COMPUTATIONAL BIOLOGY IN THE ANALYSIS OF EPIGENETIC NUCLEAR SELF-ORGANIZATION

By

Maria Mohammad Khan

A Thesis Submitted to The Honors College

In Partial Fulfillment of the Bachelors degree
With Honors in

Molecular and Cellular Biology

THE UNIVERSITY OF ARIZONA

May 2010

Approved by:

Dr. Jonathan Flax
Department of Neurology

Introduction

The function of the nucleus is central to the survival of cells and thus life as a whole. Among other processes, it is the site of gene expression, DNA repair, and genome stability. These functions are carried in the context of a complex nuclear architecture. The nucleus is compartmentalized both spatially and functionally. These compartments are proteinaceous nuclear bodies or chromatin domains, both of which are not segregated from other compartments by membranes-as are the organelles of cells. Specifically, proteinaceous nuclear bodies are characterized as regions within the nucleus with distinct sets of inhabitant proteins. Examples of such proteinaceous nuclear bodies include the nucleolus, splicing factor compartments, and the Cajal body. The nucleolus is the location of the transcription and processing of ribosomal RNA and the Cajal body is the site of snRNP assembly, while the splicing factor compartments are a storage and assembly site for spliceosomal components. [(1)]

The functions of several proteinaceous nuclear bodies are still unknown. However, three models for the function of a proteinaceous nuclear body have been purposed. The first model simply suggests that the nuclear bodies are of no functional significance and develop merely as a consequence of the aggregation of surplus, unused protein. The second model attributes proteinaceous nuclear bodies as the site of specific nuclear activities. This model exemplifies the function of the nucleolus. Lastly, the third model characterizes proteinaceous nuclear bodies as storage sites for proteins, which can be easily recruited to nearby areas to function, such is the case with the Cajal body.[(1)]

Regardless of the function of proteinaceous nuclear bodies, their development and maintenance is still not completely understood. One purposed model that provides some

insight into this phenomenon is the nuclear self-organization model. This model attributes the formation of proteinaceous nuclear bodies to transient interactions among its resident components. This model suggests that proteins rapidly and randomly move through the nucleus until they find their binding sites on DNA and then transient interactions between these proteins and DNA or other proteins associates the proteins with a particular nuclear body. In this model, nuclear bodies are both dynamic and stable steady-state structures, which rapidly exchange their component proteins with the surrounding nucleus, yet maintain a high density of particular types of proteins characteristic of that nuclear body. Support for this model comes from previous experiments conducted with proteinaceous nuclear bodies. For instance, the importance of protein-DNA interactions in the development of nuclear bodies is evident because, in the nucleolus, when DNA-specific RNA polymerase I is inhibited or ribosomal RNA genes are silenced, the nucleolus disassembles.[2] Further support comes from experiments with Cajal bodies. Cajal bodies are known to associate with histone and U2 snRNA gene clusters. The introduction of U2 snRNA promotes the formation of Cajal bodies.[3-5] Similarly, splicing factor compartments, which contain pre-mRNA splicing factors often associate with gene-dense R-bands. This facilitates the processing of RNA generated from the gene-dense R-bands.[6-7] In this study, we developed a computational model to explore the effects of specifically characterized protein-protein and protein-gene interactions on the development of nuclear bodies.

Methods

A computational model was developed to simulate the formation of proteinaceous nuclear bodies using the program, Matlab. The computational model simulates the

behavior of the components of the nucleus. This behavior is generated within a domain that represents a domain within the nucleus. The model contains a stationary gene and numerous moving protein particles. The protein particles bind to the gene with a maximum of two protein particles allowed to bind to the gene at a given time interval. Once the two protein particles are bound to the gene, they remain on the gene for a certain number of time steps and then dissociate from the gene. Two protein particles which dissociate from the gene remain together for a prescribed time period. These bound protein particles, or dimers move with a slower velocity until they dissociate from each other.

Before the computational model is run on Matlab, the user can input certain variables that are constant throughout the duration of the program. These variables include the number of time intervals the model is run, domain boundaries (which are denoted by maximum and minimum x and y values so that the domain is rectangular), the maximum displacements in the x and y directions that the particles can travel in a single time step, the number of time steps single protein particles (monomers) or protein particles bound to other protein particles (dimers) bind to the gene (monomer docking time interval and dimer docking time interval respectively), the number of time steps dimer proteins particles bind to each other (dimer association time interval), the scaling factor by which the speed of dimers is reduced (dimer speed scaling factor), the diameter of the protein particles and the diameter of the gene. Other variables, such as the location of the stationary gene are defined by the coding of the programs. In this particular study, the center of the gene is always located at the center of the domain set by the user. Once these variables are defined, the model begins by choosing random positions for a certain

number of protein particles (each particle is designated by a number). The positions are denoted by an x coordinate value and a y coordinate value and these values are randomly selected according to the boundaries outlined in Figure 1. These boundaries dictate that all protein particles must be within the preset domain, that no two protein particles can occupy the same space, and that no more than two protein particles can be inside the boundary of the gene at a given time. Once simulation starts, the proteins are randomly moved within the domain. This is accomplished by generating random numbers and scaling them (using the preset maximum x and y displacement factors) to obtain the displacement values for each protein. These displacements are added to previous positions of the protein particles; new positions are restricted by the same rules outlined above for the initial positions of the protein particles as demonstrated in Figure 3. When the proteins come in contact with a gene, which is defined by the preset gene radius, they bind to the gene. At most, two proteins can bind to the gene at a give time step. If only one protein is bound to the gene it remains bound to the gene for monomer docking time interval as explained in Figure 2. Two proteins remain bound to the gene for the dimer docking time interval. These two proteins represent a protein complex which interacts with the gene to enable events such as transcription. After remaining bound to the gene for the monomer docking time interval or the dimer docking time interval, the proteins dissociate from the gene. If two proteins were bound to the gene together, then they leave the gene bound to each other forming the dimer. These two proteins remain bound to each other for the dimer association time interval. These dimers move with slower speed than monomers and cannot bind to the gene. Reducing dimer speed is accomplished by scaling their displacement by the dimer speed scaling factor. After the dimer association

time interval has passed, the proteins no longer stick together. This process is outlined in Figure 4.

