curve\_fit$ function is $a = 726.546$.

The number of incorrect contacts(n_{INC}) rises faster than n_{CC} w.r.t the length of DNA(L). Thus, we take the function, $n_{INC} = a(x^b),$ the values of a and b obtained by the $curve\_fit$ function are $a = 0.001, b = 2.012$

The Mean Square error between the data and the fit ,

$MSE_{correct\ contacts} = 0.0014451941274426212$

$MSE_{incorrect\ contacts} = 628.860157972877$

5.3 Variation of correct contact probability wrt to the length of SARW

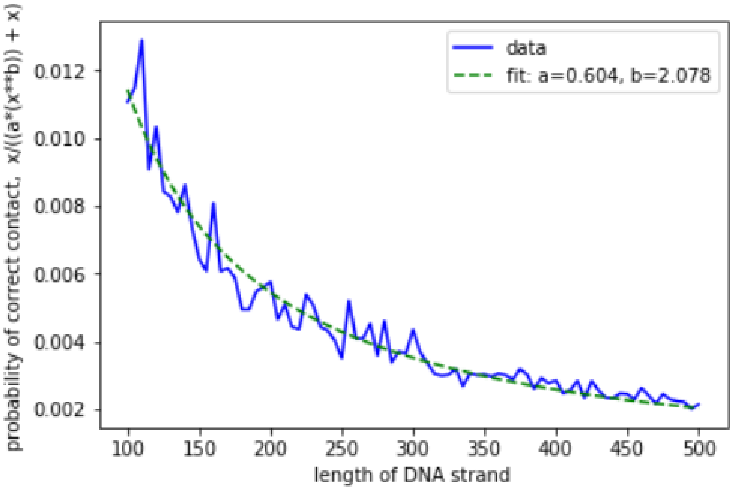


Fig 5.4 Fitting probability of correct contact against the length of DNA strand(L)

Based on functional dependency of n_{CC} and n_{INC} on L. we can take the ratio n_{CC}/ (n_{CC}+n_{INC}) to find P_{CC}.In this section, we generate a fit for the curve of P_{CC} to verify the results obtained by taking ratio of n_{CC} and total contacts n_{TC}. The probability of correct contacts reduces as the length of interacting SARWs increases in a fixed lattice box. The best fit is obtained when we take the relationship between probability of correct contacts(y) and the length of DNA strand(x) is taken as $y = x/(ax^b + x)$,the value of a obtained by $curve\_fit$ function are $a = 0.604, b = 2.078$

$MSE = 0.004153335620956$

5.4 Variation of partial contact probability wrt to complexity of the DNA

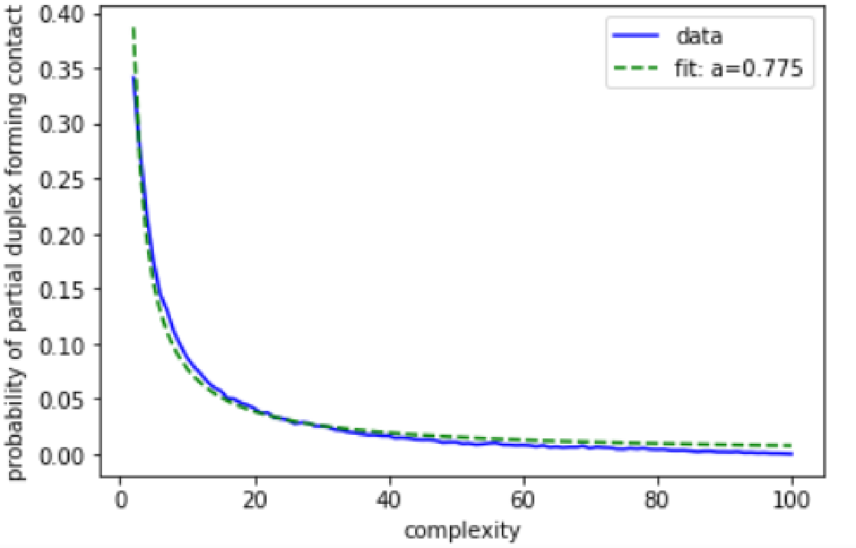


Fig 5.5 Fitting probability of partial contact against the complexity of DNA strand(c)

The probability of incorrect contacts resulting in formation of a partial duplex(partial contacts) varies asymptotically with c. The best fit is obtained when we take the dependency of probability of partial contact(y) on the complexity (x) as $y = a/x$,the value of a obtained by $curve\_fit$ function is $a = 0.775$.

The Mean Square error between the data and the fit ,

$MSE = 6.721821862285786e - 05$

6. CONCLUSION

Based on the results and analysis of the research ,various conclusions can be drawn about the functional dependency of probabilities of contacts on several parameters and consequently the dependency of rate the nucleation and zipping on those parameters.

The first significant conclusion is that the probability of correct contacts (P_{CC}) decreases as the length of DNA(L) increases. The probability is inversely proportional to the square root of length of DNA, other parameters kept constant. Based on our research we found $P_{cc} = L/(0.604L^{2.078} + L)$. Since the nucleation rate is directly proportional to this probability , we can conclude that the nucleation rate(k_n)also varies with L in similar fashion.

$$k_n \propto L/(0.604L^{2.078} + L)$$

The second important conclusion is that the probability of partial contacts ($p_{partial}$) scales inversely with the length on repeating sequence , that is the complexity(c) of the DNA. Based on our research we found that $p_{partial} = 0.775/c$. Since this probability determines the zipping rate(k_Z), we can conclude that the zipping rate is directly proportional to the number of repeats , that is the copy number(ρ) of the DNA. In essence,

$$k_z \propto L/c$$

Since the overall renaturation rate of the DNA depends on the nucleation and zipping rate i.e. the overall rate is directly proportional to nucleation rate as well as the zipping rate. Hence the final conclusion is that overall rate is directly proportional to the square root of length of DNA and inversely proportional to the complexity of the DNA.

$$\boxed{k_{renaturation} \propto L^2/\{c(0.604L^{2.078} + L)\}}$$

7. APPENDIX

1. SARW : A self-avoiding random walk is a sequence of moves on a lattice (a lattice path) that does not visit the same point more than once. This is a special case of graph theoretical notion of a path. In computational physics, a self-avoiding random walk is a chain-like path in $R^2 or\ R^3$ with a certain number of nodes, typically a fixed step length and has the property that it doesn’t cross itself or another walk. A system of SARWs satisfies the so-called excluded volume condition. In higher dimensions, the SARW is believed to behave much like the ordinary random walk.

2. MSE : In statistics, The mean squared error or MSE of an estimator measures the average of the square of errors. That is, it sums the square of differences between the actual value(data) and predicted value(fit), and averages it out over the total number of data points.

$$MSE\ =\ (1/n)\ \Sigma_n (actual_i - predicted_i)^2$$

REFERENCES

- Theory on the Mechanism of DNA Renaturation:Stochastic Nucleation and Zipping- Niranjani G, Murugan R (2016)
- A lattice polymer study of DNA renaturation dynamics- A.Ferrantini, M.Baiei, E.Carlon
- https://en.wikipedia.org/wiki/Self-avoiding_walk