THE RATE OF SITE-SPECIFIC DNA-PROTEIN INTERACTIONS IN THE PRESENCE OF MOLECULAR CROWDING

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

DUAL DEGREE (B.TECH & M.TECH)
IN
ELECTRICAL ENGINEERING

Submitted by

RAJAN RAJWANSH
EE16B146

Under the supervision of

Guide: Dr. Rajamanickam Murugan
Co- Guide: Dr Boby George

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

JUNE 2021

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

CANDIDATE'S DECLARATION

I, Rajan Rajwansh, Roll No - EE16B146, student of Dual Degree (Electrical Engineering), hereby declare that the Project Dissertation titled "The rate of site-specific DNA-protein interactions in the presence of molecular crowding" which is submitted by me to the Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu - 600036 in partial fulfillment of the requirement for the award of the degree of Dual Degree (B.Tech & M.Tech) in Electrical Engineering, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship or other similar title or recognition.

Place: Bokaro Steel City — Rajan Rajwansh

Date: 17.06.2021 — EE16B146

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

CERTIFICATE

I hereby certify that the Project Dissertation titled "The rate of site-specific DNA-protein interactions in the presence of molecular crowding" which is submitted by Rajan Rajwansh, Roll No - EE16B146, student of Dual Degree (Electrical Engineering), Indian Institute of Technology Madras, Chennai, Tamil Nadu - 600036, in partial fulfillment of the requirement for the award of the degree of Dual Degree (B.Tech & M.Tech) in Electrical Engineering, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Chennai

Date: 17.06.2021

Guide Professor: Dr Rajamanickam Murugan

Co - Guide Professor: Dr Boby George

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

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Professor Dr Rajamanickam Murugan and Professor Dr Boby George for their continuous guidance and mentorship that they provided me during the project. They showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its research relevance. They were always ready to help me and clear my doubts regarding any hurdles in this project. Without their constant support and motivation, this project would not have been successful.

Place: Bokaro Steel City — Rajan Rajwansh

Date: 17.06.2021 — EE16B146

Abstract

We develop a lattice model of site-specific DNA-protein interactions under *in vivo* conditions where the DNA is modelled as a self-avoiding random walk that is embedded in a cubic lattice box resembling the living cell. The protein molecule searches for its cognate site on the DNA via a combination of three dimensional and one dimensional random walks. Results show that an optimal one dimensional sliding length exists for smaller cellular volumes at which the search is most efficient. However such an optimal sliding length is not observed for larger cellular volumes. The number of random walk steps increases and the search efficiency reduces with increasing degrees of molecular crowding (increasing number of passive molecules moving randomly inside the cell). The reduction in efficiency is significantly greater for larger cellular volumes as compared to smaller cellular volumes.

Contents

List of Tables

List of Figures

List of Symbols, Abbreviations

- DNA: Deoxyribonucleic Acid
- 1D: One Dimensional
- 3D: Three Dimensional
- DBP: DNA binding protein
- SARW: Self-avoiding Random Walk
- RW: Random Walker
- L: Sliding Length
- $L_{optimal}$: Optimal sliding Length
- b: Side length of the cubic lattice box (The box represents the living cell)
- V : Volume of the lattice box, Volume of the cell (b^3)
- N: Size of the Self Avoiding Random Walk. Each random walk step occupies one cubic unit of the lattice box.
- U: The position of the Random Walker on the DNA $(0 \leq U \leq N)$
- U_S: Location of the specific binding site on the SARW DNA $(1 \leq U_S \leq N - 1)$
- RMSE: Root Mean Squared Error

Chapter 1: Introduction

Protein-DNA interactions occur when a protein binds a molecule of DNA, often to regulate the biological function of DNA, usually the expression of a gene. Protein-DNA interactions are mainly of two types, either specific interaction, or non-specific interaction. In site-specific interactions, the DNA-binding protein (DBP) binds to its specific site on the DNA molecule in the presence of a large number of nonspecific sites. The DBP searches for its specific site on the DNA molecule inside the cell via a combination of three-dimensional (3D) and one-dimensional (1D) diffusion. This is a two-step process where the DBP first non-specifically binds to the DNA molecule via 3D diffusion and then searches for its specific site on the DNA via 1D sliding.

The maximum number of sliding steps allowed (sliding length - L) plays an important role in determining the search efficiency of the DBP. A smaller sliding length increases 3D diffusion and decreases 1D sliding whereas a larger sliding length increases 1D sliding and decreases 3D diffusion. Increase in cellular volume negatively affects the overall efficiency of the search process. In our simulations, we study the effect of sliding length (L) on the search efficiency of the DBP and try to find if an optimal sliding length $(L_{optimal})$ exists at which

maximum search efficiency is achieved. This complete analysis is done over a range of varying cellular volumes.

The overall search efficiency is also negatively affected by the presence of other randomly moving molecules inside the cell. We analyze the effect of such molecular crowding on the search efficiency by incorporating an increasing number of randomly moving passive molecules inside the cell. This entire analysis is also done over a range of varying cellular volumes.

Chapter 2: Model Overview

We consider a cubical lattice box with side length b dimensionless units as the model for our living cell. Therefore, the cellular volume is $V = b^3$ dimensionless cubic units. The DNA is modeled as a linear lattice of size $\sqrt{3}N$ dimensionless units inside the cell volume (The DNA is composed of N steps, each step is the longest diagonal of a unit cube, i.e. each step has length $\sqrt{3}$ dimensionless units). We introduce a random walker (RW) confined within the lattice box as a model for the randomly moving DBP inside the living cell. Let $0 \leq U \leq N$ denote the position of the RW protein molecule on the SARW DNA lattice. The specific site is located at $U_S \equiv (X_S, Y_S, Z_S)$ with $1 \leq U_S \leq N - 1$.

The simulation begins with the RW protein molecule starting at a random initial position (X_0, Y_0, Z_0). The RW protein molecule, via a combination of 3D diffusion and 1D sliding, finds its specific site on the DNA lattice. When the RW hits the DNA lattice at a location other than the specific binding site, then it subsequently performs a 1D random walk along the DNA for a fixed L (sliding length) number of steps after which it dissociates and resumes 3D random walk. The simulation stops when the RW protein molecule reaches its specific binding site. The number of 1D and 3D random walk steps are computed. Their sum (Total Steps)

is the measure by which we analyse the efficiency of the search process. All the simulation results presented are averages over 10^5 iterations.