This simulation is run for a preset number of time intervals and the density of proteins for various bins around the gene is recorded to determine if there is an accumulation of protein anywhere in the domain. For each time interval, circular bins with the same radius are created in the domain radiating from the center where the gene is located; the gene is not within any bin. In this way, 11 bins of equal radius are created and numbered such that bin 1 is closest to the gene and bin 11 is furthest from the gene. Then the number of protein particles in each bin is determined and divided by the area of each bin to determine the protein density in each bin.

Results

Multiple simulations of the program were run to collect data. For each case, the program was run for 10,000 time intervals with 50 proteins particles. The domain of the simulation was a square with side length of 100 units. In a given time step, the maximum displacement a particle could move in the x or y directions was 2 units. The radius of each particle was 2 and radius of the gene was 15. The monomer docking time interval, or the amount of time a single protein remains bound to the gene, was 3 time steps. The dimer docking time interval, or the amount of time two proteins remain bound to the gene, was also 3 time steps. The dimer association time interval, or the number of time intervals that two proteins stay bound together after leaving the gene, and the dimer speed scaling factor, or the factor by which displacement is scaled down for two proteins which are bound together, were varied for the various cases (as shown in Table 1) to determine how this impacted the protein distribution throughout the domain. After each case, the

positions of all the particles during all the time intervals were recorded. The domain was divided into a number of concentric circular bins emanating from the gene. The protein density was calculated in each bin and used to calculate the average protein density for each bin over the 10,000 time intervals. This was done for each case 3 times. The average protein density for the 3 trials was determined and listed in Table 2 and graphed in Figures 5 and 6. Then an analysis of variance, or anova, was done on the averages (listed in Table 2) of the 3 trails and the P-values are listed in Table 1.

For cases 3, 6, 9, and 12, the dimer speed scaling factor was set to 1.0, while the dimer association time interval was varied from 1 to 500. When the dimer association time interval was 1 in case 3, the range of average protein densities in the bins ranged from 0.005113 to 0.005477 with a P-value of 0.314 suggesting a uniform distribution of protein particles throughout the domain. This uniform distribution was also present in case 6 (with protein densities in bins ranging from 0.005098 to 0.005412) when the dimer association time interval was 20 with a P-value of 0.697 for the average protein densities in the various bins, case 9 (with protein densities in bins ranging from 0.005112 to 0.005513) when the dimer association time interval was 100 with a p-value of 0.756 for the average protein densities in the various bins, and trial 12 (with protein densities ranging from 0.004805 to 0.005709) when the dimer association time interval was 500 with a p-value of 0.0602 for the average protein densities in the various bins.

In cases 2, 5, 8, and 11, the dimer speed scaling factor was held constant at 0.75, while the dimer association time interval was again varied from 1 to 500. With the dimer association time interval set at 1 in case 2, the range of average protein densities in the bins ranged from 0.005066 to 0.005449 with a P-value of 0.113 again suggesting a

uniform distribution of protein particles throughout the domain. This uniform distribution is also true for case 5 (when the dimer association time interval is 20) with a P-value of 0.00595 and a range of average proteins densities from 0.005175 to 0.005977 within the bins. However, in case 8, when the dimer association time interval is set to 100, the distribution of average proteins densities is no longer uniform, as can be determined by the P-value of 3.78×10^{-10}. There is aggregation of protein particles near the center of the domain near the gene with the average protein density in bin 1 being 0.006302 and the average protein density in bin 11 being 0.004854. This aggregation can also been seen in case 11 when the dimer association time interval is 500 and the P-value is 3.29×10^{-07}. In this case, the average protein density in bin 1 was 0.006429 and the average protein density in bin 11 was 0.005312.

The dimer speed scaling factor was set to 0.5 for cases 1, 4, 7, and 10. When the dimer association time interval is 1 in case 1, the protein distribution throughout the domain is again uniform, as can be seen with a P-value 0.00597 and with average protein densities in the different bins ranging from 0.5112 to 0.005393. However for cases 4 (when the dimer association time interval is 20), 7 (when the dimer association time interval is 100), and 10 (when the dimer association time interval is 500), the protein distribution is not uniform throughout the domain as is apparent in the P-values of 8.73×10^{-11}, 2.55×10^{-16}, and 1.64×10^{-19} respectively. There is an aggregation of protein particles near the center of the domain where the gene is located. The average protein density in bin 1 was 0.007396 and the average protein density in bin 11 was 0.005008 for case 4. For case 7, the average protein density in bin 1 was 0.009804 and the average protein density in bin 11 was 0.004740. However the largest range is seen in case 10 with

the average protein density in bin 1 being 0.010681 and the average protein density in bin 11 being 0.004453.

