EMERGENT STRUCTURE AND DYNAMICS FROM STOCHASTIC PAIRWISE CROSSLINKING IN CHROMOSOMAL POLYMER MODELS

Benjamin L. Walker

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Mathematics.

Chapel Hill
2021

Approved by:

Katherine Newhall

M. Gregory Forest

Kerry Bloom

Peter Mucha

David Adalsteinsson

ABSTRACT

Benjamin L. Walker: Emergent Structure and Dynamics from Stochastic Pairwise Crosslinking in Chromosomal Polymer Models
(Under the direction of Katherine Newhall)

The spatio-temporal organization of the genome is critical to the ability of the cell to store huge amounts of information in highly compacted DNA while also performing vital cellular functions. Experimental methods provide a window into the geometry of the chromatin but cannot provide a full picture in space and time. Polymer models have been shown to reproduce properties of chromatin and can be used to make simulated observations, informing biological experimentation. We apply a previously-studied model of the full yeast genome with dynamic protein crosslinking in the nucleolus which showed the emergence of clustering when the crosslinking timescale was sufficiently fast. We investigate the the crosslinking timescale at finer resolution and newly identify the presence of a *flexible clustering* regime for intermediate timescales, which maximizes mixing of nucleolar beads, of significant interest due to the role mixing plays in nuclear processes. In order to robustly identify spatio-temporal clustering structure, we map our problem to a multi-layer network and then apply the multi-layer modularity community detection algorithm, showing the presence of spatio-temporal community structure in the fast and intermediate clustering regimes. We perform analysis of the relationship between cluster size and the ensuing stability of clusters, revealing a heterogeneous collection of clusters in which cluster size correlates with stability. We view the stochastic switching as producing an effective thermal equilibrium by extending the WKB approach for deriving quasipotentials in switching systems to the case of an overdamped Langevin equation with switching force term, and derive the associated Hamilton-Jacobi equation. We apply the string method for finding most-probable transition paths, revealing previously unreported numerical challenges; we present modifications to the algorithms to overcome them. We show that our methods can correctly compute asymptotic escape times by comparison to Monte Carlo simulations, and verified an important principle: the effective force is often significantly weaker than a naive average of the switching suggests. Through this multifaceted approach, we have shown how stochastic crosslinking leads to complex emergent structure, with different timescales optimizing different properties, and shown how the structure can be analyzed using both network data based tools and through stochastic averaging principles.

To my father, my mother, and my sister, for their unfailing support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

CHAMP	Convex Hull of Admissible Modularity Partitions
CTMC	Continuous-Time Markov Chain
DNA	Deoxyribonucleic Acid
GMAM	Geometric Minimum Action Method
Hi-C	High-throughput conformation capture
kbp	Kilo basepairs
SMC	Structural Maintenance of Chromosomes (protein)
TAD	Topologically Associated Domain
WKB	Wentzel–Kramers–Brillouin (asymptotic approximation)

CHAPTER 1: INTRODUCTION AND REVIEW OF LITERATURE[1]

1.1 Polymer Modeling of Chromosomes

Chromosomes inside the nucleus are long, complex molecules consisting of DNA combined with a variety of proteins in a complex structure referred to as chromatin. The chromatin experiences an extraordinary level of compaction, with a typical human cell containing approximately two meters of DNA in a nucleus with a radius of mere micrometers. Despite this fact, the cell is able to use the DNA in a variety of cellular processes, including duplicating and dividing it in the cell division process, as well as nuclear processes such as DNA transcription and repair that make direct use of the information contained in the base pairs. The spatial geometry of the DNA must be able to facilitate all of these processes occurring simultaneously, and so in order to understand how the cell performs these processes with complex and competing objectives, we wish to understand the spatio-temporal organization of the DNA inside the cell nucleus.

Due to the microscopic scale of DNA and this extreme compaction, direct observation of its dynamics is not possible. Instead, a common approach to develop understanding of the dynamics and organization of DNA is to model the chromatin as a polymer, allowing application of previously models from the field of polymer physics, dating back to early work in (Marko & Siggia 1994*a*, Marko & Siggia 1994*b*, Marko & Siggia 1995, Marko & Siggia 1997). This work applied theory from polymer physics to models of DNA as a helical polymer, and showed that experimental observations of DNA supercoiling could be explained based on these simple and fundamental physical properties of polymers. This observation that despite the highly complicated macromolecular compound structure of the chromatin, many experimental observations can be explained by simple polymer physics has set the stage for a diverse field of research in understanding the nuclear processes and chromosome geometry through an integrated approach of experimental research which guides development of polymer models, the results from which then in return guide future experimental research.

