Modelling the phase separation of transcription factors on DNA

Virginia d'Adamo
virginia.dadamo@epfl.ch

Fall Semester 2024

Contents

1 Abstract

This project employs a one-dimensional Monte Carlo simulation to investigate the cluster size distribution of a transcription factor (protein A) on a linear DNA strand. The primary objective is to evaluate whether the introduction of a second protein, an intrinsically disordered protein (IDP, referred to as protein B), induces a phase transition in the cluster size distribution of protein A. The simulation starts with protein A binding exclusively to the DNA, followed by the gradual incorporation of protein B to assess its influence on the system. The study explores phase separation in this complex system through dimensionality reduction of the parameter space. Key outputs of the simulations include the cluster size distribution for protein A and the residence times of the transcription factor. The analysis identifies critical parameters influencing cluster growth and establishes optimal parameter ranges to streamline future computational models. The results confirm scale-invariance within the system and demonstrate behavior consistent with phase separation.

2 Introduction

Transcription factors (TFs) are sequence-specific DNA-binding proteins that play a fundamental role in regulating the spatiotemporal dynamics of gene expression. These proteins modulate transcription by facilitating or hindering the recruitment of RNA polymerase to specific target genes, thereby altering transcriptional rates. This study focuses on elucidating the mechanisms underlying transcriptional activation. TFs are pivotal in biological processes such as cell differentiation, where they govern developmental patterns through precise gene regulation. The study of TF behavior holds significant importance for human health, as disruptions in TF activity often lead to aberrant gene expression, which can result in various diseases. Understanding these mechanisms at a deeper level has the potential to advance healthcare and foster the development of targeted therapeutic strategies. Gene-specific TFs exert their regulatory functions by recognizing and binding to specific DNA sequences within promoter or enhancer regions. These interactions facilitate the recruitment of transcriptional co-activators and general transcription factors, enabling transcription and ensuring proper gene regulation. Recent evidence underscores the significance of combinatorial control in transcriptional regulation. Clusters of DNA motifs that recruit multiple TFs act as hubs for additional regulatory molecules, enhancing the efficiency of TFs in locating their target sites. Notably, transcription factors frequently form nuclear clusters that facilitate transcription. Although the precise mechanisms underlying cluster formation remain unclear, it is hypothesized that these clusters are initiated by the sequence-specific binding of TFs to DNA and subsequently expand through interactions mediated by intrinsically disordered regions (IDRs) of TFs. IDRs within transactivation domains have been implicated in forming biomolecular condensates via various mechanisms. The binding of a TF to a specific site can

nucleate site-specific transcriptional condensates, which serve as attractors for other TFs, leading to their local accumulation near the enhancer or promoter of a target gene. These interactions contribute to the formation and stabilization of transcriptional clusters that regulate gene expression. Such processes often involve complex, energy-dependent dynamics, where multiple proteins bind to and dissociate from specific DNA sites, influencing cellular functions. To investigate these molecular interactions, we employ Hamiltonian Monte Carlo simulations to model the stochastic nature of protein-DNA binding and unbinding events.

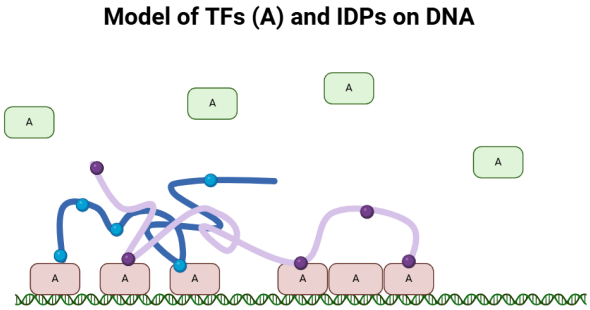


Figure 1: *The figure illustrates the simulation model. The DNA is represented as a linear strand. Transcription factors (TFs), denoted by A, are either bound to the DNA (shown in red) or free in solution (shown in green). Two B proteins are depicted in blue and violet. The circles represent binding sites, which can either be occupied by an A or remain free.*

As it is depicted in Figure 1 in this study we consider two types of proteins: Protein A (a transcription factor) that binds nonspecifically to DNA, and Protein B (an intrinsically disordered protein or IDP) that serves as a platform with multiple binding sites for Protein A. The interactions between these proteins and their effects on DNA binding are governed by energy-based rules, incorporating binding affinities, competitive dynamics, and potential cooperative effects. To further explore the dynamics of protein clustering, concepts from the Ising model, widely used in statistical physics to describe phase transitions, are applied. In this context, local binding interactions between proteins A and B on DNA may result in emergent behaviors such as phase transitions in the size distribution of Protein A clusters. Monte Carlo simulations allow us to assess how variations in system parameters influence cluster growth and identify key drivers of these transitions. Additionally, we analyze the residence times of transcription factors—defined as the duration a TF remains bound to DNA between binding and unbinding events. Evidence suggests that longer residence times correlate with increased transcriptional output of target genes. Studies, such as Donovan et al. (2019), Loffreda et al. (2017), and Stavreva et al. (2019), have shown that high-affinity binding sites significantly prolong TF residence time, impacting transcriptional outcomes. For example, mutating a Gal4 binding site in yeast demonstrated that Gal4 binds longer to high-affinity motifs, enhancing transcriptional activity. There is growing interest in understanding how TF residence time, bound fraction, and concentration influence transcriptional bursting. However, the complexity of this model, with its numerous parameters, necessitates dimensional reduction to facilitate computational simulations. By identifying the critical pa-

rameters affecting cluster growth and limiting their ranges, we aim to make this intricate system more tractable for in-depth analysis.

In this report, we will explore the various aspects of our simulation study, focusing on the impact of different parameters on the system's behavior. The methods section will provide an overview of the key parameters used in the simulation, the Monte Carlo simulation steps employed, and the assumptions made to simplify the model. We will also discuss the key observables analyzed throughout the study, including the system size, energy values, and their effects on the observed parameter space. The results and discussion section will delve into the analysis of the average B-occupied fraction, cluster size, and the distribution of B at the final time step, along with temporal analysis to understand the system's evolution over time. The findings will be framed within a broader context to highlight the general observations derived from the simulations. Finally, the we will conclude with insights into the implications of our results and future directions for further research

3 Methods

This Monte Carlo simulation, implemented in Python, models discrete steps in which proteins bind to and unbind from specified sites along a linear DNA chain. The code's core functions include methods to conduct Monte Carlo steps for each protein, compute the energy requirements for binding and unbinding events, and update the proteins' binding states over time. To optimize performance, numpy arrays were used wherever possible to manage memory efficiently. When arrays couldn't be applied, lists were used and subsequently deleted to free memory after use. The simulations were conducted on the Helvetios computing system. We will now summarize the main structure of the simulation, for further detail please refer to the README or the docstrings of the functions in the GitHub repository (https://github.com/virginiadadamo/Modelling-the-phase-separation-of-transcription-factors-on-DNA.git)

3.1 Parameters

Main parameters are summarized in Table 1. A central aspect of the simulation involves modeling DNA as an array, where each site can exist in one of two states: unoccupied or bound by protein A. The ensemble of As is represented as a set of transcription factors, each capable of interacting with both the DNA and protein B, while protein B is depicted as a set of binding sites, each capable of accommodating a single protein A molecule. One of the assumptions of the model is that protein B can only bind to the DNA only if an A is already present on the DNA. The hierarchical structure of the model introduces a directional perspective to the interactions. From the DNA's viewpoint, only the presence of protein A is

recognized. Protein A, in turn, perceives its interactions with both the DNA and protein B. Conversely, protein B interacts exclusively with protein A and does not directly perceive the DNA. This structured interaction framework ensures a clear directionality within the system, facilitating an organized and tractable representation of the complex dynamics involved.

DNA structure: Represented as 1xN array where 0 denotes an empty site, 1 indicates binding by protein A. N are the number of sites on the DNA.