In our first set of simulations, while keeping b constant (cellular volume is kept constant), we change the sliding length (L) and try to find if an optimal sliding length $(L_{optimal})$ exists, at which the Total Steps for the search is minimized. The same procedure is repeated over a range of varying cellular volumes.

Several other molecules may be present and be performing random motion inside the cell (molecular crowding), and so this may impact the search efficiency of the RW protein molecule inside the cell. Through our second set of simulations we try to better understand this relationship. Keeping cellular volume constant, the sliding length is set to $L_{optimal}$ (obtained from the previous set of simulations). Next, an increasing number of randomly moving passive molecules are introduced inside the cell along with the RW protein molecule. These passive molecules move randomly in a similar fashion to the RW protein molecule.

All the molecules perform 3D random motion in accordance with the **excluded volume effect**. Thus, any two molecules do not occupy the same location (X, Y, Z) during 3D diffusion. If during this collective random motion, any two molecules which are in 1D sliding motion on the DNA lattice happen to collide,

the collision results in dissociation and both molecules switch over to 3D diffusion from that point. The impact of increasing degrees of molecular crowding on the Total Steps is analyzed. The same procedure is repeated over a range of varying cellular volumes.

Chapter 3: Simulation Methods

3.1 To find the optimal sliding length in the case of a single protein molecule

In our stochastic simulations, we have defined a cubical box of size $b = 10$ over $(X, Y, Z) \in [0, b]$ and generated a self avoiding random walk (SARW) of size $N = 100$ within this box boundaries using the inverse restricted sampling method (The Rosenbluth-Rosenbluth Algorithm). This SARW represents the DNA lattice. The specific site of the protein molecule was positioned at $U_S \equiv (X_S, Y_S, Z_S)$ with $1 \leq U_S \leq N - 1$. The initial position of the random walker (RW) protein molecule is at (X_0, Y_0, Z_0). For each simulation, the starting point of SARW, the location of the binding site on the SARW DNA (U_S), and the starting point of the random walker protein (X_0, Y_0, Z_0) were chosen randomly.

Let us assume that the current position of the protein molecule is (X, Y, Z). The next 3D random walk movement of the RW protein will be decided in the following manner:

- If $X + 1 \leq b$ then $(X + 1, Y, Z)$ is a point inside the cubic lattice and movement in $+ X$ $(positive\ X)$ direction is feasible.

- If $X - 1 \geq 0$ then $(X - 1, Y, Z)$ is a point inside the cubic lattice and movement in $-X$ $(negative\ X)$ direction is feasible.
- If $Y + 1 \leq b$ then $(X, Y + 1, Z)$ is a point inside the cubic lattice and movement in $+Y$ $(positive\ Y)$ direction is feasible.
- If $Y - 1 \geq 0$ then $(X, Y - 1, Z)$ is a point inside the cubic lattice and movement in $-Y$ $(negative\ Y)$ direction is feasible.
- If $Z + 1 \leq b$ then $(X, Y, Z + 1)$ is a point inside the cubic lattice and movement in $+Z$ $(positive\ Z)$ direction is feasible.
- If $Z - 1 \geq 0$ then $(X, Y, Z - 1)$ is a point inside the cubic lattice and movement in $-Z$ $(negative\ Z)$ direction is feasible.

The new location of the protein molecule is decided by taking a step in a random feasible direction after checking for the possible feasible directions using the above six steps. The next location of the protein molecule is checked against the coordinates of the SARW at each step. If there is a match that resembles a lattice point on the SARW DNA, then the protein molecule undergoes a 1D random walk over the DNA lattice for L(sliding length) number of steps. Let $U_i \equiv (X_i, Y_i, Z_i)$ be the current position of the RW protein molecule on the SARW DNA.

The next position of the RW protein molecule on the DNA lattice is obtained in the following manner:

- When the current position of the RW protein molecule is not one of the terminals of the DNA lattice:
    - We generate a random number r that takes the value $0\ or\ 1$.
    - If $r = 0$, then the RW takes a step in the backward direction along the DNA lattice, i.e. $U_i \equiv (X_i, Y_i, Z_i) \rightarrow U_{i-1} \equiv (X_{i-1}, Y_{i-1}, Z_{i-1})$.
    - If $r = 1$, then the RW takes a step in the forward direction along the DNA lattice, i.e. $U_i \equiv (X_i, Y_i, Z_i) \rightarrow U_{i+1} \equiv (X_{i+1}, Y_{i+1}, Z_{i+1})$.

- When the protein molecule reaches the terminals of the DNA lattice:
    - With probability 0.5 the RW protein dissociates and switches over to 3D random walk.
    - With probability 0.5 the RW protein returns back to the DNA lattice. That is, if the RW protein is at the forward end $(i = N)$, then it takes a backward step $U_i \equiv (X_i, Y_i, Z_i) \rightarrow U_{i-1} \equiv (X_{i-1}, Y_{i-1}, Z_{i-1})$. If it is at the backward end $(i = 0)$, then it takes a forward step $U_i \equiv (X_i, Y_i, Z_i) \rightarrow U_{i+1} \equiv (X_{i+1}, Y_{i+1}, Z_{i+1})$.

As the entire SARW DNA $(U_i,\ 0 \leq i \leq N)$ is stored in memory, the next 1-D sliding step can be obtained by direct memory access. The simulation stops when the RW protein molecule reaches the location $U_S \equiv (X_S, Y_S, Z_S)$. The

average of the total number of 3D and 1D steps over several trajectories gives us the required Total Steps as the measure of search efficiency. Here the main parameters which decide the Total Steps are:

- Volume of the cell ($V = b^3$)
- Length of the SARW DNA lattice (N)
- Sliding Length (L)

We study the influence of the sliding length (L) on the Total Steps over a varying range of cellular volumes ($V = b^3$) and try to find the existence of an optimal sliding length ($L_{optimal}$) at which the Total Steps are minimized in each of those cases.

3.2 To study the effect of molecular crowding

The simulation method is similar to the one described above although with some changes to accommodate for multiple randomly moving passive molecules inside the cell. The passive molecules inside the cell move in the exact same manner as the RW protein molecule, i.e. they perform 3D diffusion along with 1D sliding. There is only one binding location on the DNA lattice, which is the specific site for the RW protein molecule. However, in these simulations, since we are trying to analyse the impact of molecular crowding, we keep the sliding length (L) constant for a particular cellular volume. The sliding length for a particular cellular

volume is set to the $L = L_{optimal}$ where the $L_{optimal}$ is obtained from the previous set of simulations for that particular cellular volume. All the molecules collectively move inside the cell via a combination of 3D diffusion and 1D sliding. While in 3D diffusion, the molecules move randomly in accordance with the Excluded Volume Effect.