[1] Section 1.4 contains material adapted from work previously published in PLOS Computational Biology under the CC-BY license (`https://creativecommons.org/licenses/by/4.0/`). The original citation is: Walker, Benjamin, Dane Taylor, Josh Lawrimore, Caitlin Hult, David Adalsteinsson, Kerry Bloom & M Gregory Forest. 2019. "Transient crosslinking kinetics optimize gene cluster interactions." *PLoS Computational Biology* 15(8):e1007124.

Because the DNA cannot be observed directly on a microscale level, a core consideration is the type of experimental data that can be obtained, and the ways in which this can be used to inform and verify modeling. An experimental technique that is frequently used as a benchmark for understanding chromosome dynamics, and correspondingly can be applied in modeling, is the tagging of DNA with fluorescent proteins that can be observed through a microscope (Belmont & Straight 1998, Belmont, Li, Sudlow & Robinett 1999, Pearson, Maddox, Salmon & Bloom 2001, Fisher, Ballenger, O'Brien, Haase, Superfine & Bloom 2009, Hajjoul, Mathon, Ranchon, Goiffon, Mozziconacci, Albert, Carrivain, Victor, Gadal, Bystricky & Bancaud 2013). This creates a visible spot whose motion can be measured and characterized. Comparisons of these observed fluctuations with mathematical predictions of what fluctuations would be observed are then used to show agreement between the model and the biology. In a similar note, observations of segregation of chromosomes; that is, the tendency of separate chromosomes to each occupy separate "compartments" in the nucleus, as opposed to being intertwined together, has also been shown to be explained by polymer models of chromosomes (Rosa & Everaers 2008, Vasquez, Hult, Adalsteinsson, Lawrimore, Forest & Bloom 2016).

Experimental measurements of chromosome structure were revolutionized by the introduction of "Chromosome Conformation Capture" (3C) (Dekker, Rippe, Dekker & Kleckner 2002). In chromosome conformation methods, the DNA inside the nuclei of a population of cells is subjected to formaldehyde, which links together nearby genetic loci. Then, the DNA is chopped into small pieces, creating bound pairs of loci that were nearby during the formaldehyde phase. In the early 3C approach, data is only extracted about the frequency with which a single selected pair was observed, over the population of cells. This was extended in 4C to compare a single selected locus with all other loci, and finally in Hi-C to compare simultaneously all possible pairs of loci across the entire genome.

The allow for the construction of full-genome "contact maps," representations of which sections of the genome tend to be close to, or in contact with, which other sections (Dekker 2008, Lieberman-Aiden, Van Berkum, Williams, Imakaev, Ragoczy, Telling, Amit, Lajoie, Sabo, Dorschner et al. 2009, Duan, Andronescu, Schutz, McIlwain, Kim, Lee, Shendure, Fields, Blau & Noble 2010, Tanizawa, Iwasaki, Tanaka, Capizzi, Wickramasinghe, Lee, Fu & Noma 2010, Pope, Ryba, Dileep, Yue, Wu, Denas, Vera, Wang, Hansen, Canfield et al. 2014, Vian, Pękowska, Rao, Kieffer-Kwon, Jung, Baranello, Huang, El Khattabi, Dose, Pruett et al. 2018). As, Hi-C methods in particular allow for the construction of such maps simultaneously across the entire genome, this can be thought of as a representation of the population-averaged spatial geometry in terms of pairwise distances. This approach has revealed information about chromatin structure such as the existence of “Topologically Associated Domains,” regions of chromosomes that tend to be associated closer together than other parts of the genome.

One drawback of Hi-C methods is that the accumulation of large quantities of DNA pairs is done using a large population of cells, which means that the resulting data is a population average which may not be representative of the geometry of the DNA of a single cell at any particular point in time. This sets the stage for a renewed interest in computational modeling – using the model to allow for investigation of the spatio-temporal structure that is hidden by these population averages.