Protein A: Stored in an nA x 2 array, where each row represents an individual protein A. The first column marks -1 if the transcription factor is unbound, or the site number if it is bound. The second column marks -1 if it is not bound to protein B, or stores the index of the corresponding protein B if it is. nA is the number of As in the simulation, which can be computed by multiplying the density of As with N, the number of DNA sites.

Protein B: Represented in an nB x k matrix where k represents the number of binding sites. Each entry contains -1 if the site is free, or the corresponding protein A if bound. nB represents the number of Bs present in the simulation. Another key parameter for the structure of B protein is L, which represents the length (in DNA sites) that there is between two adjacent binding sites. This means that if a site of a B is bound to an A to the DNA sites, the adjacent site of the B can bind on As on the DNA within a maximum distance of L.

The energy scales in the simulation incorporate several critical factors: the binding energy of protein A to DNA (E_{ad}), the interaction strength between adjacent DNA-bound transcription factors (E_{aa}), and the mutual influence of proteins A and B on each other's unbinding behaviors (respectively E_{ba} and E_{ab}). To emphasize the role of protein B in facilitating cluster formation, an idealized assumption is introduced: the interaction energy between protein B and protein A (E_{ab}) is considered infinite. Under this assumption, any protein A bound to protein B remains permanently bound and cannot unbind. However, if the corresponding protein B unbinds (with a probability proportional to the interaction energy E_{ba}), the associated protein A is released and becomes free to leave the DNA.

This framework can be conceptually linked to the original Ising model. The latter describes how temperature affects interactions between "spins" on a lattice, where each spin can be in an "up" or "down" state. At high temperatures, the spins are oriented randomly, resulting in no net magnetization. Below a critical temperature (the Curie temperature), the interaction energy between neighboring spins becomes dominant, overcoming thermal entropy, and all spins align, producing a net magnetization. Although this is framed in terms of temperature, the concept can also be explored by varying the interaction energy, which can have a similar effect. Other types of phase transitions can also occur depending on the concentration or interaction energy, not just temperature. In our case, by adjusting the the value of the energy parameters, we can induce a phase transition without changing temperature.

3.2 The simulation

The initial configuration of the system begins with all DNA sites empty, and both transcription factors (A and B) are free in solution. The simulation employs a MC algorithm to determine the binding status of protein A and protein B, incorporating probabilistic adjustments based on the various energy parameters in the system. Two configurations of the simulation are available: one with the presence of protein B and one without. Separate MC steps are executed for each configuration to represent these distinct scenarios. The simulation progresses by updating the system's state at each step. To ensure that the system reaches a steady-state or equilibrium before sampling data, we introduce an "ignoring time" period at the beginning of the simulation. During this initial phase, no data is collected, as the system is allowed to evolve towards equilibrium. This step is crucial because the first few iterations typically involve the system settling into its characteristic behavior, where TFs begin to bind to the DNA and interact with each other. Once this equilibrium state is reached, we start taking samples to analyze the system's dynamics, ensuring that the collected data accurately reflects the long-term behavior rather than transient effects from the initial configuration. However it is important to note how although most of the samples are collected exclusively during these sampling steps, time-related variables are the only exception. Residence times and binding events for each transcription factor are updated continuously, whenever a binding or unbinding event occurs for any TF. This ensures accurate tracking of dynamic interactions and time-dependent behaviors within the system.

3.3 Assumptions

To simplify the model and improve computational efficiency, several assumptions were made.

In terms of binding availability, if a protein A molecule is free and there is an unoccupied binding site on the DNA, it is assumed that the protein will always locate and bind to that site. This assumption streamlines the binding process by removing the need for additional computations related to protein searching or diffusion time of the TF.

For residence time calculations, only transcription factors that both bind to and eventually unbind from DNA are included in calculating residence times—the duration a transcription factor remains bound. If a transcription factor binds but does not detach, it is excluded from the calculation of mean residence time.

To reduce artifacts stemming from these assumptions, the simulation was extended to a prolonged duration of 2 million steps (or 4 million for longer DNA). Running the simulation longer ensures the convergence to a steady-state of the system, whereas shorter simulations would allow the assumptions to exert a greater influence on the overall outcomes.

3.4 Monte Carlo Simulation Steps

The steps of the Monte Carlo simulation involve calculating energies that take into account binding affinities, intermolecular interactions and the intrinsic properties of the molecules involved. These steps are structured as follows, with separate components for handling the actions of A and B:

For A:

1. A random A is selected, and a random empty site on the DNA is chosen.
2. A random number between 0 and 1 is generated. If the number is less than 0.5, the event chosen is the addition of A; otherwise, it is the removal of A.
    - **Add_A_event:** If the chosen A is free this event always succeeds based on the initial assumptions, otherwise the step is ignored. When A is added, its residence time on the DNA is initiated and tracked. An `add_A_to_DNA_site` function is called to execute the actual binding logic and update structures.
    - **Remove_A_event:** The removal of A is subject to a probabilistic condition, which is based on the energy. The energy term if the A is not bound to a B is calculated by summing E_{ad} for the current site or E_{aa} for neighboring sites if they are occupied by a transcription factor. This logic is supported by the fact that the values in the DNA array are either 0 or 1, indicating the absence or presence of an A on the DNA. The presence of B, as already mentioned, sets the energy to infinity, making it impossible for the corresponding A to unbind. In this case the removal event will never succeed. In case of success of the move a `remove_A_from_DNA_site` function is called to execute the actual removal logic and update structures.

For B:

1. A random instance of B is selected.
2. A random number between 0 and 1 is generated to decide between the adding or removal event for B.
    - **Add_B_event:** Handles the event of adding a B protein to the DNA. If the selected B protein is free, a random binding site, among those where an A is bound, is selected. If the B protein is partially bound, it selects a free site with a bound A within a defined region $[-L, +L]$ around an already bound site. If a suitable site on the DNA satisfies the required assumptions (presence of an A or distance L), the binding of B will always succeed. An `add_B_to_DNA_site` function is called to execute the actual binding logic and update structures.

- **Remove_B_event:** This event handles the removal of B by selecting a bound site and calculating the energy required for its release. If the removal conditions are met, B is removed, and the state is updated. The energy for this event is proportional to E_{ba}, depending on the number of A sites bound to B. The computation of the total energy is now set to be proportional to the number of Bs, within a distance -L, L. In this way the more Bs are bound in a neighborhood, the less probable it will be for the B to leave. We explored two different approaches for calculating the energy associated with B unbinding. Initially, we computed the energy to be proportional to the number of binding sites on the same B. Later, to emphasize a clustering effect among Bs and encourage them to stay close together, we implemented a function where the energy is proportional to the number of Bs bound to As within a distance L on either side. A `remove_B_from_DNA_site` function is called to execute the actual binding logic and update structures

3.5 Key observable analyzed

The key observables analyzed in the study include cluster size and residence time. A cluster is defined by the number of consecutive As on the DNA. At each time step sampled the clusters sizes are collected. At the end of the simulation an average is computed across all time steps sampled. This helps to quantify the spatial organization of TFs over time. Residence time, on the other hand, is a crucial metric for understanding DNA "stickiness". In fact it varies with different energy values, offering insight into how changes in energy conditions affect DNA adhesion.

3.6 Plots used for analysis

The plots used are summarized in the Table 2. In our analysis, we utilized a series of plots to examine the behavior and characteristics of TFs binding and unbinding dynamics under varying conditions of the parameters. These plots can be selectively computed based on the specific focus of the analysis. For each plots the corresponding txt files are also saved. In addition other txt file generated include the sorting of the A matrix for two distinct time points: the final time step and the midpoint between the final time step and the sampling start point. For each of these time points, we applied two sorting methods:

1. Sorting by the second column: This grouped identical Bs together, allowing us to examine which sites they bind to and assess whether these sites tend to cluster nearby.

2. Sorting by the first column: This organized the data by DNA sites, enabling us to investigate whether adjacent DNA sites are bound to the same B.