Let us assume that the current position of a molecule is (X, Y, Z). The next 3D random walk step of the molecule will be decided in the following manner:

- If $X + 1 \leq b$ and if $(X + 1, Y, Z)$ is not currently occupied by another molecule then $(X + 1, Y, Z)$ is a is a possible next location and movement in $+ X$ $(positive\ X)$ direction is feasible.
- If $X - 1 \geq 0$ and if $(X - 1, Y, Z)$ is not currently occupied by another molecule then $(X - 1, Y, Z)$ is a is a possible next location and movement in $- X$ $(negative\ X)$ direction is feasible.
- Similarly we can check if the movements in $+ Y$ $(positive\ Y)$, $- Y$ ($negative\ Y$), $+ Z$ $(positive\ Z)$ and $- Z$ $(negative\ Z)$ directions are feasible.
- The new location of the molecule is decided by taking a step in a random feasible direction after checking for the possible feasible directions using the steps mentioned above.

The next location of the protein molecule is checked against the coordinates of the SARW at each step. If there is a match that resembles a lattice point on the SARW DNA, then the protein molecule undergoes a 1D random walk over the DNA lattice for L(sliding length) number of steps in a similar manner to as described before. However, there is the following key modification:

- If during simulation any two molecules (let's say molecule X and molecule Y) encounter each other during 1D sliding random walk, then they both dissociate and switch to 3D random walk from that point after the collision.

The simulation stops when the RW protein molecule reaches the location $U_S \equiv (X_S, Y_S, Z_S)$. The average of the total number of 3D and 1D steps over several trajectories gives us the required Total Steps as the measure of search efficiency. Here the main parameters which decide the Total Steps are:

- Volume of the cell $(V = b^3)$
- Length of the SARW DNA lattice (N)
- Degree of molecular crowding (Number of passive molecules)

We study the influence of the number of passive molecules on the Total Steps over a range of varying cellular volumes (V).

Chapter 4: Results and Discussion

4.1 The optimal sliding length in the case of a single protein molecule

Through the first set of simulations, we tried to find the existence of an optimal sliding length for which the Total Steps is minimum for a given cellular volume (V). Results are obtained over a range of varying cellular volumes. A part of the simulation results is presented in the tables below.

4.1.1 Results for smaller cellular volumes

We present the results for smaller cellular volumes. The results list the number of 1D sliding steps, the number of 3D diffusion steps and the total number of steps that the protein molecule took to find its specific binding location, as the sliding length was increased gradually. The values obtained are averages over 10^5 iterations. The tables and figures presented below are for the following cellular volumes:

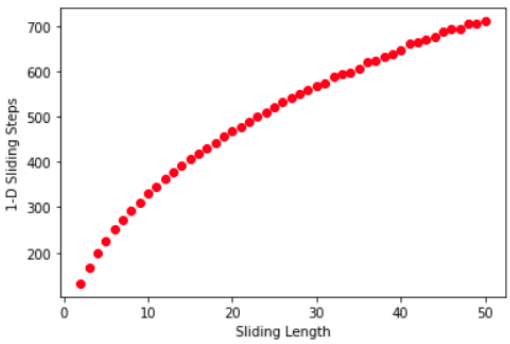
- $V = 1000\ cubic\ units\ (b = 10\ units)$
- $V = 8000\ cubic\ units\ (b = 20\ units)$

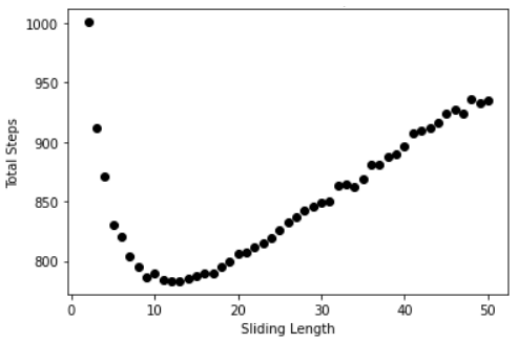
Sliding Length	1-D Steps	3-D Steps	Total Steps
0	0	2060	2060
1	83	1081	1164
3	168	744	912
5	225	605	830
7	273	531	804
9	311	475	786
11	346	438	784
13	377	406	783
15	406	381	787
17	429	360	789
19	456	343	799
21	478	329	807
23	499	316	815
25	520	306	826
27	541	296	837
29	559	287	846
31	573	277	850
33	593	271	864
35	606	263	869
37	624	257	881
39	639	251	890

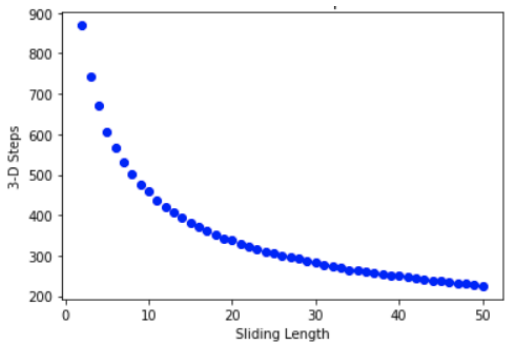
Table 1.1: Optimal Sliding Length $(b = 10, V = 1000)$

Sliding Length	1-D Steps	3-D Steps	Total Steps
0	0	14446	14446
10	342	3774	4116
30	610	2580	3190
50	776	2165	2941
70	906	1945	2851
90	1002	1799	2801
110	1082	1691	2773
130	1150	1600	2750
150	1211	1541	2752
170	1255	1489	2744
190	1312	1451	2763
210	1346	1408	2754
230	1377	1370	2747
250	1411	1348	2759
270	1441	1320	2761
290	1468	1300	2768
310	1493	1281	2774
330	1521	1264	2785
350	1536	1254	2790
370	1565	1234	2799
390	1573	1213	2786