These techniques have revealed the presence of complex hierarchical looping structure in the genome, with chromosomes forming a fractal-like loops-within-loops structure (Dekker et al. 2002, Le, Imakaev, Mirny & Laub 2013, Rao, Huntley, Durand, Stamenova, Bochkov, Robinson, Sanborn, Machol, Omer, Lander et al. 2014). The 4D Nucleome Project, described in (Dekker, Belmont, Guttman, Leshyk, Lis, Lomvardas, Mirny, O'shea, Park, Ren et al. 2017), provides an overview of how this variety of methods, including Hi-C and other experimental approaches along with spatio-temporal modeling, can be combined to provide the next level of understanding of mechanics of the genome.

However, the static pictures of the geometry given by Hi-C are blind to the specifically dynamic aspects of the organization, and these averages may not be representative of the state of the chromosomes at any particular point. One particularly important dynamic consideration is that of mixing or communication of the genome. Cellular processes such as homology search rely on finding matching genes, which requires parts of the genome to communicate with a wide range of other genes. Recent literature has shown the importance of this communication (Dekker & Mirny 2016), sought to develop extensions to Hi-C to capture this sort of information (Schmitt, Hu, Jung, Xu, Qiu, Tan, Li, Lin, Lin, Barr et al. 2016), and develop models of that help explain how these temporal variations contribute to nuclear processes (Ghosh & Jost 2018).

Hi-C experiments provide detailed data about chromosomal geometry, albeit in the form of pairwise compact maps, which act as a sort of proxy for pairwise distance. This suggests what is in essence an inverse problem setup, seeking underlying 3-dimensional geometries that would agree with observed Hi-C data. Of course, the fact that the Hi-C data contains experimental noise and represents an average across a population of cells complicates this aspect. An approach that has arisen to address this inverse problem is a hybrid approach of polymer modeling and constrained or informed by experimental contact maps (Fudenberg & Mirny 2012, Serra, Di Stefano, Spill, Cuartero, Goodstadt, Baù & Marti-Renom 2015). In a similar idea, (Di Pierro, Zhang, Aiden, Wolynes & Onuchic 2016) constructs an empirical free-energy function by treating the chromosome as a base polymer with the imposition of additional interaction energies to capture the heterogeneity seen in the Hi-C data. Choosing these interaction energies to agree with the contact map data, application of the maximum entropy principle allows for a representation of an "idealized chromosome" to be created. Overall, polymer modeling approaches to understanding chromosome conformation can be viewed as coming from two angles: those that fundamentally derive from experimental data, as described in this paragraph, and those that are based in biological and physical principles.

We note that the organization of the DNA changes significantly during the course of the cell cycle – most notably, during the cell division process, the entire genetic material must be duplicated, and then divided evenly into what will become each of the daughter cells. Motivated by a desire to understand the complex nuclear processes that underly this division, models of the mitotic genome have been extensively studied relative to experimental observations (Belmont 2006, Maeshima, Hihara & Takata 2010, Naumova, Imakaev, Fudenberg, Zhan, Lajoie, Mirny & Dekker 2013, Cheng, Heeger, Chaleil, Matthews, Stewart, Wright, Lim, Bates & Uhlmann 2015). However, in this work we will restrict the focus to interphase portion of the cell cycle, in which the cell is not undergoing division. During this period, the DNA is expected to be organized in order to facilitate different nuclear processes, including most notably transcription and gene expression, in which the DNA is used as a blueprint for the cell to manufacture proteins.

Any understanding of the spatial organization of the chromosomes requires understanding the processes that cause heterogeneity in its structure. One particular unique area of the genome is the nucleolus, a section of DNA on one chromosome consisting of a repeated sequence of the same base pairs. Hi-C experimental methods are blind to the structure of the nucleolus, as the repetition means it is impossible to identify where in the nucleolus a particular DNA fragment came from. This shortcoming of Hi-C methods underscores a unique need for biologically-inspired modeling to understand structure in the nucleolus. (Albert, Mathon, Shukla, Saad, Normand, Léger-Silvestre, Villa, Kamgoue, Mozziconacci, Wong et al. 2013) showed agreement between experimental imaging data and a polymer model of the nucleolus, suggesting that polymer models would continue to be effective in this area. Experimental observations show the nucleolus to be denser and segregated from the rest of the genome, able to be viewed as a "membraneless organelle" (Feric, Vaidya, Harmon, Mitrea, Zhu, Richardson, Kriwacki, Pappu & Brangwynne 2016), separated by a dynamic phase separation instead of a physical boundary. Further experiments have shown that mutations to genes that control SMC proteins also affect the observed structure of the nucleolus, suggesting that the SMC proteins are responsible for this phase separation (Albert, Colleran, Leger-Silvestre, Berger, Dez, Normand, Perez-Fernandez, McStay & Gadal 2013).