Discussion

The computational model was run with the same parameters for each case, except the dimer association time interval and dimer speed scaling factor were varied. The graphs show a significant aggregation of protein particles toward the center of domain near the gene when the dimer association time interval is 100 or 500 and the dimer speed scaling factor is 0.5 (Figure 5 and Figure 7). This is seen by high average protein density in bins close to the gene compared with low average protein density in the bins far from the gene. However, no aggregation is seen when the dimer association time interval is 1 or the speed scaling factor is 1 (Figure 5 and Figure 6). These results, which are seen graphically, were also confirmed with the p-values calculated using analysis of variance (Table 1). For instance, the p-values suggest there is about a 1.64 ×10-17% chance that the aggregation seen when the dimer association time interval is 500 and the dimer speed factor is 0.5 is due to random events. However, there is 31.4% possibility that the protein particle distribution seen when the dimer association time interval is 1 and the dimer speed factor is 1 is due to random chance. Thus, it is conceivable that a higher dimer association time interval and a lower speed scaling factor, may lead to the development of a region of high protein density around the gene. This implies that simple factors such as the characteristics of the interactions among proteins and of the interactions between proteins and DNA, can cause a uniformly distributed protein population to become aggregated near a gene offering insight into the notion of nuclear self-organization.

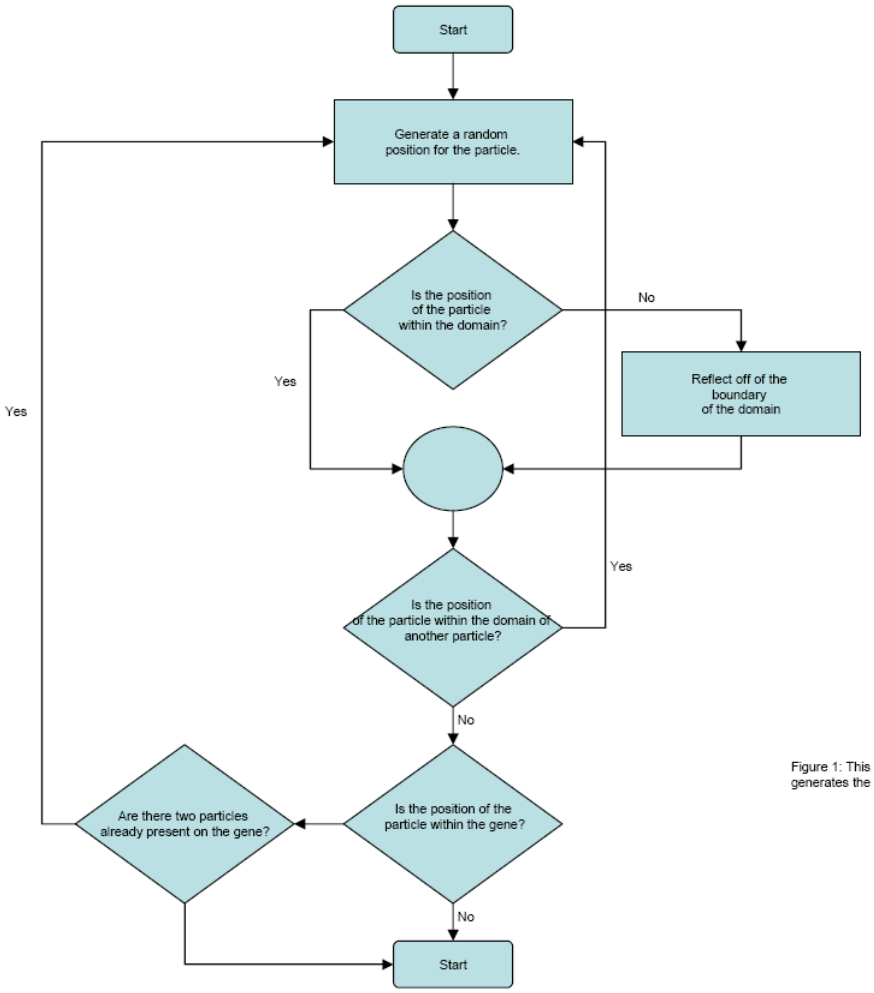

Figure 1: This outlines how the program generates the initial positions of the particles.