By applying both sorting methods to both time points, we gained a comprehensive understanding of binding patterns and how they evolve over time.

4 Exploration of parameters space

Due to the model's complexity, dimensional reduction was necessary to simplify parameter interactions. Thus, our first goal was to examine the relative influence of each parameter on cluster formation, determining parameter ranges to enhance computational efficiency.

4.1 System Size Analysis

In our system size analysis, we investigated the effect of varying the number of DNA sites (N =3000, N=6000) on the dynamics of transcription factor binding. To ensure consistency, we kept the TF density constant across system sizes, enabling us to analyze how the number of available binding sites impacts overall behavior. Additionally, we explored different values for alpha, selecting those that struck a balance: sufficient TFs to enable clustering while avoiding excessive concentrations that would result in unrealistic biological scenarios with DNA overcrowding. By systematically adjusting both the system size and TF density, we aimed to understand how these factors interplay to influence TF clustering and binding dynamics. This approach provided valuable insights into the relationship between system size, TF density, and the emergence of phase transitions, shedding light on the conditions that drive collective TF behaviors or shifts in binding regimes.

No differences in behavior were observed using $N = 6000$ or by varying the density. Therefore, for computational efficiency, it was decided to focus the majority of the simulations on the case $N = 3000$ and $\alpha = 0.15$, while still retaining larger simulations as a check for the system's scale invariance.

The number of B proteins was kept constant, maintaining an intermediate value to ensure a balanced system. Instead, we focused on exploring variations in the structure of individual Bs by adjusting the ratio k/L. This allowed us to investigate how changes in the intrinsic properties of Bs influence the system's behavior, providing insights into the impact of structural variations on the dynamics of clustering and binding.

4.2 Energy Value Impact on Parameter Space

The impact of energy values on the parameter space was explored by analyzing the behavior of bound transcription factors over time. Specifically, we examined the number of A's bound

over time, which revealed minimal differences in the binding dynamics for varying E_{aa} and E_{ad} beyond a certain energy threshold. This indicated that, beyond a specific energy value, the system's behavior becomes relatively stable and insensitive to further changes in energy.

5 Results and Discussion

All the plots are available in the GitHub repository within their respective folders. For clarity, only the key plot relevant to the discussion will be presented here.

5.1 Average B occupied fraction

For a given set of parameters, all B sites exhibit similar behavior, indicating that the system has reached steady state. In the random case (all interaction parameters set to 0), the average occupied fraction is low. When E_{ad} is increased to a weakly bound state, the average occupancy shows a slight but non-significant increase, as A factors adhere to the DNA more effectively, making it marginally easier for B sites to locate and bind. Progressive increases in E_{ba} (e.g., to 1, 2, and 4) lead to a corresponding rise in the average occupied fraction. Similarly, higher E_{aa} values contribute to an increase in average occupancy by promoting clustering of A factors, which enhances the likelihood that, once a B site binds, it can more readily locate nearby A factors for additional binding. The most pronounced effect is observed in the maximal energy configuration (E_{ad} = 3, E_{aa}=2, E_{ba}=4), where nearly all B sites are occupied.

5.2 Cluster Size

It is important to note that the reported values are averages across the sampled time steps. Therefore, we do not have a direct measure of the cluster dynamics over time. For example, the same cluster at different time steps may split into smaller clusters or merge into larger ones, but we currently lack a method to capture these temporal changes.

A particularly noteworthy observation was the influence of the fragmentation ratio (k/L) in combination with the energy levels. We hypothesize that this ratio is closely correlated with the conformation of protein B, affecting its flexibility.

In the case where $k/L = 0.2$, introducing only E_{ba} (while keeping all other energy parameters at 0) did not produce any significant effects. Similarly, adding E_{ad} did not cause substantial changes. However, when E_{aa} was introduced with all other energy parameters set to 0, a reduction in the number of one-sized clusters was observed. Further introduction of E_{ba}

alongside E_{aa} resulted in a progressive effect, characterized by a decrease in one-sized clusters and an increase in the size of the maximal cluster. Additionally, when E_{ba} was held constant and E_{ad} was varied, we observed that lower values of E_{ad} led to an increase in the maximal cluster size and a rise in intermediate clusters, particularly at an intermediate value of $E_{ad} = 1$.

For higher k/L values, large clusters were already present when $E_{ba} = 4$ and all other energy parameters were set to 0. Again, the introduction of $E_{aa} = 2$ plays a role in the drop of one-sized clusters. It is noteworthy that in both cases, high values of E_{ad} negate the effects observed when introducing higher values of $E_{aa} = 2$ and $E_{ba} = 4$. We hypothesize that excessively high values of E_{ad} lead to a system where the single-clustered A species bind too strongly to the DNA, preventing sufficient unbinding and re-binding near the clusters, thereby limiting the system's fluidity. Excessively large E_{ad} values, particularly in combination with other energy parameters, cause the system to become too confined.

From this initial analysis, we confirm the role of E_{aa} in promoting the formation of A clusters. We also observe that E_{ba} does not exert a strong effect when k/L is too low and all other energies are set to zero.

Therefore, it is crucial to introduce E_{aa} and E_{ad} to observe more significant effects. Having explored these observations, we now focus on a fixed case for E_{ad}, selecting a value that is neither 0 (to avoid random binding) nor too high (as in the case of $E_{ad} = 3$, which limits fluidity). Specifically, we consider $E_{ad} = 1$ and $E_{aa} = 2$ to allow stronger formation of A-clusters and to observe the effects on B-distributions more clearly. We will investigate how these conditions vary with different k/L ratios, comparing the effects of weak versus strong E_{ba} values (Figures 4, 5, 6).

For the case of low k/L and low E_{ba} (Figure 4a), the system consists of predominantly high one-sized clusters, with low maximum cluster sizes and few intermediate clusters. For intermediate k/L and low E_{ba} (Figure 5a), fewer one-sized clusters are observed, with the maximum cluster size around 80 and still few intermediate clusters. For high k/L and low E_{ba} (Figure 6a), the number of one-sized clusters decreases further, the maximum cluster size reduces, and the number of intermediate clusters increases.

As the k/L ratio increases in conjunction with higher E_{ba} values, a different effect in the distribution of cluster sizes are observed. For low k/L and high E_{ba} (Figure 4b), the system exhibits a relatively high number of one-sized clusters, with a maximum cluster size of around 100, and few intermediate-sized clusters. In contrast, when the k/L ratio is intermediate (Figure 5b), the system shows a reduction in the number of one-sized clusters (slightly more than 20), accompanied by an increase in intermediate-sized clusters, with the maximum cluster size decreasing to around 60. For high k/L values and high E_{ba} (Figure 6b), the number of one-sized clusters drops further (fewer than 20), the maximum cluster size decreases to approximately 40.

5.3 Distribution of B at the last time step

Let us now focus on the distribution of B, with particular attention to the conditions previously analyzed (Figures 7, 8, 9). For the scenario with low k/L and low E_{ba} (Figure 7a), the distribution of B is highly fragmented, exhibiting scattered and disorganized patterns. As k/L increases to an intermediate level, while maintaining low E_{ba} (Figure 8a), the distribution becomes more clustered, although it remains relatively fragmented compared to higher values of k/L. With a high k/L and low E_{ba} condition (Figure 9a), the distribution of B shows even greater clustering, with more compact structures compared to the intermediate condition.

In contrast, when E_{ba} is increased while keeping k/L low (Figure 7b), the distribution becomes very clustered, with tightly grouped patterns and fewer fragments. When both k/L and E_{ba} are increased to intermediate levels (Figure 8b), the distribution of B shifts toward a more fragmented pattern. Finally, for high k/L and high E_{ba} (Figure 9b) conditions, the distribution remains fragmented, but with a higher degree of disorganization compared to the intermediate k/L and high E_{ba} scenario.