The work in this dissertation begins applies a model that was developed to represent the genome of a yeast cell during the interphase segment of the cell cycle, in which it is not undergoing division. Early work over (Wong, Marie-Nelly, Herbert, Carrivain, Blanc, Koszul, Fabre & Zimmer 2012, Wong, Arbona & Zimmer 2013, Vasquez & Bloom 2014, Wang, Mozziconacci, Bancaud & Gadal 2015) developed an entropic polymer bead-spring model for DNA, discretizing with beads representing approximately 5kbp of DNA. In addition to the polymer spring force, beads are also subject to an excluded-volume repulsion force that takes the form of a pairwise energy penalty that is Gaussian in the distance between the two beads. As the 5kbp of DNA that each bead represents is itself a floppy polymer, this Gaussian excluded volume force treats the bead as a soft volume. The beads are also constrained to remain within a spherical nucleus, modeled using a hard wall. The polymer representing the DNA is also tethered to this wall at both ends; one end a centromere and the other end a telomere site.

Subsequent work in (Vasquez et al. 2016) extended this polymer model to the entire yeast genome, representing it using 24 arms containing 2803 beads. Computational investigation showed that this entropic polymer model was able to reproduce observations of the DNA, such as the segregation of different chromosome arms into different regions (Tjong, Gong, Chen & Alber 2012, Verdaasdonk, Vasquez, Barry, Barry, Goodwin, Forest & Bloom 2013, Ea, Baudement, Lesne & Forné 2015, Vasquez et al. 2016).

The spatio-temporal organization of the genome is understood to be highly influenced not only by the DNA behaving like a simple polymer, but by the action of Structural Maintenance of Chromosomes (SMC) Proteins, such as condensin and cohesin, on the genome. The specific actions of these proteins are not yet well understood, but are hypothesized to include crosslinking and loop extrusion. Experimental results have shown them to affect observable traits of the structure (Sakai, Mochizuki, Kinoshita, Hirano & Tachikawa 2018), and in particular it has been shown that specifically dynamic binding-unbinding action of proteins, as opposed to just lasting binding, is responsible for these effects (Brackley, Liebchen, Michieletto, Mouvet, Cook & Marenduzzo 2017).

Inspired by these results and a desire to better understand the dynamics of the nucleolus, (Hult, Adalsteinsson, Vasquez, Lawrimore, Bennett, York, Cook, Yeh, Forest & Bloom 2017) builds on the model from (Vasquez et al. 2016) by adding SMC proteins modeled as stochastic pairwise crosslinkers acting on nucleolar beads. In addition to the previous forces in the model, pairs of beads that are sufficiently close will sometimes form a dynamic bond, modeled as an additional strong force between them. This spring will hold the beads together for some period of time before breaking. The formation and breaking is governed by an active/inactive system where beads enter an active state for some time, allowing them to form bonds with other nearby active beads, and when a bound bead switches to the inactive state, the bond breaks, and that bond remains unable to form bonds until it returns to the active state.

The timescale of this active/inactive switching is governed by the dynamic timescale parameter μ, which corresponds to the mean duration that beads remain in the active state. The mean time in the inactive state is coupled to this parameter as $\mu/9$. Waiting times in this model are normally distributed, with a standard deviation set to one fifth of the mean.

The work in (Hult et al. 2017) considered values of $\mu \in \{0.09, 0.9, 90\}$ (expressed in simulation seconds) and observed that for $\mu = 0.09$, the nucleolar beads condensed into large stable clusters, reminiscent of the membraneless phase separation previously seen (Feric et al. 2016). However, at larger values of μ, this same behavior is not observed.

The position, shape, and movement of the nucleolus can be observed experimentally by tagging nucleolar DNA with fluorescent proteins and capturing photos or videos through a microscope. In order to allow for direct comparisons to be made between simulations of the model and this experimental imaging data, (Hult et al. 2017) also includes a microscope simulator, which convolves the point-spread function generated by a 5kbp section of fluorescent DNA under a microscope with the positions of the nucleolar beads, and thereby generates a simulated image representing what the current nucleolar structure would look like under a microscope. This link allows for the observed imaging data to inform choices of modeling parameters, and conversely for modeling results to inform future experimentation.