Table 1.2: Optimal Sliding Length $(b = 20)$

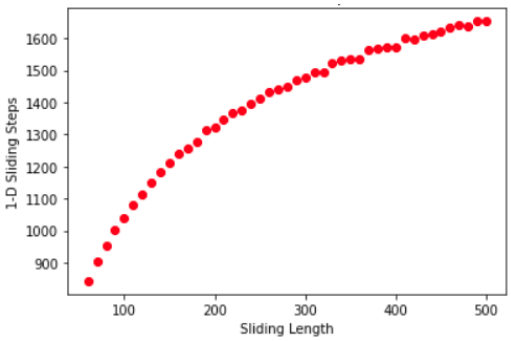


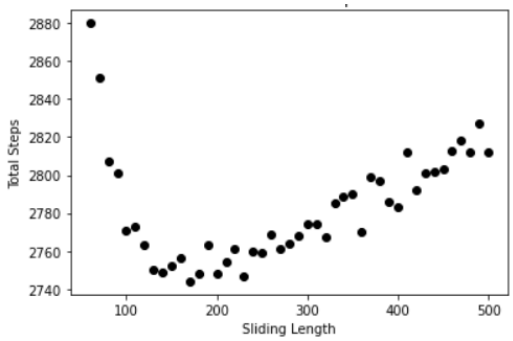


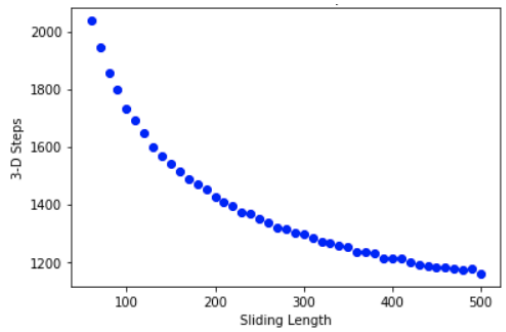


1-D Steps Total Steps 3-D Steps

Figure 1.1: Optimal Sliding Length ($b = 10$)







1-D Steps Total Steps 3-D Steps

Figure 1.2: Optimal Sliding Length ($b = 20$)

4.1.2 Results for larger cellular volumes

For larger cellular volumes we did not observe the existence of an optimal sliding length. We performed simulations with with cellular volumes $27000\ cubic\ units$ $(b = 30\ units)$, $64000\ cubic\ units$ $(b = 40\ units)$ and $125000\ cubic\ units$ $(b = 50\ units)$. The total number of steps kept on decreasing as the sliding length increased and a clear minimum number of total steps was not observed in each of these cases.

4.1.3 Discussion and Analysis

The limiting values of Total number of steps for each of these cellular volumes was obtained by setting the sliding length to infinity (actually set to INT_MAX = 2,147,483,647 in our simulations). These values are presented below:

b(Side of cubic cell)	$V = b^3$ (Cellular Volume)	1D Sliding Steps	3D Diffusion Steps	Total number of Steps
10	1000	1900	65	1965
20	8000	1991	960	2951
30	27000	1973	3771	5744
40	64000	1948	9342	11290
50	125000	1935	18541	20476

Table 2: Limiting Values of Total Steps with Sliding Length approaching infinity

From the observations in the Table 1.1 and Table 1.2 we can see that the optimal sliding lengths for $b = 10$ and $b = 20$ occur at sliding lengths 13 and 170 respectively. In both of these cases, we can see that the total number of steps consists of an approximately equal number of 1D and 3D steps. Thus we can say that the optimal sliding length occurs when the RW protein molecule has an equal number of 1D and 3D steps in the search process. This explains why we do not observe an optimal sliding length for larger cellular volumes $(b = 30,\ b = 40\ and\ b = 50)$.

When the sliding length is small, there is very little 1D sliding and 3D steps > 1D steps irrespective of the cellular volume. However as the sliding length approaches infinity, 1D steps > 3D steps for smaller cellular volumes, while 3D steps > 1D steps for larger cellular volumes. Therefore, a switch from a predominant 3D mode of diffusion to a predominant 1D mode of diffusion is observed for smaller cellular volumes, while such a switch does not happen in larger cellular volumes. So, the criteria for an optimal sliding length (1D steps = 3D steps) is never satisfied for larger cellular volumes and thus an optimal sliding length is not observed in those cases.

4.2 The effect of molecular crowding

Through the second set of simulations we studied the effect of molecular crowding on the Total Steps taken by the RW protein molecule to find its specific binding site. The results were obtained over a varying range of cellular volumes.

4.2.1 Tables and Figures

The sliding length for a particular cellular volume setting was set to the optimal sliding length obtained from the previous set of simulations. In settings where the optimal sliding length did not exist the sliding length was set to 1000. The number of passive molecules was varied from 0 to 20 and the number of 1D Steps, the number of 3D Steps and the number of Total Steps were computed. The values obtained are averages over 10^5 iterations. The parameters used to generate the tables and plots are listed below:

- $b = 10,\ V = 1000,$

  $Sliding\ Length = Optimal\ Sliding\ Length = 13$

- $b = 20,\ V = 8000,$

  $Sliding\ Length = Optimal\ Sliding\ Length = 170$

- $b = 30,\ V = 27000,\ Sliding\ Length = 1000$

- $b = 40,\ V = 64000,\ Sliding\ Length = 1000$

- $b = 50,\ V = 125000,\ Sliding\ Length = 1000$

Passive molecules	1D Steps	3D Steps	Total Steps
0	377	406	783
1	373	416	789
2	366	422	788
3	364	433	797
4	359	440	799
5	355	449	804
6	351	457	808
7	345	463	808
8	343	474	817
9	338	480	818
10	335	488	823
11	329	494	823
12	328	505	833
13	327	516	843
14	319	518	837
15	317	527	844
16	315	536	851
17	311	542	853
18	309	552	861
19	304	556	860
20	300	562	862

Table 3.1: Effect of molecular crowding
$(b = 10,\ sliding\ length = 13)$

Passive molecules	1D Steps	3D Steps	Total Steps
0	1262	1488	2750
1	1199	1582	2781
2	1144	1676	2820
3	1102	1761	2863
4	1064	1839	2903
5	1024	1906	2930
6	995	1989	2984
7	974	2076	3050
8	935	2105	3040
9	914	2173	3087
10	890	2234	3124
11	865	2286	3151
12	850	2346	3196
13	836	2397	3233
14	810	2435	3245
15	803	2499	3302
16	781	2525	3306
17	769	2564	3333
18	761	2624	3385
19	741	2653	3394
20	739	2719	3458

Table 3.2: Effect of molecular crowding
$(b = 20,\ sliding\ length = 170)$

Passive molecules	1D Steps	3D Steps	Total Steps
0	1817	4052	5869
1	1709	4429	6138
2	1608	4712	6320
3	1533	4993	6526
4	1475	5226	6701
5	1419	5423	6842
6	1362	5602	6964
7	1314	5787	7101
8	1281	5973	7254
9	1239	6120	7359
10	1204	6224	7428
11	1188	6390	7578
12	1162	6498	7660
13	1142	6638	7780
14	1110	6734	7844
15	1096	6846	7942
16	1069	6914	7983
17	1053	7040	8093
18	1032	7093	8125
19	1024	7263	8287
20	1007	7309	8316