5.4 Temporal Analysis

The distribution of binding events and residence times is strongly influenced by the energy values within the system. To begin, we analyze the most random scenario, where all energy values are set to zero and no protein B is involved (Figure 2). In this case, binding and unbinding events occur with equal probability and once they are selected they never fail. Let's consider a system with 3000 DNA sites and 450 As over 2 million steps.

Time to Bind: To compute the average number of steps required for a specific A to bind, we calculate the probability of selecting that A (1/450) and multiply it by 1/2 (the probability of binding). The inverse of this probability gives the expected number of steps, $1/(1/450 \times 1/2)=900$.

Time to Unbind: Similarly, the average number of steps required for the specific A to unbind is calculated in the same way. Since the probability of selecting the A and the probability of unbinding are identical to those for binding (in this random case with no energy contributions), the expected time for unbinding is also 900 steps.

Bind/Unbind Cycle: A full bind/unbind cycle therefore requires 900+900=1800 steps. Over 2 million timesteps, the total number of bind/unbind events is approximately $(2 \times 10^6)/1800$ which is about 1100. This aligns with the results shown in the histogram. (Figure 2a)

Mean Residence Time: Since the binding and unbinding states are equally likely, the bound state duration constitutes half of the cycle. This gives an average residence time of 900 steps, consistent with the histogram. (Figure 2b)

Increasing E_{ad} has a distinct effect on the binding and unbinding dynamics. The binding time remains unaffected by E_{ad}, as it depends solely on the initial selection probability. However, the unbinding time is prolonged due to the enhanced stability of the bound state, increasing by a factor proportional to $e^{E_{\text{ad}}}$ (Figure 3).

In the absence of interaction energies (i.e., when there is no protein B or E_{aa}), the mean residence time can be generalized by the following expression:

$$\text{Mean Residence Time} = n_A \cdot 2^2 \cdot e^{E_{\text{ad}}}$$

Introducing additional energy values, such as E_{aa} or E_{ba}, complicates the dynamics further. It is important to note that the case where protein B has no energy interaction was also considered (and used as a check), and it was found to be similar to the random case, as the effect of protein B without interaction energy is negligible.

In general, higher energy values result in fewer binding events, reflecting a more energy-constrained system. This leads to an increase in the mean residence time, as transcription factors tend to bind and unbind more slowly, reflecting a reduction in interaction stability. As shown in the distribution of residence times, increasing the energy typically results in both an increase in the mean residence time and a higher standard deviation. This indicates that while the mean residence time increases with energy, the variability also rises, suggesting that the residence time for transcription factors becomes more diverse. For example, the residence time for an unbound transcription factor or one not involved in a cluster will be much shorter, while those that are more tightly bound or part of a cluster tend to have longer residence times.

It is also worth noting the effect of different energies on the ratio between the mean residence time and the standard deviation, which provides additional insight into the analysis of residence time distributions. It is important to clarify that this standard deviation differs from the standard deviation of the residence time distributions; rather, it is calculated by averaging the standard deviations of residence times across all transcription factors. As shown in Figure 10, the ratio of the mean to the standard deviation is highly sensitive to the E_{aa} value. For the most random case, the mean is equal to the standard deviation, introducing E_{aa}, reduces this factor by almost an half (Figure 10a). These findings suggest that elevated energy levels for E_{aa} induce more heterogeneous behavior among transcription factors. The inclusion of E_{ba} further reduces both curves, increasing the system's heterogeneity. Additionally, the curves no longer remain constant, and the ratio exhibits a nearly linear increase as E_{ad} values increase (Figure 10b).

We also observed the effect of varying fragmentation values while keeping E_{ba} constant. As the k/L ratio increases, the Mean Residence Time also increases. Additionally, with higher k/L values, the ratio between the Mean Residence Time and the standard deviation (of the distribution) decreases.

6 General observations

Bringing all these observations together, the results we discussed provide valuable insights into the interplay between k/L, E_{ba}, and their effects on the cluster size (resumed in Table 3), distribution of B (resumed in Table 4) and mean residence time.

From our analysis, we conclude that for achieving more clustered behavior at low energy levels (low E_{ba}), increasing k/L proves to be more effective. This strategy enhances the number of intermediate clusters and increases the mean residence time. By increasing k/L, we allow B-clusters to exert a larger influence, where higher intermediate cluster formation and longer residence times suggest that, at low energy levels, enhancing stability and reducing fluidity actually facilitates the formation of larger clusters.

In contrast, the opposite behavior is observed at higher energy levels. Lowering k/L, while maintaining a higher number of one-sized clusters with fewer intermediate clusters and reducing the mean residence time, appears to be more effective. This suggests that, at higher energy values, it is beneficial to enhance the fluidity of the system, allowing for more frequent binding-unbinding events, which may ultimately promote more dynamic cluster formation.

7 Data Management

The results in the repository are systematically organized to ensure easy access and interpretation. The repository is divided into two main folders: *Simulations_protein_A* and *Simulations_proteins_A_B*, which separate simulations involving only protein A from those that include both proteins A and B. Within each main folder, subfolders are categorized by different values of the parameter α. Further subdivision is provided based on the number of DNA sites (3000 or 6000). For simulations involving protein B, an additional organizational level is introduced, with sub folders further divided according to the parameter K.

To avoid file overwriting, all simulation outputs are saved with filenames that include the corresponding plot title and the key parameters used in the simulation. Additionally, each figure contains a legend specifying all relevant parameters, enabling straightforward tracking and identification of the simulation conditions.

8 Conclusions

This study began with an analysis of various energy parameters, which provided valuable insights into the limitations of the parameter space and identified regions where the system exhibits distinct behaviors. By investigating the effects of energy values on binding patterns, we pinpointed areas where significant changes in transcription factor binding dynamics occur. We also identified regions where further adjustments to the energy parameters no longer meaningfully influence the system's behavior.

The observation of scale invariance was crucial in selecting a computationally efficient system size for the subsequent analysis.

To observe a phase transition, it was essential to maintain E_{ad} at a low enough value to facilitate rapid unbinding in cases where A proteins were not bound to B proteins. While E_{aa} mainly contributed to the cohesion of clusters, the parameters k/L and E_{ba} proved to be crucial for observing phase transitions and more clustered behavior. Fine-tuning these parameters is essential for achieving the desired system dynamics.

Looking ahead, the implementation of a *Cluster class* has been initiated, which could allow to track the residence time of clusters and monitor their dynamics over time. At present, the simulation does not distinguish whether a 10-sized cluster results from a single TF unbinding within a 20-sized cluster or from other interactions.

9 Tables

Parameters	**Brief Explication**	**Range of Values Investigated**	**Comments**
N	Total number of binding sites in the DNA	Short DNA runs: 3000, Long DNA runs: 6000	Length of a promoter region on the DNA is typically 100-1000 bp
alfa	Density of As on the DNA	Short DNA runs: 0.15, 0.3, Long DNA runs: 0.15	
nA	Number of As in the simulation (computed as int(alfa * N))	Short DNA runs: 450, 900, Long DNA runs: 900	
nB	Number of Bs	Short DNA runs: 100, Long DNA runs: 200	
k	Number of B interacting sites in the DNA	Short DNA runs: 2, 5, 10, Long DNA runs: 2, 10	
L	Distance (in terms of binding sites in the DNA) from one binding site in B protein to the other	Short DNA runs: 10, Long DNA runs: 10	
stop_time	Time to stop the simulation	Short DNA runs: 2×10^6, Long DNA runs: 4×10^6	
ignoring_time	Time to ignore before taking sample	Short DNA runs: 1×10^6, Long DNA runs: 2×10^6	
m	Take sample every mth step	Short DNA runs: 50, Long DNA runs: 50	m must be a divisor of stop_time - ignoring_steps
E_ad	Energy of A binding to DNA	Short DNA runs: 0, 1, 3, Long DNA runs: 1	Typically in the order of k_BT
E_aa	A influence on its neighbors	Short DNA runs: 0, 2, Long DNA runs: 2	
E_ab	B influence on A unbinding event	Set to infinity	
E_ba	A influence on B unbinding event	Short DNA runs: 0, 1, 2, 4, Long DNA runs: 0, 4	