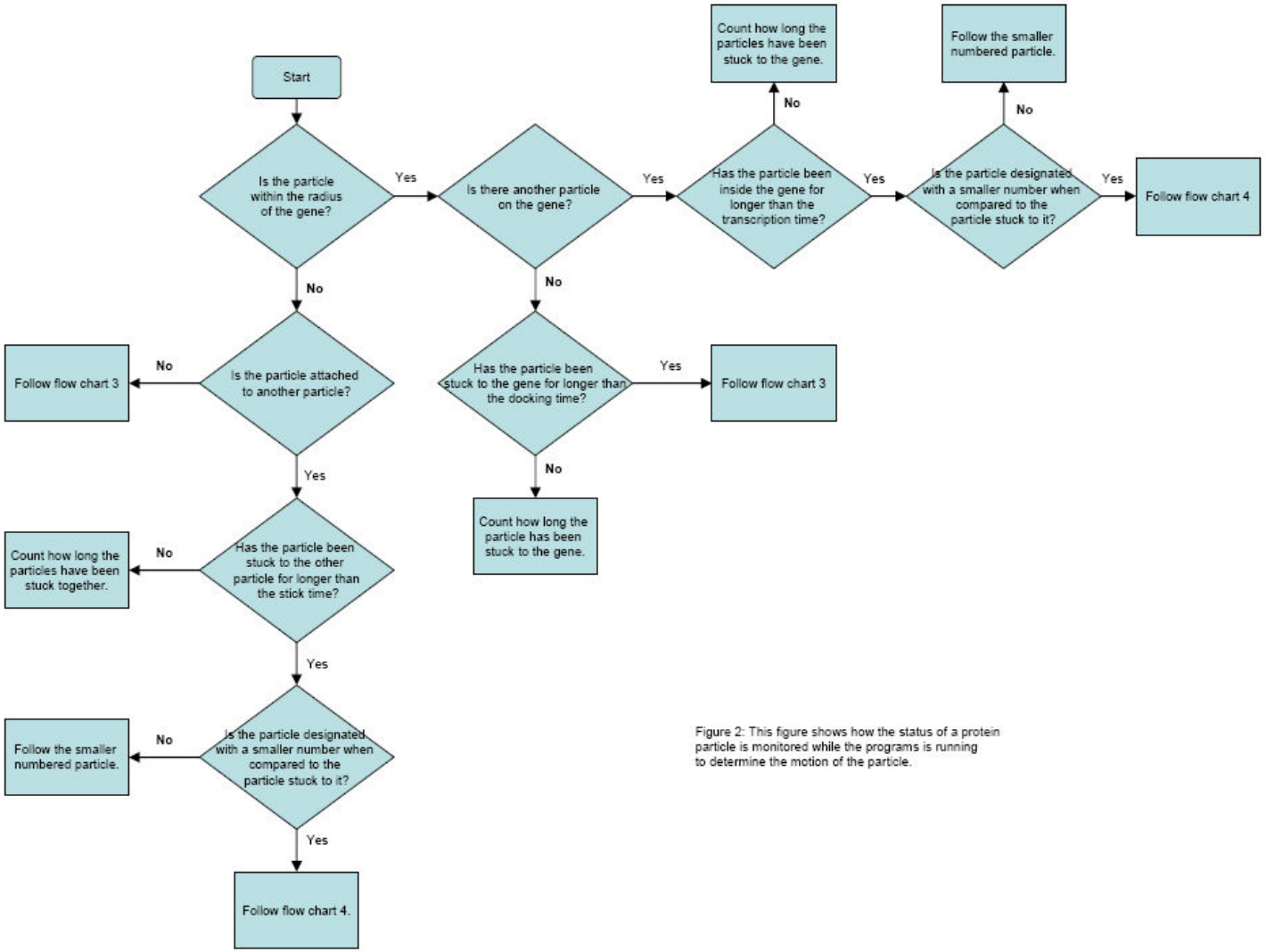

Figure 2: This figure shows how the status of a protein particle is monitored while the programs is running to determine the motion of the particle.

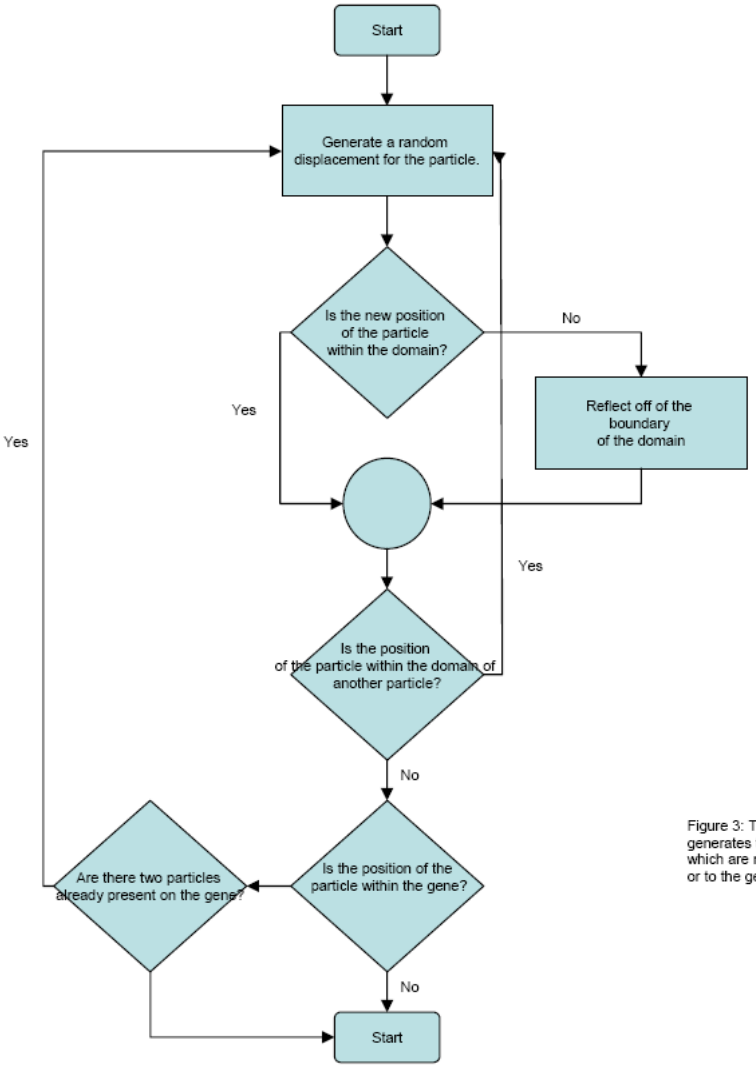

Figure 3: This outlines how the program generates the motion of protein particles which are not bound to any other particle or to the gene.