The first goal of this dissertation is to map out on a fine resolution the relationship between the timescale of crosslinking action in the model from (Hult et al. 2017) and the resulting emergent spatio-temporal clustering. In the first half of Chapter 2, we will perform simulations on a set of 20 values for the timescale parameter μ spaced logarithmically on the range $[0.09, 90]$ and develop a set of metrics to measure the spatio-temporal behavior for each result, allowing us to map out the dependence of this behavior on μ. We will show that the clusters identified for $\mu = 0.09$ quickly begin to dissolve as $\mu > 0.1$, but that this paves the way for an intermediate μ, flexible clustering regime that maximizes the cross-communication of beads. For $\mu > 1$ we see these flexible clusters disappear as the nucleolus becomes essentially homogeneous. These results set the stage for the next goal of the dissertation, which is to develop tools for automatically extracting spatio-temporal clusters from this data, and use this to further understand the dynamics of the emergent structure.

1.2 Network Analysis of Genomic Interaction Data

The results of (Hult et al. 2017) showed presence of emergent spatio-temporal clustering behavior in the nucleolus based on dynamic binding-unbinding crosslinking. This motivates the development of tools to extract spatio-temporal clustering information from the model simulations, so that bead dynamics can be analyzed in terms of the clusters present. We will look for inspiration in the field of network science.

Literature on analysis of Hi-C contact maps has made frequent use of network modeling based approaches (Rajapakse, Scalzo, Tapscott, Kosak & Groudine 2010, Rajapakse & Groudine 2011, Rajapakse, Groudine & Mesbahi 2011). In this framework, we represent sections of the genome as nodes in a network, connected by edges whose weight represents the frequency or strength of interaction. This approach allows for the existing library of network analysis literature to be extended to analysis of chromosomal geometry, such as through measures of centrality (Liu, Chen, Ronquist, Seaman, Ceglia, Meixner, Muir, Chen, Higgins, Baldi et al. 2017, Liu, Chen, Hero & Rajapakse 2018).Striving to identify TADs and other regions of association in the data, another common network analysis applied is that of community detection (Chen, Hero III & Rajapakse 2016, Cabreros, Abbe & Tsirigos 2016). In the framework of networks, communities refer to subsets of nodes that are more heavily connected within than they are to nodes outside the community. When connections represent Hi-C interactions, this would correspond to sections of the genome that are exhibit more frequent within-group contacts than outside-group contacts, as would be expected of TADs. The application of community detection approaches has the key advantage of being more principled and robust than attempting to identify such groups by visual inspection of heatmaps.

Network community detection is a rich field with a diverse collection of established algorithms (Girvan & Newman 2002, Porter, Onnela & Mucha 2009, Fortunato 2010, Fortunato & Hric 2016). The output of such an algorithm is a *partition*, a division of all of the nodes in the network into disjoint groups corresponding to the communities. Additionally, there are algorithms that may allow communities to overlap (soft partitioning) or nodes to not be present in any community; however, in this work, we will restrict our focus to hard community partitions.

Network community detection algorithms stand in contrast to another type of TAD detection algorithms present in the literature that explicitly looks for partitions where groups consist only of sets of nodes representing contiguous segments of chromosomes (Oluwadare & Cheng 2017). Under a framework in which TADs are defined specifically as such contiguous regions, this assumption may increase the ability of the algorithm to identify high-quality partitions or simplify the execution. However, results such as the split-nucleus simulations of (Hult et al. 2017), which showed that the actions that formed the nucleolus were not significantly affected if the nucleolus was split into two discontiguous segments, suggest the possibility of association beyond simple contiguous regions.

The literature has explored a number of algorithms for identifying community partitions in networks. The simplest such algorithm is to split the network into its connected components, ensuring that all connected nodes are in the same community and no inter-community edges exist. In the common case of weighted networks that have large numbers of low-weight connections such that the entire network is overall connected from an unweighted perspective, the connected components algorithm can be extended by applying a thresholding, keeping only edges above a certain weight, before extracting the connected components.

The modularity criterion was introduced in (Newman & Girvan 2004) to represent quality of partitions by counting the number of edges present within communities and comparing to that which would be expected under a null model. Recall the original loose definition of communities as groups of nodes that exhibit stronger connectivity within the group than to other nodes outside the group. By setting a null model, typically that of an Erdos-Renyi graph with the same average degree, the modularity value measures the degree to which this is true for a particular partition. Viewing the modularity as a value function, one can define a community detection algorithm that seeks to find a partition that maximizes the modularity (Newman & Girvan 2004, Duch & Arenas 2005, Brandes, Delling, Gaertler, Gorke, Hoefer, Nikoloski & Wagner 2008, Good, de Montjoye & Clauset 2010).