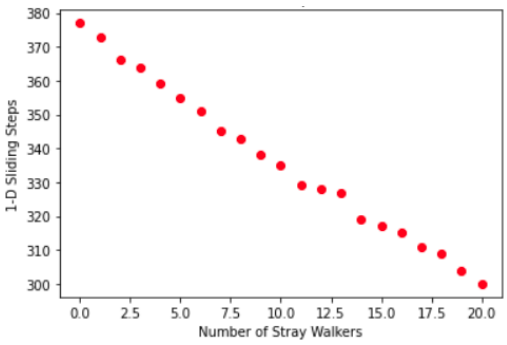
Table 3.3: Effect of molecular crowding
$(b = 30,\ sliding\ length = 1000)$

Passive molecules	1D Steps	3D Steps	Total Steps
0	1802	10031	11833
1	1729	10522	12251
2	1676	10905	12581
3	1618	11356	12974
4	1564	11628	13192
5	1536	12047	13583
6	1501	12265	13766
7	1457	12494	13951
8	1444	12808	14252
9	1402	13053	14455
10	1386	13303	14689
11	1361	13440	14801
12	1334	13678	15012
13	1311	13854	15165
14	1297	14001	15298
15	1278	14311	15589
16	1258	14446	15704
17	1244	14688	15932
18	1226	14779	16005
19	1207	14890	16097
20	1201	15119	16320

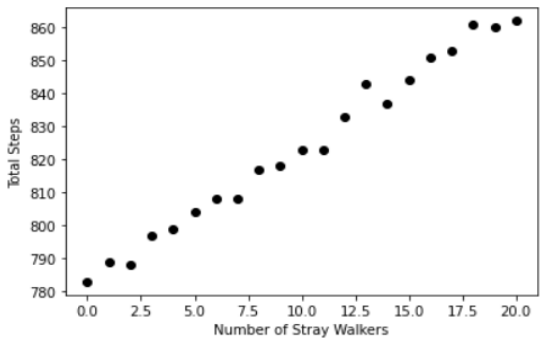
Table 3.4: Effect of molecular crowding
$(b = 40,\ sliding\ length = 1000)$

Passive molecules	1D Steps	3D Steps	Total Steps
0	1778	19674	21452
1	1750	20400	22150
2	1712	20919	22631
3	1670	21334	23004
4	1627	21778	23405
5	1614	22214	23828
6	1578	22693	24271
7	1561	22990	24551
8	1538	23308	24846
9	1514	23722	25236
10	1490	23950	25440
11	1473	24177	25650
12	1456	24616	26072
13	1435	24760	26195
14	1429	25114	26543
15	1411	25362	26773
16	1393	25549	26942
17	1377	25968	27345
18	1361	26026	27387
19	1353	26269	27622
20	1333	26700	28033

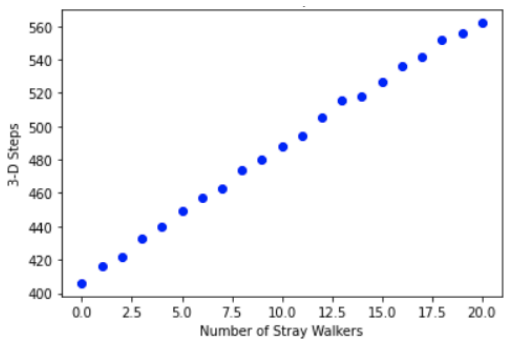
Table 3.5: Effect of molecular crowding
$(b = 50, \ sliding\ length = 1000)$



1-D Steps

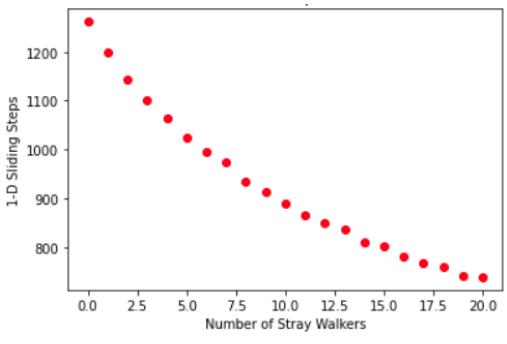


Total Steps

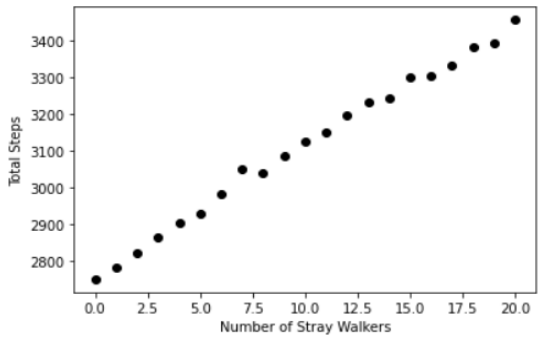


3-D Steps

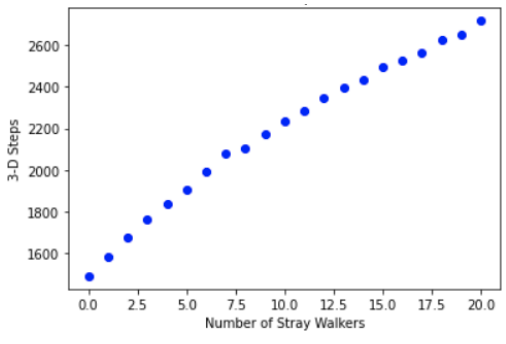
Figure 2.1: Effect of molecular crowding $(b = 10,\ sliding\ length = 13)$