Table 1: Simulation Parameters and Their Values

Table 2: Plots Used in the Analysis

Plots	**Brief Description**	**Axis**	**Used For**	**Comments**
First binding time of As that don't unbind	Plot tracking the binding status of each TF over time, identifying those that remain bound throughout the simulation.	Y: Binding Time; X: Index of each TF (0 to $n_A - 1$)	Investigate TFs with extremely stable binding; determine if they were first bound at the start or end of the simulation.	Results not in GitHub repository; used for initial analysis but deemed irrelevant.
Frequency of cluster sizes of As	Histogram. For each time sampled the sizes of the consecutive As on the DNA are collected. At the end frequency of each size is computed and then divided for the number of time step collected .	Y: Frequency; X: Cluster Size	Analyze cluster distribution across different parameter combinations.	
Frequency of mean residence time per A	Histogram of the frequency of mean residence times computed for each TF at the end of the simulation.	Y: Frequency; X: Mean Residence Time	Analyze the distribution of mean residence times.	
Scatter plot of Std Dev of residence time for each A	Scatter plot showing the standard deviation of residence times for each TF.	Y: Std Dev; X: Index of each TF (0 to $n_A - 1$)	Display variability of residence times across different TFs and parameter sets.	Results not in GitHub repository; used for initial analysis but deemed irrelevant.
Time average of B occupied site fraction	Scatter plot of the average fraction of occupied sites per B over time.	Y: Occupied Sites / Total Sites; X: Index of each B (0 to $n_B - 1$)	Show average occupancy of Bs.	

Continued on the next page

Table 2 – Continued from previous page

Plots	**Brief Description**	**Axis**	**Used For**	**Comments**
Bs bound to As on DNA for two different time steps	Histograms showing the presence (1) or absence (0) of Bs bound to DNA sites at intermediate and final time steps.	Y: Frequency; X: Index of DNA Sites	Show clustering and distribution of Bs at different time steps under various parameters.	
Frequency of number of binding events per A	Histogram showing the distribution of the number of binding events for all As.	Y: Frequency; X: Number of Binding Events	Show the impact of parameters on binding events and fluidity of the system.	
Time series of bound As	Plot showing the progression of the number of As bound on the DNA over time.	Y: Number of A Bound; X: Time Step	Show if the system reached a steady state and explore the parameter space for energies.	
Mean cluster size vs. max cluster size for each E_{aa}	Plot comparing mean and max cluster sizes of As for each E_{aa}.	Y: Cluster Size; X: E_{aa}	Analyze cluster formation dynamics.	Results not in GitHub repository; used for initial analysis but deemed irrelevant.
Log(mean residence time) vs. E_{ad} for several E_{aa}	Plot of natural log of mean residence time for each E_{ad} value, showing values for different E_{aa} values.	Y: ln(Mean Residence Time); X: E_{ad}	View how TF binding dynamics change with different energy scales.	
Std Dev of residence time vs. E_{ad} for several E_{aa}	Plot of the mean standard deviation of residence times for different E_{ad} values across all TFs.	Y: Mean (Stdev Residence Time); X: E_{ad}	View how TF binding dynamics change with different energy scales.	

Continued on the next page

Table 2 – Continued from previous page

Plots	**Brief Description**	**Axis**	**Used For**	**Comments**
Ratio of mean / Std Dev of residence time vs. E_{ad} for several E_{aa}	Plot showing the ratio of mean to standard deviation (refer to the plot above) of residence times for different E_{ad} values.	Y: Mean / Stdev; X: E_{ad}	View how TF binding dynamics change with different energy scales.	

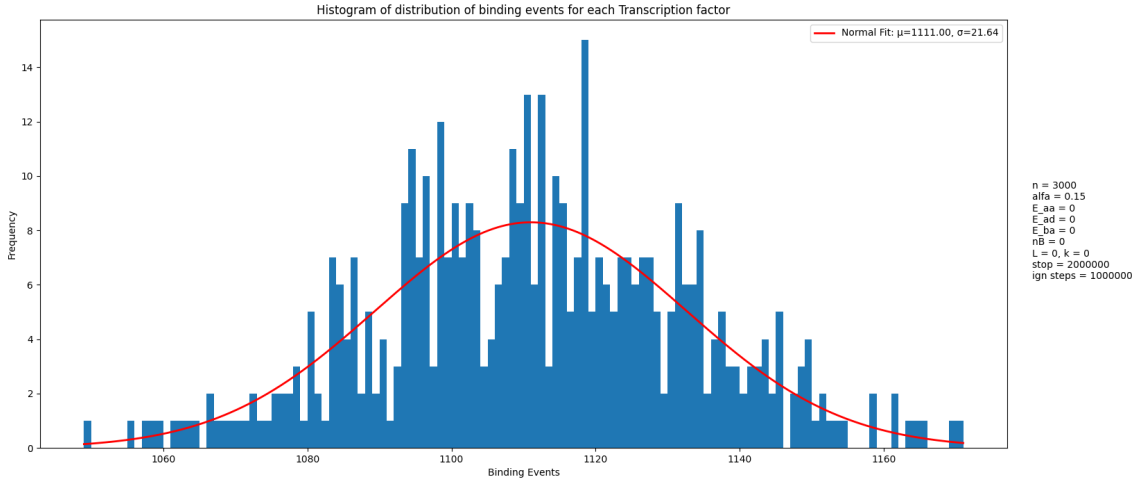
Condition	**Cluster Size**
Low k/L**, Low** E_{ba}	High one-sized, low maximum cluster size, low intermediate-sized clusters
Intermediate k/L**, Low** E_{ba}	Fewer one-sized clusters, maximum cluster size around 80, low intermediate-sized clusters
High k/L**, Low** E_{ba}	Fewer one-sized clusters, maximum cluster size around 60, more intermediate clusters
Low k/L**, High** E_{ba}	Quite high one-sized clusters, maximum cluster size around 100, low intermediate-sized clusters
Intermediate k/L**, High** E_{ba}	Slightly more than 20 one-sized clusters, maximum cluster size around 60, high intermediate-sized clusters
High k/L**, High** E_{ba}	Slightly fewer than 20 one-sized clusters, maximum cluster size around 40, high intermediate-sized clusters

Table 3: Cluster size across different conditions of k/L and E_{ba}.

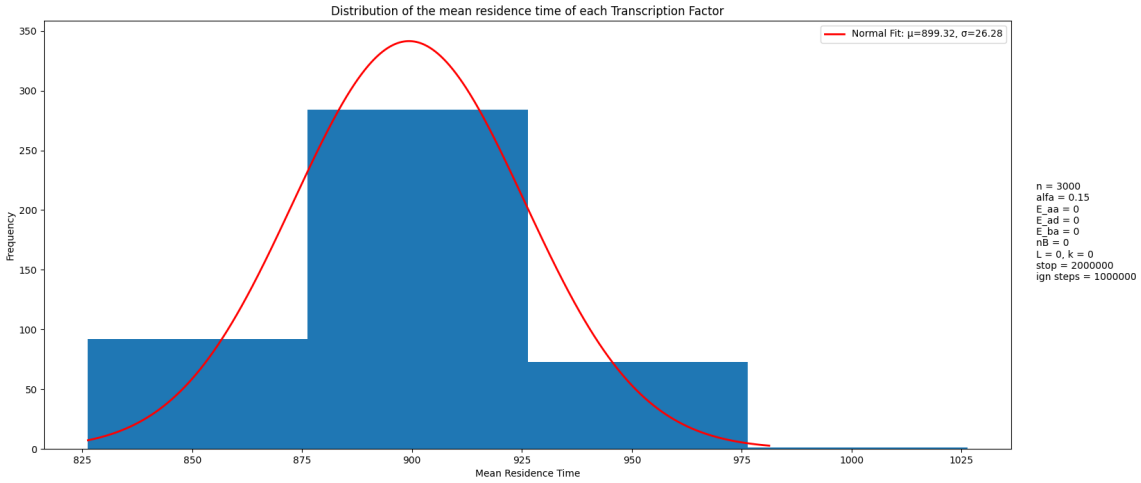
Condition	**B Distribution**
Low k/L**, Low** E_{ba}	Very fragmented
Intermediate k/L**, Low** E_{ba}	More clustered than above
High k/L**, Low** E_{ba}	Even more clustered
Low k/L**, High** E_{ba}	Very clustered
Intermediate k/L**, High** E_{ba}	Quite fragmented
High k/L**, High** E_{ba}	More fragmented

Table 4: B distribution across different conditions of k/L and E_{ba}.