The modularity is a measure that correlates with quality of partition, but is not a perfect measure. Results have shown that even random networks with no structure can have high-modularity partitions (Guimera, Sales-Pardo & Amaral 2004), and so, as with any community detection algorithm, we must consider the question of verifying significance of the extracted community structure.

Further analysis of the original definition of modularity revealed that community detection by modularity optimization may not capture communities on a small scale relative to the size of the network, even in situations where the communities were qualitatively obvious (Fortunato & Barthelemy 2007), referring to this problem as the "resolution limit." In order to overcome this challenge, the modularity function was modified by the inclusion of a scale parameter, referred to as γ, with $\gamma = 1$ reducing to the original definition. γ scales the effect of the null-model contribution, and increasing γ has the effect of biasing modularity optimization towards smaller communities. In networks with hierarchical or multiscale clustering, different choices of γ can simultaneously correspond to significant community structure on different scales, and this approach has led to multiresolution analysis (Granell, Gómez & Arenas 2011, Traag, Krings & Van Dooren 2013) in which a range of values of the scale parameter are considered and the results are analyzed in order to find and extract community structure at the scale or scales at which it exists.

In addition to the question of spatial scale, the structure we are interested in also changes in time, and so we also wish to consider the case of networks that change over time. In this case, we can represent the data using a multi-layer networks. Multi-layer networks can be in general thought of as a sequence or collection (ordered or unordered) of networks, each of which represents a single "layer," along with some set of interlayer edges connecting nodes in different layers (De Domenico, Solé-Ribalta, Cozzo, Kivelä, Moreno, Porter, Gómez & Arenas 2013, Kivelä, Arenas, Barthelemy, Gleeson, Moreno & Porter 2014). Our interest in this work is specifically in representing networks that change in time, in order to extend these network tools to dynamic data. In particular, we will consider temporal networks, a subset of multi-layer networks in which layers correspond to subsequent time values in series, and inter-layer edges connect nodes with the same identity over subsequent timesteps, and we can apply similar principles to identify communities in temporal networks (Bassett, Porter, Wymbs, Grafton, Carlson & Mucha 2013, Han, Xu & Airoldi 2015).

The modularity criterion can be extended to multi-layer networks by adding a coupling parameter ω that increases the modularity when a node remains in the same community across multiple layers (Newman 2006, Mucha, Richardson, Macon, Porter & Onnela 2010). Note that because inter-layer edges are different from intra-layer edges, the null model is not compatible with "flattening" the multi-layer network into a single layer and then applying the single-layer modularity criterion, necessitating this additional step. However, with this modified form, the algorithm can be applied with little additional practical consideration except for the inclusion of the new temporal coupling parameter ω. Similar to single layer modularity, numerous algorithms exist for identifying partitions that optimize the multi-layer modularity, such as the GenLouvain algorithm (Jutla, Jeub & Mucha 2011-2019), or the multi-layer modularity belief propagation algorithm of (Weir, Emmons, Gibson, Taylor & Mucha 2017) that uses a statistical physics inspired approach to tie convergence of the algorithm to presence of significant community structure, and in this way the identification of communities is integrated with the analysis of significance.

The CHAMP algorithm of (Weir et al. 2017) provides a method to assess the significance of the partitions produced by modularity optimization by measuring robustness with respect to the parameters γ and ω. To apply this method, one first computes a large number of partitions for various values of γ and ω. As mentioned above regarding multiresolution analysis of networks, different values of γ correspond to looking for community structure at different scales, while different values of ω correspond for looking for structure that changes temporally at different rates, and so this collection of partitions may produce a significant diversity of results. Because γ and ω affect the value criterion, different partitions can simultaneously be optimal in different regions of (γ, ω) space, and we note that due to the randomness in the optimization algorithm, many partitions in the collection may not be optimal over the collection for any choice of (γ, ω). By harnessing the linearity of modularity with respect to these two parameters and applying convex hull methods, CHAMP extracts the subset of partitions that are optimal over some region of the parameter space, and identifies exactly what region each partition is optimal over. By observing the rate at which the partition changes with respect to variance in the parameters, one can identify which parameter values lead to robust partitions.