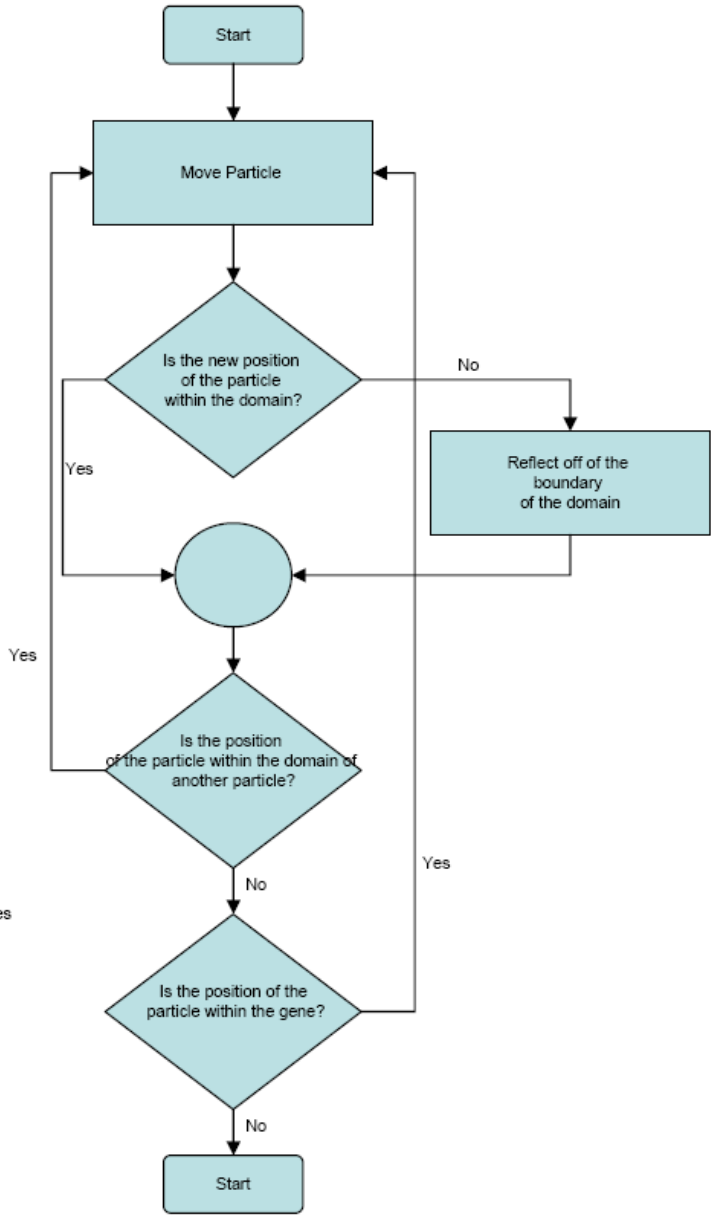

Figure 4: This outlines how the program generates the motion of two protein particles which are bound to each other.