10 Figures



(a) Distribution of binding events in the random case, where all energies are set to zero.



(b) Distribution of mean residence time in the random case, where all energies are set to zero.

Figure 2: Binding events and mean residence time in the random case, where all energies are set to zero. For both distributions, a Gaussian fit is applied, with the mean and standard deviation displayed. The parameter values used are provided in the legend on the right side of the figure.

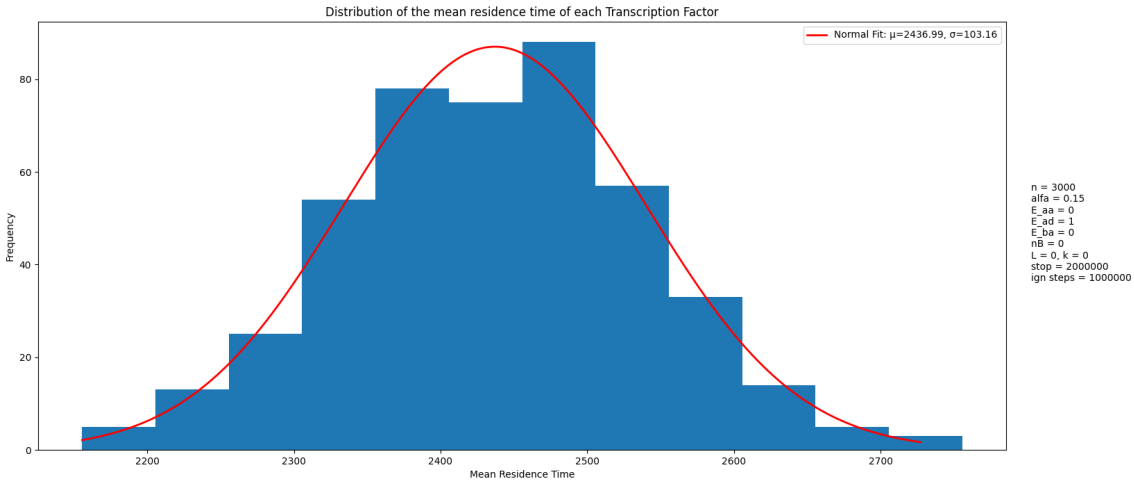
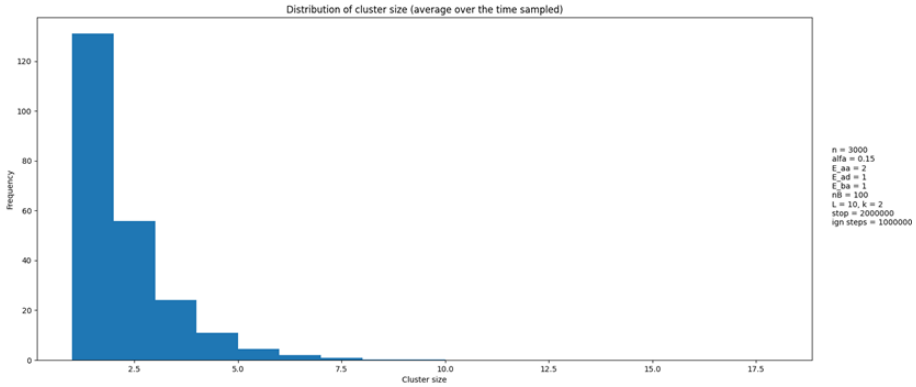
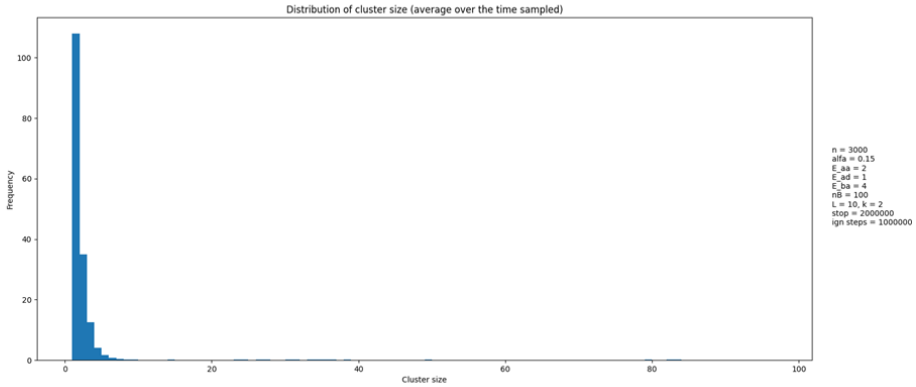


Figure 3: Distribution of mean residence time with $E_{\text{ad}} = 1$, all the other energies are set to zero. A Gaussian fit is applied, with the mean and standard deviation displayed. The parameter values used are provided in the legend on the right side of the figure.

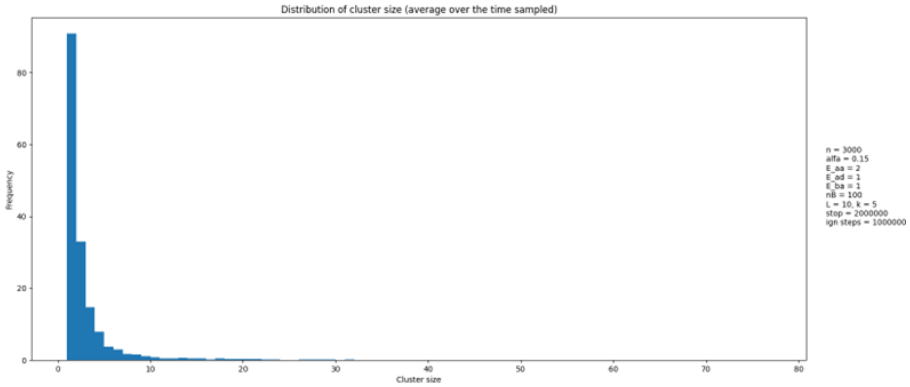


(a) Cluster size distribution for $k/L = 0.2$ and $E_{ba} = 1$

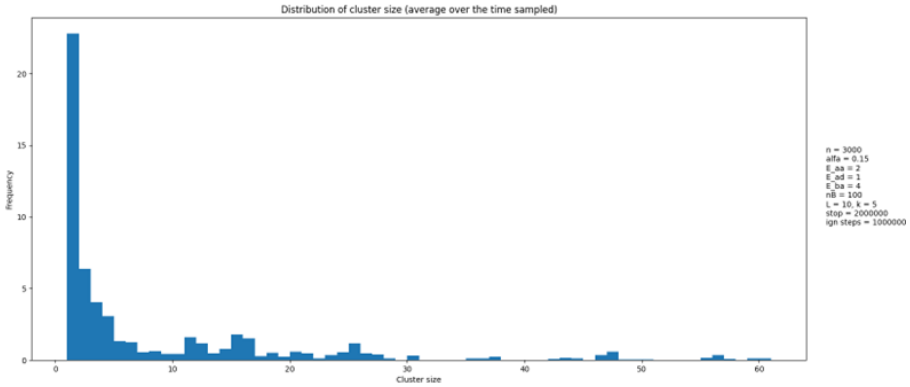


(b) Cluster size distribution for $k/L = 0.2$ and $E_{ba} = 4$

Figure 4: Comparison of cluster size distributions for $k/L = 0.2$ at different values of E_{ba}: weak ($E_{ba} = 1$) and strong ($E_{ba} = 4$) interactions.The parameter values used are provided in the legend on the right side of the figure.