1-D Steps

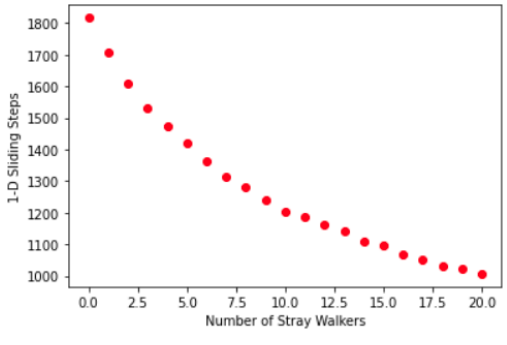


Total Steps

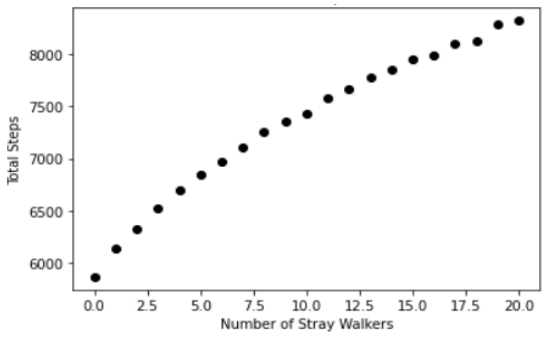


3-D Steps

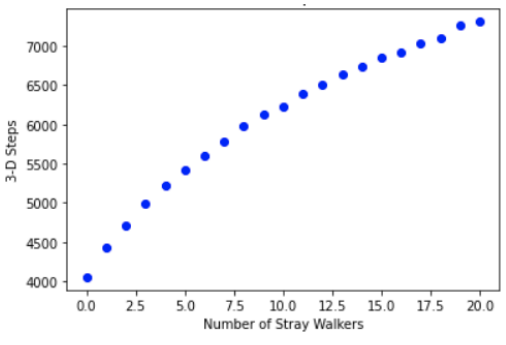
Figure 2.2: Effect of molecular crowding $(b = 20,\ sliding\ length = 170)$



1-D Steps

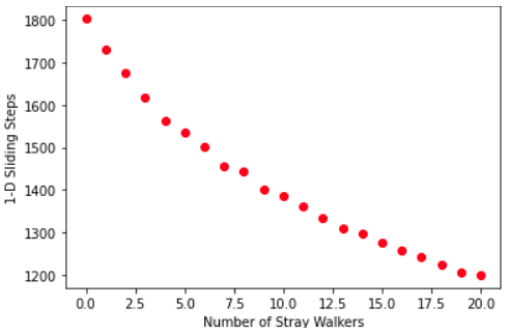


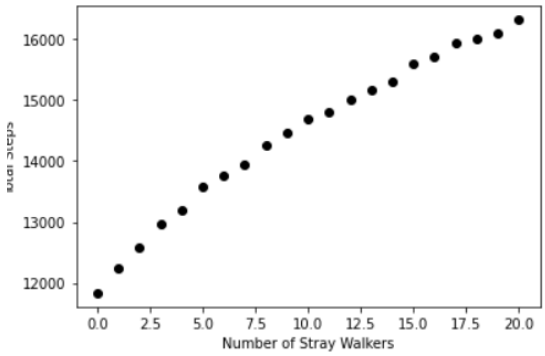
Total Steps

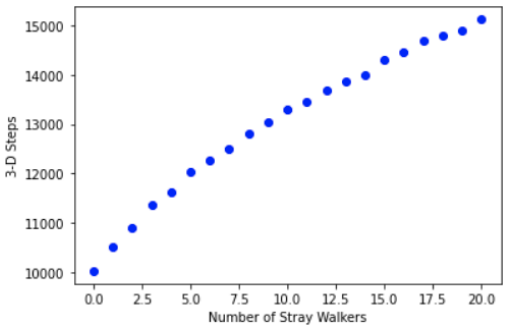


3-D Steps

Figure 2.3: Effect of molecular crowding $(b = 30,\ sliding\ length = 1000)$

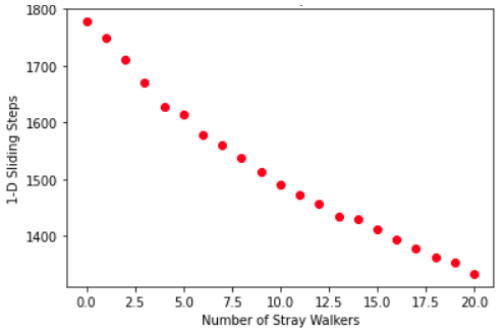


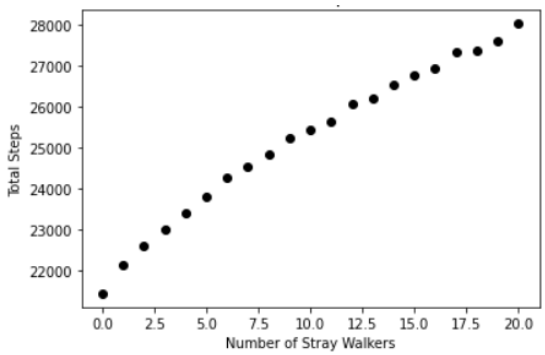


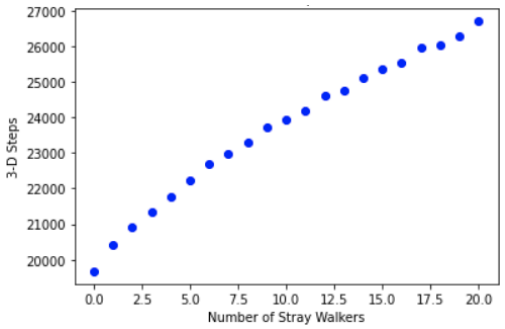


1-D Steps Total Steps 3-D Steps

Figure 2.4: Effect of molecular crowding ($b = 40$, $sliding\ length = 1000$)







1-D Steps Total Steps 3-D Steps

Figure 2.5: Effect of molecular crowding ($b = 50$, $sliding\ length = 1000$)

4.2.2 Discussion and Analysis

Several conclusions about the effect of molecular crowding on the rate of site-specific binding of the protein molecule can be drawn from the results presented in the tables and figures above:

- As the number of passive molecules increases the number of 1D sliding steps decreases. According to our model a collision of the protein molecule with another passive molecule while both are in 1D sliding motion on the DNA lattice results in dissociation, and both the protein molecule and the passive molecule continue 3D diffusion from that point after collision. Due to collisions, the RW protein molecule is unable to slide on DNA lattice for L (sliding length) steps and thus its effective sliding length is reduced. The frequency of collisions increases as the number of passive molecules inside the cell increases, reducing the effective sliding length. Thus, the number of 1D steps decreases as the number of passive molecules inside the cell increases.

- The number of 3D steps, on the other hand, increases with the increase in the number of passive molecules. This increasing trend can be attributed to two factors:
    - Firstly, due to an increase in the frequency of collisions due to increase in molecular crowding, the frequency of dissociation

increases and the RW protein molecule spends more time in 3D diffusion as the number of passive molecules inside the cell increases.