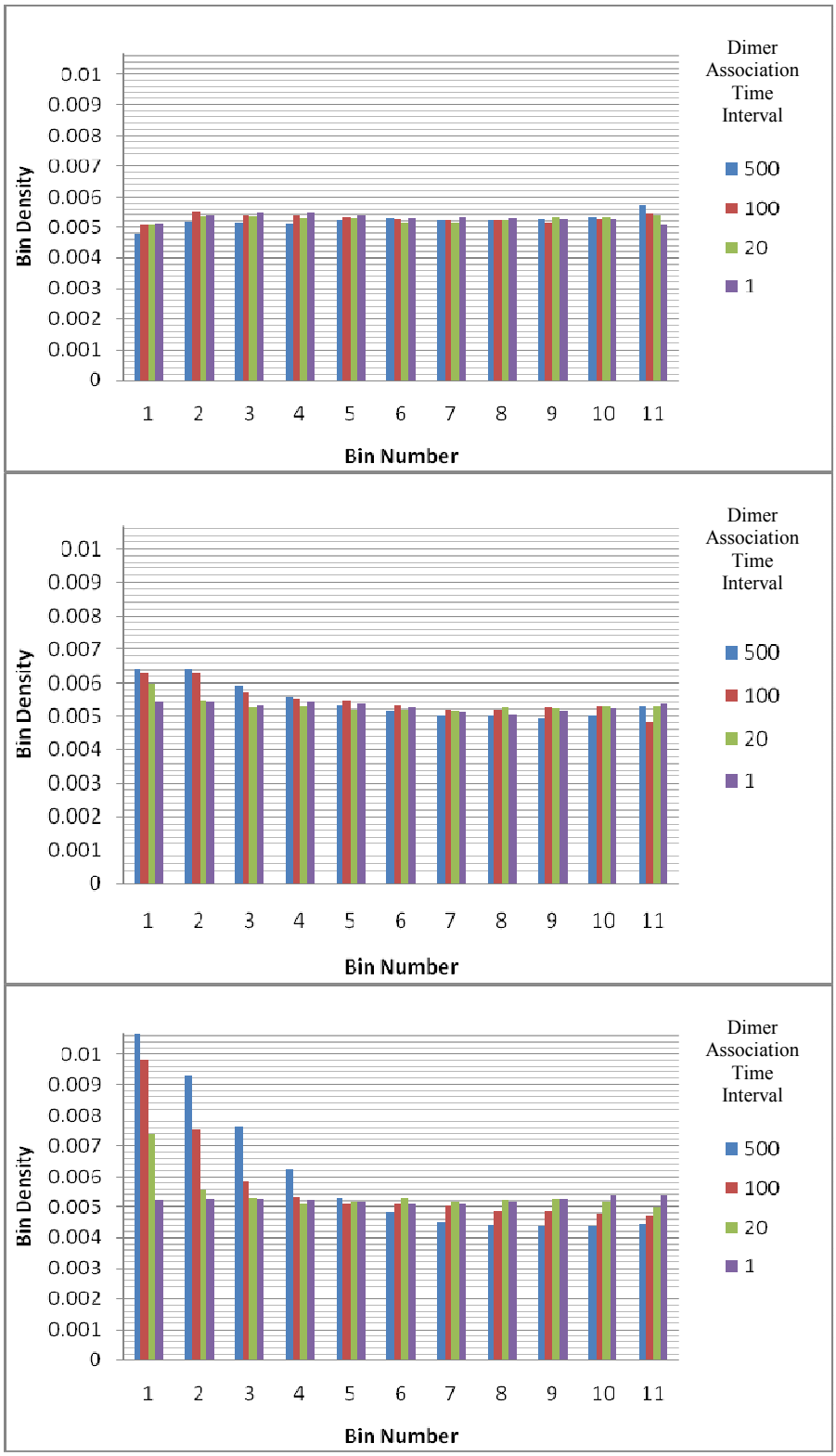

Figure 5: Plots of the average protein densities in each of the 11 bins for all of the different cases. (A) The dimer speed scaling factor was 1. (B) The dimer speed scaling factor was 0.75. (C) The dimer speed scaling factor was 0.5.

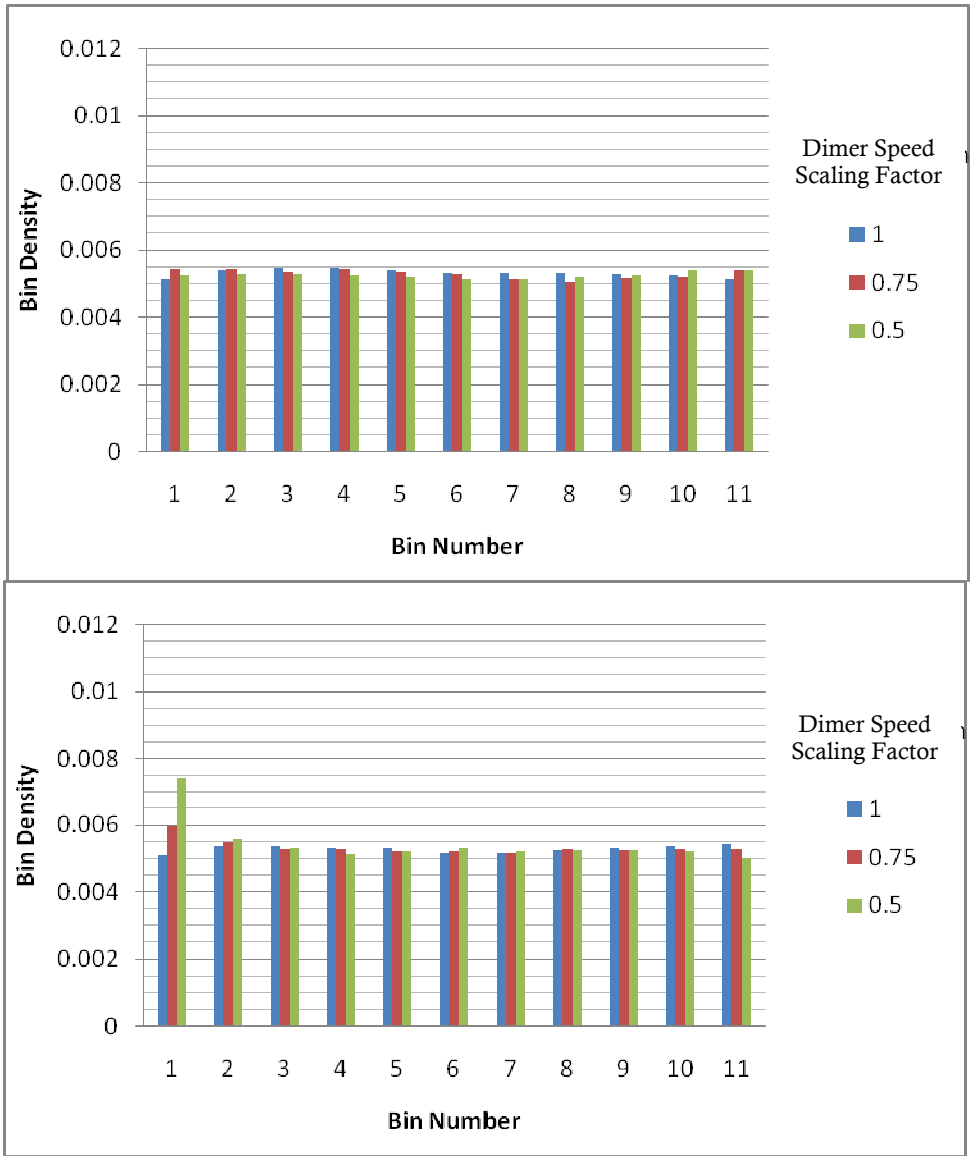

Figure 6: Plots of the average protein densities in each of the 11 bins for all the different cases. (A) For these cases the dimer association time interval is 1. (B) For these cases the dimer association time interval is 20.