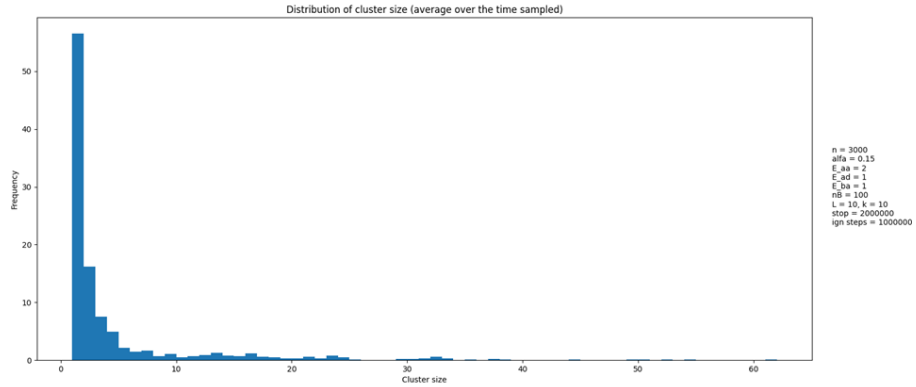


(a) Cluster size distribution for $k/L = 0.5$ and $E_{ba} = 1$

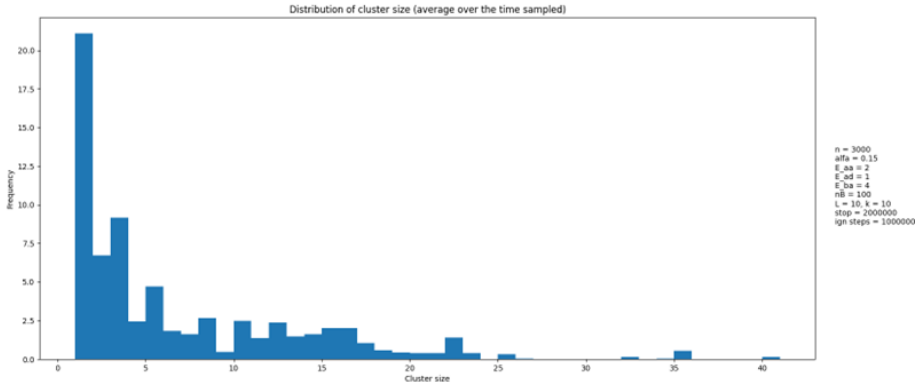


(b) Cluster size distribution for $k/L = 0.5$ and $E_{ba} = 4$

Figure 5: Comparison of cluster size distributions for $k/L = 0.5$ at different values of E_{ba}: weak ($E_{ba} = 1$) and strong ($E_{ba} = 4$) interactions. The parameter values used are provided in the legend on the right side of the figure.

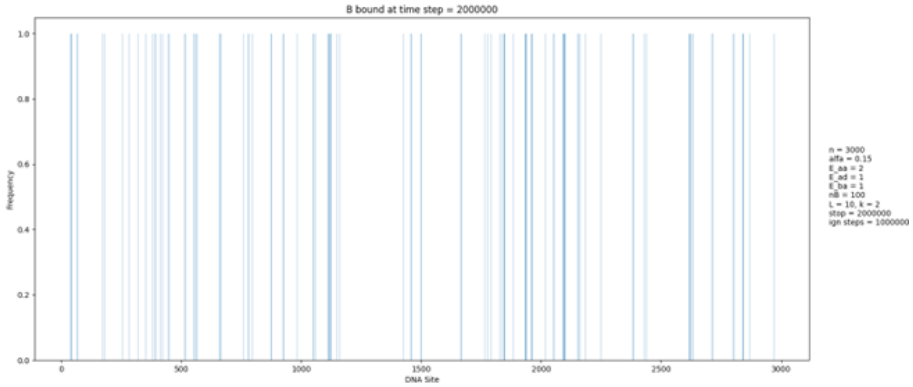


(a) Cluster size distribution for $k/L = 1$ and $E_{ba} = 1$

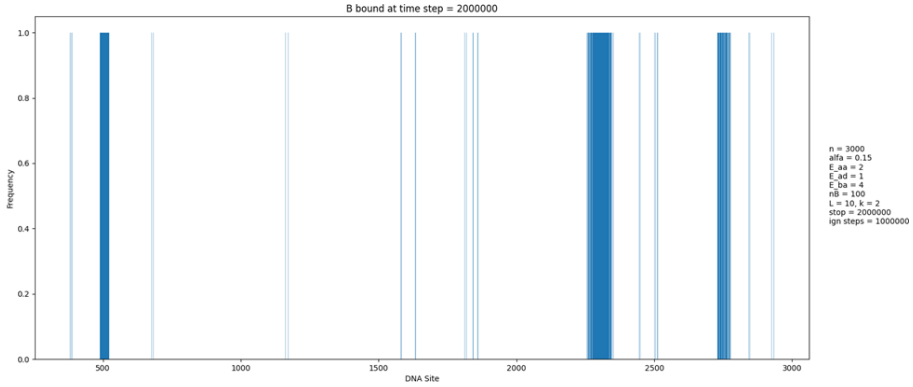


(b) Cluster size distribution for $k/L = 1$ and $E_{ba} = 4$

Figure 6: Comparison of cluster size distributions for $k/L = 1$ at different values of E_{ba}: weak ($E_{ba} = 1$) and strong ($E_{ba} = 4$) interactions. The parameter values used are provided in the legend on the left side of the figure.

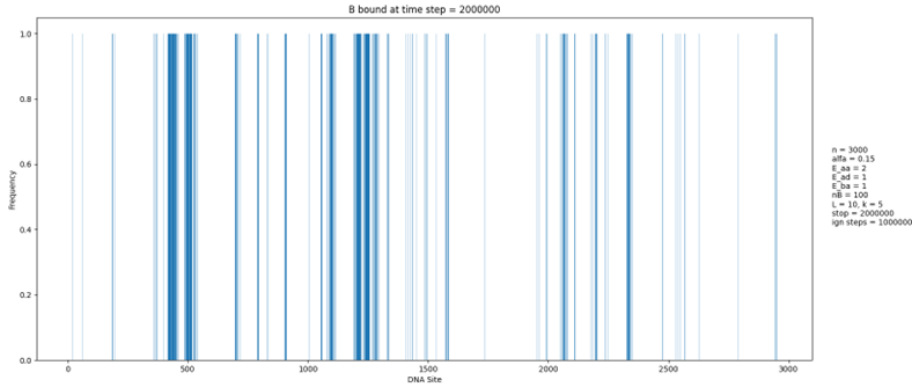


(a) Binding of B at the final time step for $k/L = 0.2$ and $E_{ba} = 1$.

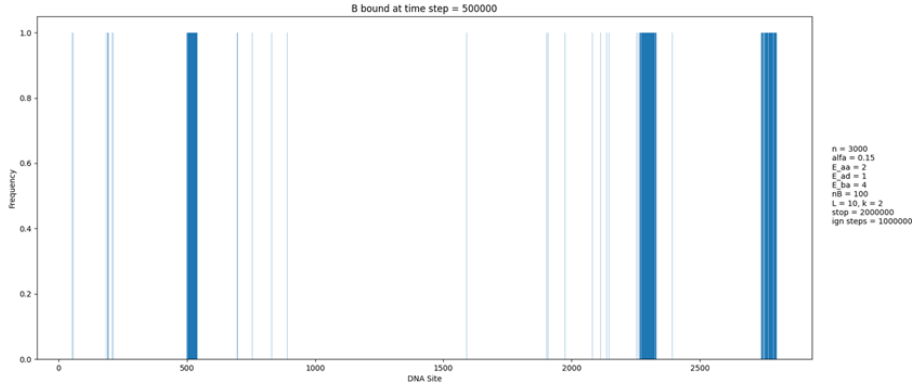


(b) Binding of B at the final time step for $k/L = 0.2$ and $E_{ba} = 4$.