The data from our model is in the form of spatio-temporal position data for the beads that make up our polymers. In order to map this into a network problem, we will develop a mapping from a bead point-cloud into a network, using a similarity measure that monotonically decreases with the pairwise Euclidean distance. Combined with a thresholding step, this will produce a weighted network that represents the same information as the position data. Combining across timesteps, this will produce a multi-layer, temporal network as described above. We will then apply the GenLouvain multi-layer modularity algorithm to extract spatio-temporal community partitions. We note that this algorithm will natively use all timestep values simultaneously in computing communities, instead of separately computing communities at each timestep and then stitching together. We also use the CHAMP algorithm in order to determine what values of γ and ω lead to the most robust community partitions.

We will then extend our analysis of the emergent structure informed by the spatio-temporal community partitions provided by the multi-layer modularity method, demonstrating that the community partition captures the same notion of “interactions” that we originally defined spatially. We then show that the previously observed stable community structure in the fast crosslinking regime can be demonstrated mathematically as the existence of a small number of large communities that change very little from layer to layer in the network, and that the flexible crosslinking can be characterized in opposition as a very heterogeneous structure of many communities of many different sizes, including both larger communities that are more stable in time and small transient communities. The lack of a similar trend in the non-crosslinking regime agrees with our expectation that the community structure extracted does not represent meaningful clustering.

1.3 Energy Landscape Analysis of Stochastic Systems

The third and final goal is to understand how to mathematically represent the effect of stochastically switching pairwise crosslinking, as seen in the model in section 1.4, as an effective force. Because pairs of beads dynamically form and break these bonds, the system looks on a macroscale level as though it is in a state of thermal equilibrium, even though in reality the forces the beads are subject to are constantly changing, keeping the system out of a strict thermal equilibrium.

Further, stochastic switching in models of biological systems is not limited to chromosome models. Further examples include molecular motors (Welte 2004), crosslinked biopolymer networks (Cao & Forest 2019), and transient antibody crosslinking of antigens to mucus protein networks (Jensen, Wang, Lai, Forest & McKinley 2019, Schroeder, Newby, Schaefer, Subramani, Tubbs, Forest, Miao & Lai 2020). Modeling these active agents in combination with passive diffusion leads to mathematical models with two sources of noise – stochastically switching forces combined with stochastic Brownian motion.

Given the observations of Chapter 2 combined with a desire to better understand mathematical methods for stochastically switching systems, we would like to find an effective thermal equilibrium than can explain the mechanism behind the emergent clustering behavior we have observed. Recall that if the forces in a system, $v(x) = -\nabla U(x)$, are the gradient of a potential function $U(\vec{x})$, the dynamics governed by

$$dX = v(X)dt + \sqrt{2k_BT}dW \tag{1.1}$$

approach, in the long time limit, a Boltzmann thermal equilibrium distribution given by

$$p(x) \sim \exp\left(-\frac{U(x)}{k_BT}\right). \tag{1.2}$$

States x that minimize $U(x)$ are long-lived stable configurations at temperatures small enough relative to the energy barriers of $U(x)$ separating such states, defined as the gap ΔU between the energy at the minimum and the energy at the lowest saddle point on the region of attraction of the minimum. The mean transition time between the wells surrounding the energy-minimizing states can be computed asymptotically for vanishing temperature ($k_B T \to 0$): the time τ taken for the system to escape from a potential well under Brownian noise relates asymptotically to the energy barrier of the well following the Arrhenius equation given by

$$\mathbb{E}\left[\log \tau\right] \sim \frac{\Delta U}{k_B T}. \tag{1.3}$$

The most probable path the system traverses as it makes one such transition can also be found asymptotically for as the temperature approaches 0; it is a path that is everywhere parallel to the gradient of the energy landscape $U(x)$. The study of energy landscapes and large deviations allows for valuable insights to be made by viewing systems through the lens of statistical thermodynamics, and thus we seek an effective equilibrium that takes into account not only the thermal fluctuations but also the stochasticity induced by the switching forces in the above-mentioned DNA bead-spring model.

As mentioned above, in the asymptotic regime transitions will follow the the most-probable transition path, which in the gradient system case is perpendicular to the potential. Research seeking to understand transitions in gradient systems frequently applies numerical schemes for finding such transition paths between stable states. In order to understand such transitions, we must first note that in the small-noise regime, the system will spend most of its time near local minima of the potential function, and rarely undergo transitions between such minima. The most-probable transition path will start at the initial minimum, ascend up the energy landscape to a saddle point on the edge of the basin of attraction of the initial minimum, cross over, and then descend down into the new minimum.