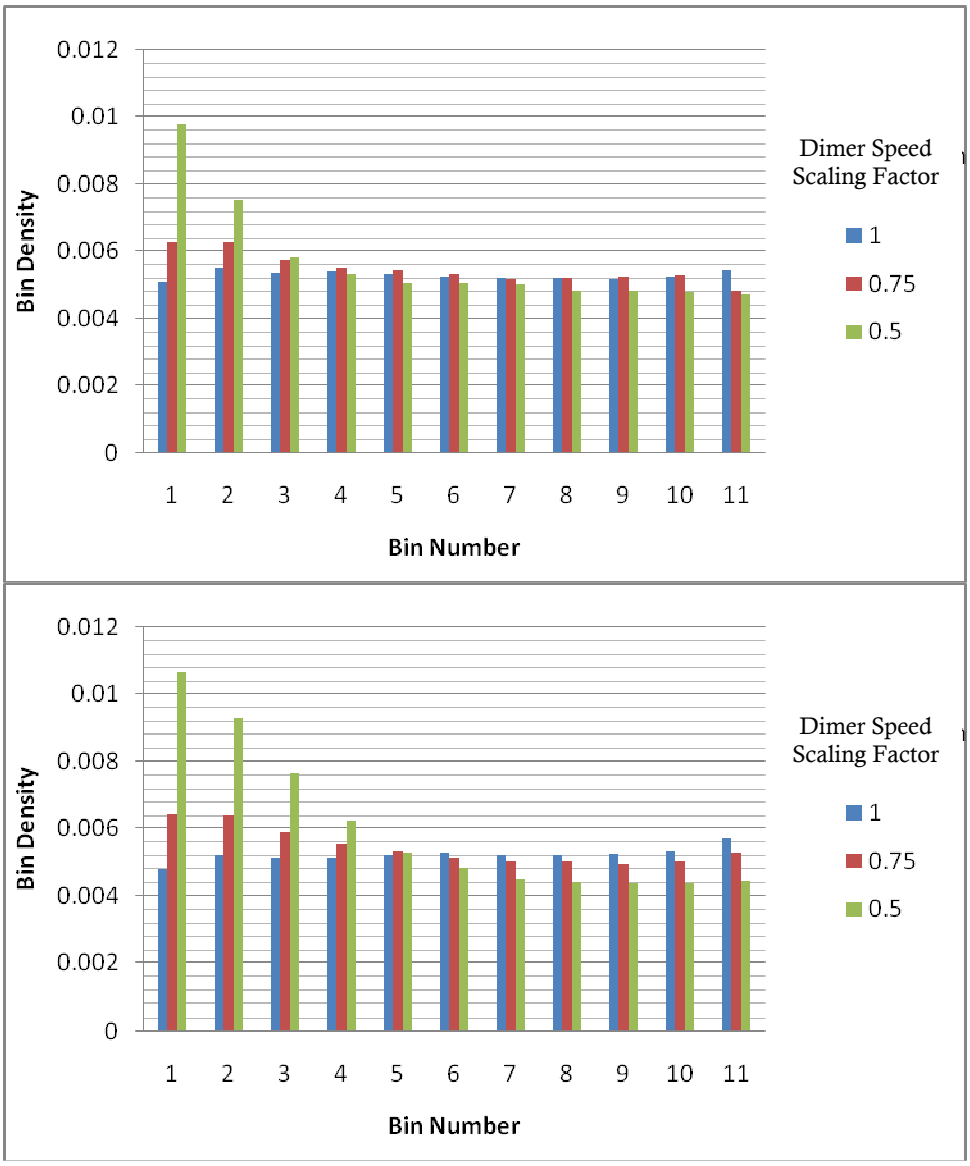

Figure 7: Plots of the average protein densities in each of the 11 bins for all the different cases. (A) For these cases the dimer association time interval is 100. (B) For these cases the dimer association time interval is 500.