Figure 7: Comparison of the binding of B at the final time step for $k/L = 0.2$ with different E_{ba} values: weak ($E_{ba} = 1$) and strong ($E_{ba} = 4$) interactions. The X-axis represents DNA sites, with a bin indicating where a B is bound. The parameter values used are provided in the legend on the right side of the figure.

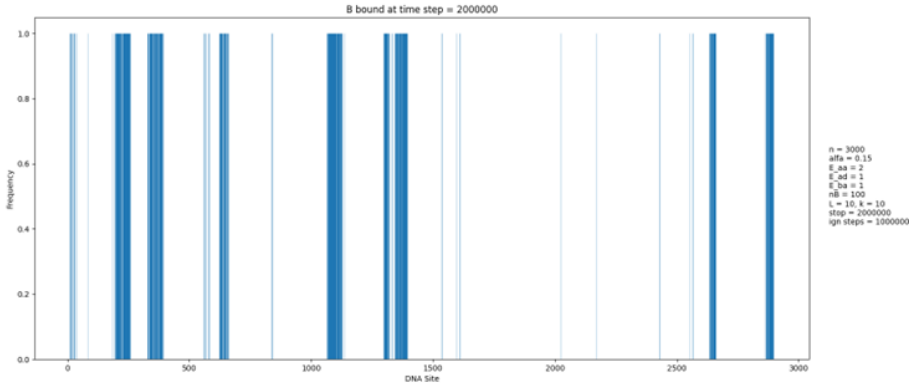


(a) Binding of B at the final time step for $k/L = 0.5$ and $E_{ba} = 1$.

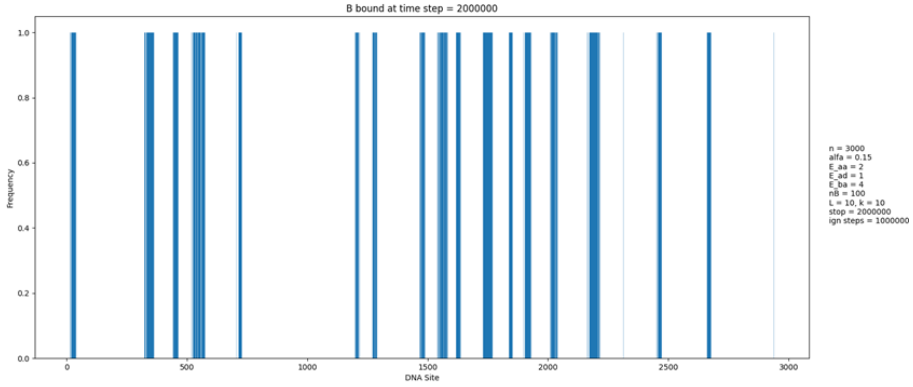


(b) Binding of B at the final time step for $k/L = 0.5$ and $E_{ba} = 4$.

Figure 8: Comparison of the binding of B at the final time step for $k/L = 0.5$ with different E_{ba} values: weak ($E_{ba} = 1$) and strong ($E_{ba} = 4$) interactions. The X-axis represents DNA sites, with a bin indicating where a B is bound. The parameter values used are provided in the legend on the right side of the figure.

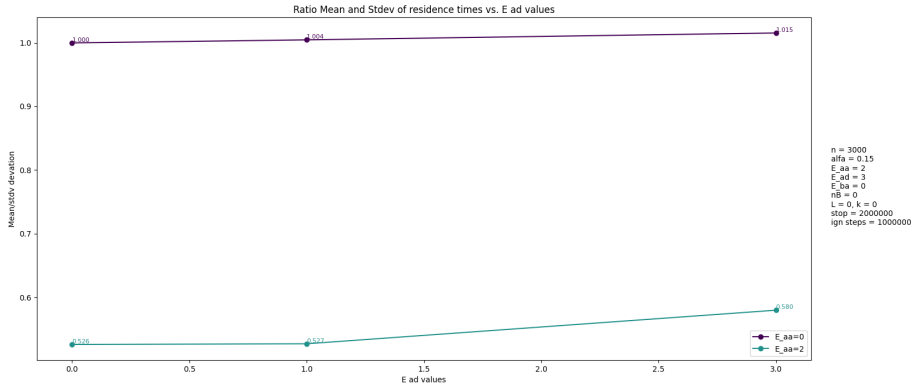


(a) Binding of B at the final time step for $k/L = 1$ and $E_{ba} = 1$.

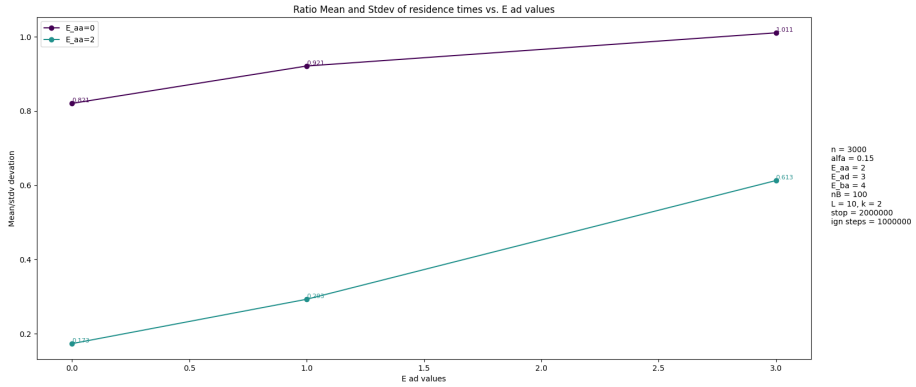


(b) Binding of B at the final time step for $k/L = 1$ and $E_{ba} = 4$.

Figure 9: Comparison of the binding of B at the final time step for $k/L = 1$ with different E_{ba} values: weak ($E_{ba} = 1$) and strong ($E_{ba} = 4$) interactions. The X-axis represents DNA sites, with a bin indicating where a B is bound. The parameter values used are provided in the legend on the right side of the figure.



(a) Ratio of Mean Residence Time to Mean Standard Deviation for the no B case.



(b) Ratio of Mean Residence Time to Mean Standard Deviation when $E_{\text{ba}} = 4$.

Figure 10: Comparison of the ratio of Mean Residence Time to Mean Standard Deviation. In both panels, two curves are shown: $E_{\text{aa}} = 0$ (purple) and $E_{\text{aa}} = 2$ (light blue). The parameter values used are provided in the legend on the right side of the figure.

References

[1] Annual Review of Cell and Developmental Biology, "Transcription Factors and Gene Regulation," *Ann. Rev. Cell Dev. Biol.*, 2023, `annurev-cellbio-022823-013847.pdf`

[2] Bioinformatics, "Title of the Paper," *BMC Bioinformatics*, 2022, `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-05090-2`

[3] "Transcription Factors and Their Role in Gene Expression," *BioRxiv*, 2024, `https://www.biorxiv.org/content/10.1101/2024.11.01.621483v1#:~:text=Transcription%20factors%20(TFs)%20often%20form,that%20bring%20in%20more%20TFs.`

[4] Donovan et al. 2019, "Study on Transcription Factors," *PMC*, https://pmc.ncbi.nlm.nih.gov/articles/PMC10482752/

[5] "The Length of Promoter Region on DNA," *PMC*, https://pmc.ncbi.nlm.nih.gov/articles/PMC6848157/#:~:text=Promoters%20can%20be%20about%20100,DNA%20transcription%20and%20RNA%20polymerase.

[6] "Biological Order and k_BT," *ScienceDirect*, 2023, https://www.sciencedirect.com/science/article/pii/S0006349523006616?via%3Dihub