    - Secondly, during 3D diffusion with molecular crowding the molecules move randomly in accordance with the Excluded Volume Effect. Thus, the 3D random walk of the RW protein molecule gets more and more constrained as the number of passive molecules inside the cell increases, increasing the number of 3D steps.

- The total number of steps shows an increasing trend with an increasing number of passive molecules. This is because the decrease in the number of 1D Steps is offset by the increase in the number of 3D Steps. The rate of increase in the number of 3D Steps is more than the rate of decrease in the number of 1D Steps, therefore the total number of steps increases with increasing number of passive molecules. The increase in Total Steps is more prominent in larger volumes as 3D diffusion is the predominant mode of motion in those cases.

4.2.3 Least squares fitting

We use least squares fitting to try to obtain a functional relation between the number of passive molecules and the number of total steps taken by the RW protein molecule to find its specific-site (separately for different cellular volumes). We use a quadratic function ($f(x) = mx^2 + nx + p$) to obtain the least squares fit. Let the number of passive molecules be the dependent variable x here and $f(x)$ is the function that gives us the total number of steps for a particular cellular volume. The following functional parameters were obtained after performing the least squares fit.

Side Length - b	Cellular Volume - V	LSTSQ fit Parameter - m	LSTSQ fit Parameter - n	LSTSQ fit Parameter - p
10	1000	0.0047	4.0298	782.9198
20	8000	-0.3210	41.0354	2744.7053
30	27000	-3.5346	186.2760	5958.8707
40	64000	-5.6247	325.8975	11959.7403
50	125000	-7.1483	449.8828	21686.4867

Table 4: Least Squares Fit Parameters

The figures below show the least squares fit obtained for different cellular volumes. The fit is shown as a solid blue line while the data points are shown as black dots.