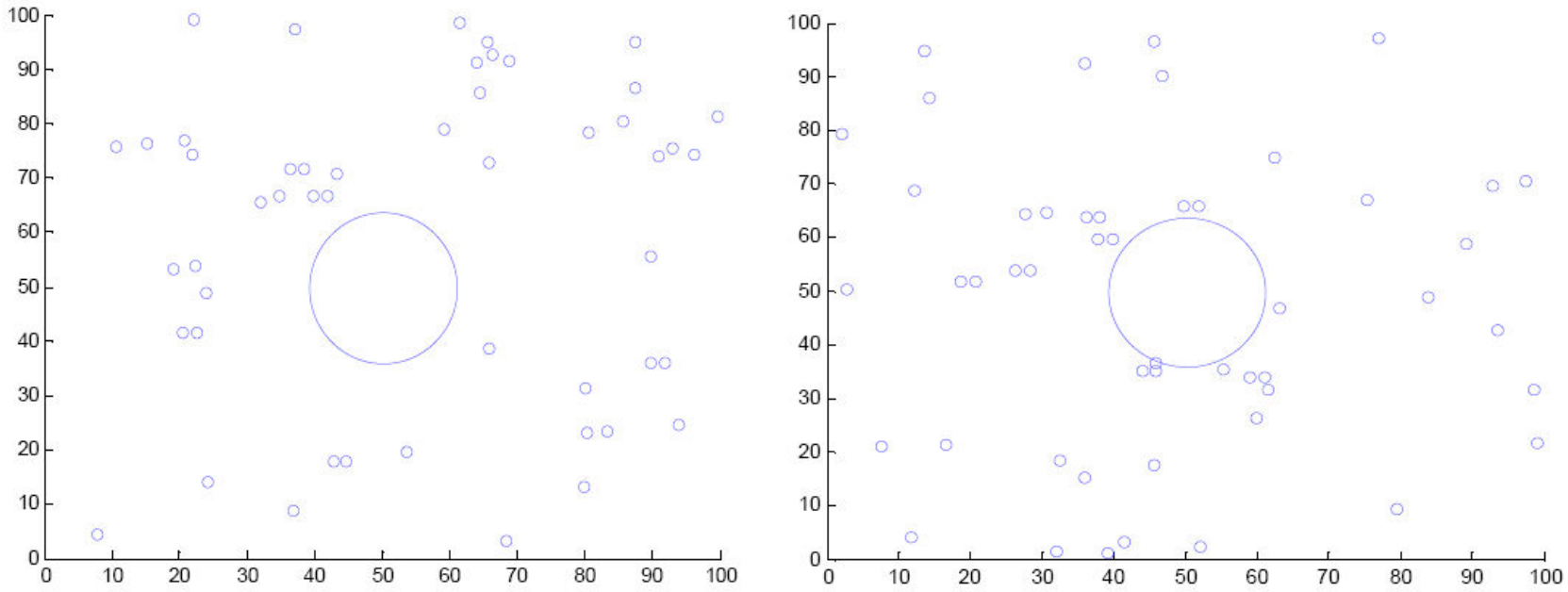

Figure 8: Scatter plots depicting the distribution of protein particles in a given time step throughout the domain which is 100 units in width by 100 units in length. The large circle in the center represents the gene while the smaller circles represent the protein particles. In this specific run, the dimer association time interval is 500 and the dimer speed scaling factor is 0.5. (A) This is the initial protein particle distribution at the start of the run of the computational model. A uniform distribution of protein particles throughout the domain is seen. (B) This is the protein particle distribution at the end (or after 10,000 time steps) of the run of the computational model . An aggregation of protein particles near the gene can be seen.

Case Number	Dimer Association Time Interval	Dimer Speed Scaling Factor	P-value
Case 1	1	0.5	0.0597
Case 2	1	0.75	0.113
Case 3	1	1.0	0.314
Case 4	20	0.5	8.73×10^{-11}
Case 5	20	0.75	0.00595
Case 6	20	1.0	0.697
Case 7	100	0.5	2.55×10^{-16}
Case 8	100	0.75	3.78×10^{-10}
Case 9	100	1.0	0.756
Case 10	500	0.5	1.64×10^{-19}
Case 11	500	0.75	3.29×10^{-07}
Case 12	500	1.0	0.0602

Table 1: Dimer association time interval and dimer speed scaling factor used in each case.

Protein Densities In Bins											
	Bin Number										
Case	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Bin 8	Bin 9	Bin 10	Bin 11
1	0.005240	0.005283	0.005268	0.005227	0.005211	0.005131	0.005112	0.005216	0.005255	0.005393	0.005390
2	0.005449	0.005449	0.005345	0.005434	0.005354	0.005260	0.005131	0.005066	0.005154	0.005219	0.005383
3	0.005117	0.005401	0.005477	0.005477	0.005408	0.005312	0.005322	0.005316	0.005271	0.005254	0.005113
4	0.007396	0.005589	0.005303	0.005132	0.005209	0.005312	0.005195	0.005230	0.005251	0.005189	0.005008
5	0.005977	0.005486	0.005270	0.005288	0.005203	0.005207	0.005175	0.005263	0.005245	0.005294	0.005292
6	0.005098	0.005365	0.005355	0.005306	0.005310	0.005181	0.005148	0.005248	0.005335	0.005336	0.005412
7	0.009804	0.007529	0.005847	0.005334	0.005084	0.005083	0.005026	0.004849	0.004845	0.004798	0.004740
8	0.006302	0.006278	0.005756	0.005512	0.005459	0.005343	0.005186	0.005213	0.005264	0.005297	0.004854
9	0.005112	0.005513	0.005389	0.005411	0.005342	0.005256	0.005228	0.005227	0.005164	0.005275	0.005451
10	0.010681	0.009295	0.007644	0.006243	0.005286	0.004821	0.004504	0.004413	0.004372	0.004380	0.004453
11	0.006429	0.006413	0.005896	0.005568	0.005336	0.005151	0.005039	0.005014	0.004971	0.005021	0.005312
12	0.004805	0.005200	0.005149	0.005133	0.005221	0.005305	0.005224	0.005222	0.005254	0.005344	0.005709

Table 2: Average protein densities within each bin during the different cases of running the program. The dimer association time interval and dimer speed scaling factor for each case are defined in table 1.

References

1. Misteli, Tom. 2005. Concepts in nuclear architecture. Bioessays 27:477-487.

2. Olson MOJ, Dundr M, Szebeni A. 2000. The nucleolus: An old factory with unexpected capabilities. Trends Cell Biol 10:189–196.

3. Schul W, van Driel R, de Jong L. 1998. Coiled Bodies and U2 snRNA Genes Adjacent to Coiled Bodies Are Enriched in Factors Required for snRNA Transcription. Mol Biol Cell 9:1025–1036.

4. Shopland LS, Byron M, Stein JL, Lian JB, Stein GS, et al. 2001. Replication-dependent histone gene expression is related to Cajal body (CB) association but does not require sustained CB contact. Mol Biol Cell 12:565–576.

5. Frey MR, Matera AG. 2001. RNA-mediated interaction of Cajal bodies and U2 snRNA genes. J Cell Biol 154:499–509.

6. Wang J, Shiels C, Sasieni P, Wu PJ, Islam SA, et al. 2004. Promyelocytic leukemia nuclear bodies associate with transcriptionally active genomic regions. J Cell Biol 164:515–526.

7. Shopland LS, Johnson CV, Byron M, McNeil J, Lawrence JB. 2003. Clustering of multiple specific genes and gene-rich R-bands around SC-35 domains: evidence