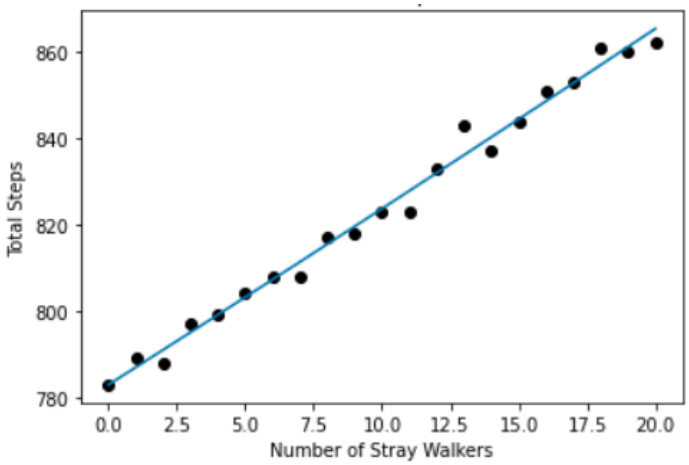


Figure 3.1: Least Squares Fit for Total number of steps $(b = 10)$

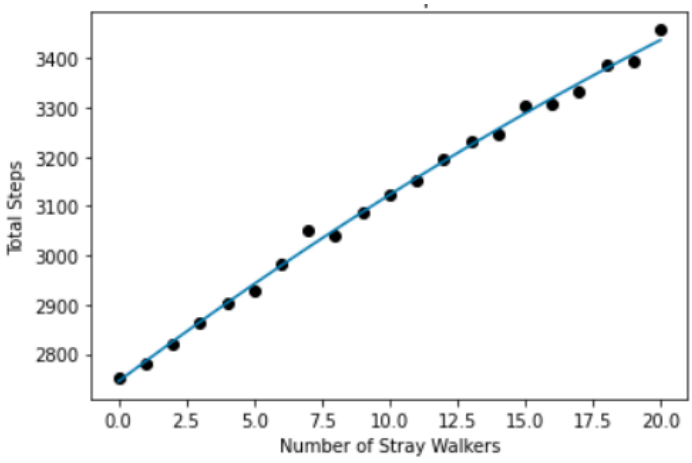


Figure 3.2: Least Squares Fit for Total number of steps $(b = 20)$

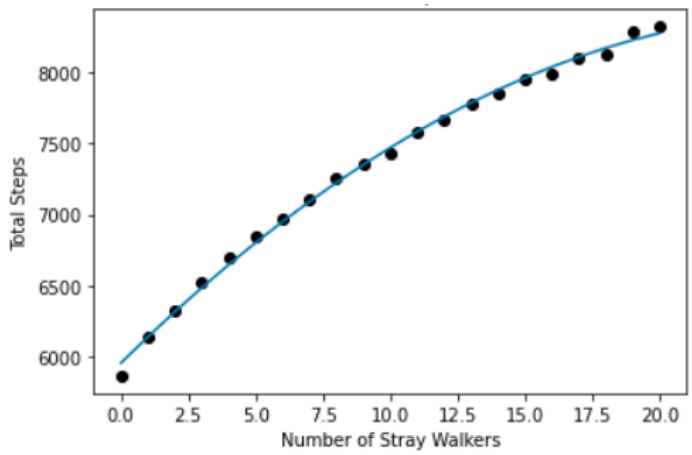


Figure 3.3: Least Squares Fit for Total number of steps $(b = 30)$

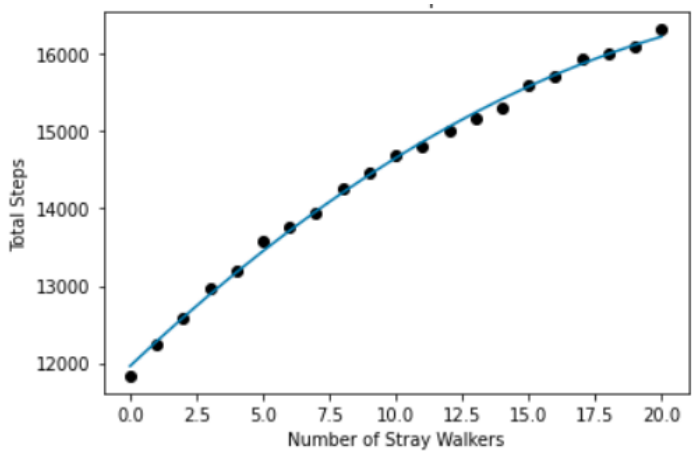


Figure 3.4: Least Squares Fit for Total number of steps $(b = 40)$

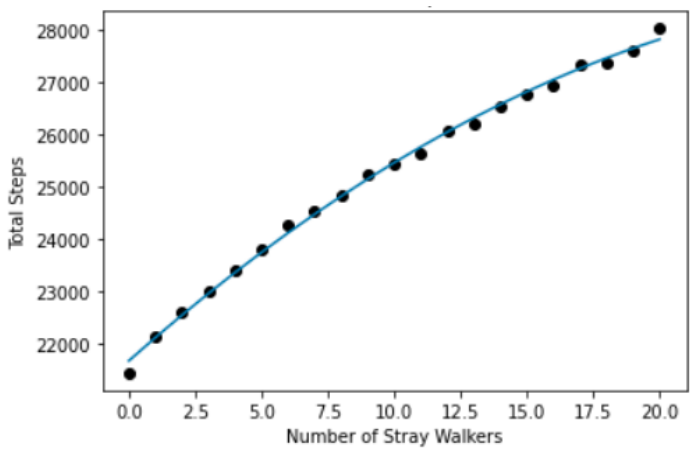


Figure 3.5: Least Squares Fit for Total number of steps $(b = 50)$

4.2.4 Error Analysis

The standard errors for the Least Squares Fit coefficients obtained are presented in the tables below. The parameter is significant if its confidence level is greater than 95% (i.e. significance level$< 5\%$ or p-value< 0.05).

Parameter	Optimal Value	Standard Error	p-Value	Confidence Level	Is Parameter Significant
m	0.0047	0.0194	0.8122	18.78%	No
n	4.0298	0.4024	< 0.0001	> 99.99%	Yes
p	782.9198	1.7366	< 0.0001	> 99.99%	Yes

Table 5.1: Error Analysis $(b = 10)$

Parameter	Optimal Value	Standard Error	p-Value	Confidence Level	Is Parameter Significant
m	-0.3210	0.0882	0.0017	99.83%	Yes
n	41.0354	1.8268	< 0.0001	> 99.99%	Yes
p	2744.7053	7.8839	< 0.0001	> 99.99%	Yes

Table 5.2: Error Analysis $(b = 20)$

Parameter	Optimal Value	Standard Error	p-Value	Confidence Level	Is Parameter Significant
m	-3.5346	0.2716	< 0.0001	> 99.99%	Yes
n	186.2760	5.6261	< 0.0001	> 99.99%	Yes
p	5958.8707	24.2805	< 0.0001	> 99.99%	Yes

Table 5.3: Error Analysis $(b = 30)$

Parameter	Optimal Value	Standard Error	p-Value	Confidence Level	Is Parameter Significant
m	-5.6247	0.4735	< 0.0001	> 99.99%	Yes
n	325.8975	9.8084	< 0.0001	> 99.99%	Yes
p	11959.7403	42.3299	< 0.0001	> 99.99%	Yes

Table 5.4: Error Analysis $(b = 40)$

Parameter	Optimal Value	Standard Error	p-Value	Confidence Level	Is Parameter Significant
m	-7.1483	0.7147	< 0.0001	> 99.99%	Yes
n	449.8828	14.8064	< 0.0001	> 99.99%	Yes
p	21686.4867	63.8996	< 0.0001	> 99.99%	Yes

Table 5.5: Error Analysis $(b = 50)$

Chapter 5: Conclusion

As the sliding length increases the number of 3-D steps decreases and the number of 1-D steps increases. With increasing sliding length the mode of diffusion for the protein molecule changes from predominant 3D diffusion to predominant 1D sliding for smaller cellular volumes. However, 3D diffusion is always the predominant mode of diffusion for larger cellular volumes, irrespective of the sliding length. The optimal sliding length is observed when the number of 1-D steps and the number of 3-D steps are approximately equal. For smaller cellular volumes an optimal sliding length exists for which the total number of steps is at a minimum. But, for larger cellular volumes an optimal sliding length does not exist as the number of 3D Steps is always greater than the number of 1D Steps (irrespective of the sliding length).

The search efficiency of the protein molecule is affected by molecular crowding and the efficiency decreases as the number of passive molecules inside the cell increases. The number of 1D Steps decrease as the effective sliding length decreases due to collisions between the protein molecule and the passive molecules while both are in 1D sliding motion on the surface of the DNA lattice. The number of 3D Steps on the other hand increases with increase in the degree of molecular crowding due to two reasons. First, as the number of passive

molecules inside the cell increases, the number of collisions during 1D sliding motion of the protein increases. This decreases the effective sliding length and increases the frequency of dissociation. Thus, the protein molecule spends more time in 3D diffusion. Secondly, during 3D diffusion with molecular crowding the molecules move randomly in accordance with the Excluded Volume Effect. Due to this the 3D motion of the protein molecule becomes more and more constrained as the number of passive molecules increases and thus the number of 3D steps increases with increasing number of passive molecules. The Total Steps shows an increasing trend with increasing molecular crowding. This is because the rate of increase in the number of 3D Steps is more than the rate of decrease in the number of 1D Steps. Therefore the total number of steps increases with the increasing number of passive molecules. The increase is more prominent in larger cellular volumes as 3D diffusion is the predominant mode of diffusion in those cases.

Appendix

1. **Excluded Volume Effect:** In polymer science, excluded volume refers to the idea that one part of a long chain molecule can not occupy space that is already occupied by another part of the same molecule.

2. **Self Avoiding Random Walk:** In mathematics, 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. 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 excluded volume condition. In higher dimensions, the SARW is believed to behave much like the ordinary random walk.

3. **The inverse restricted sampling method (Rosenbluth-Rosenbluth Algorithm) to generate SARW:**

   Presented below is a figure that illustrates the pseudocode for the inverse restricted sampling method which can be used to generate SARWs effectively.

```
      title Inversely restricted sampling.
      function irsamp(N)
      comment This routine returns an N-step SAW and its weight factor.

      ω_0 ← 0
start: weight ← 1/[2d(2d − 1)^{N−1}]   (this is merely a convenient normalization)
      for i = 1 to N do
          S_i ← set of all nearest neighbors of ω_{i−1} not contained in {ω_0, ..., ω_{i−1}}
          if S_i = ∅ goto start   (the walk is "trapped")
          ω_i ← a random element of S_i
          weight ← weight × |S_i|
      enddo
      return (ω, weight)
```

Figure 4: Inverse Restricted Sampling (Rosenbluth-Rosenbluth Algorithm)

References

1. A lattice model on the rate of *in vivo* site-specific DNA-protein interactions - *R Murugan 2021*
2. Theory of Site-Specific DNA-Protein Interactions in the Presence of Conformational Fluctuations of DNA Binding Domains - *R Murugan 2010*
3. Monte Carlo Methods For The Self-Avoiding Walk - *Alan D. Sokal*