In order to numerically represent these transitions, we consider collections of "images," each of which represents one particular system state x. In this way, a transition path can be discretized as a sequence of such images lying along the path.

A variety of methods exist for numerically characterizing transitions (Olsen, Kroes, Henkelman, Arnaldsson & Jónsson 2004), either by simply finding the saddle point or a series of images representing the full transition path. Earlier approaches include the dimer method (Henkelman & Jónsson 1999), in which a two images are considered together as a "dimer," climbing up towards a saddle point. An extension of this is the elastic band method (Henkelman, Uberuaga & Jónsson 2000), in which a series of images starts from the minimum, with the final image climbing up to the saddle point as the dimer did. These images are connected using a virtual spring force, ensuring they stay together. However, this additional force can distort the path the images describe. As an alternative, the string method (Weinan, Ren & Vanden-Eijnden 2002) replaces the elastic bands with re-interpolation of the images to ensure constant arc length between images along the path,

In order to fully map out the transition between two states, the string method (Weinan, Ren & Vanden-Eijnden 2002) stores a series of images that interpolate a path from the beginning state to the end state, interpolated to ensure uniform spacing along the path. Because a most-probable transition path is parallel to the direction of the gradient everywhere, the string method uses an iterative descent method in which all images are moved a small distance in the direction of the gradient, and then re-interpolated to satisfy the equal arc-length criterion. Repeating this process until convergence produces a discretized most-probable transition path. A modified version of the string method (Ren & Vanden-Eijnden 2013) exists in which the final image of the string has the component of its descent direction (given by the gradient) parallel to the string reversed, so that it descends in the directions perpendicular to the string, but ascends in the direction parallel to the string. This modification means that the final image now has a stable fixed point at saddle points, and so can "climb" out of a basin of attraction to a nearby saddle point. Another extension of the string method is the Geometric Minimum Action Method (GMAM) (Vanden-Eijnden & Heymann 2008, Heymann & Vanden-Eijnden 2008) which adds a preconditioning step that may improve performance.

In addition to computing the transition path, one can also compute the transition energy barrier as the difference in the potential between the minimum and the saddle point. Combined with the existence of a most-probable transition path connecting the minimum and saddle that is monotonic in the value of the potential along its path, which demonstrates the saddle point to be on the boundary of the basin of attraction of that minimum, we know that the system will exhibit a transition from that minimum through that saddle, at a rate asymptotically given by Arrhenius' law eq. (1.3). Finally, in the case that there are multiple such saddle points on the basin of attraction of the minimum in question, we know that the overall transition rate will be asymptotically given by the lowest such energy barrier. This allows us to characterize the stability of a minimum in terms of the depth of its associated energy well.

We would like to be able to apply these methods for finding transition paths and energy barriers to systems with stochastically switching crosslinking forces such as in our model section 1.4. However, as mentioned above, such a potential $U(x)$ does not exist – this motivates us to seek an "effective potential" that can stand in its place. A first thought to find an effective potential function $U(x)$ is to simply time-average the switching in the force while holding x constant, thereby removing the randomness in the switching. Intuitively, this can be thought of as assuming that the timescale of the switching is significantly faster than the timescale on which x, allowing us to do iterative averaging.

However, this assumption does not hold in general for our systems of interest, including the system from section 1.4, for values of μ that we are interested in. In particular, this approach significantly overestimates the effective strength of a strong force when the switching is not so fast. This can be intuitively understood by considering a thought experiment of an infinitely high potential barrier that is only sometimes on, as we can see that no matter the manner of the switching, the barrier remains infinitely high after averaging. Yet, as the particle could cross the barrier due to diffusion while it is off, transmission is clearly possible, contrary to the expectation of the naive time-averaging. While this example is extreme, the general principle that such naive averaging overestimates the effective height of the barrier can be observed.

In order to understand the emergent clustering behavior we are studying, especially in the intermediate regime, we thus cannot simple use a naive averaging approach which takes the noise from the stochastic switching to zero first, and then considers the effects of thermal noise. Rather, we seek a distinguished limit that takes the switching timescale to zero simultaneously with the thermal noise. To simultaneously consider both sources of randomness while leveraging the power of the energy landscape framework, we compute in Chapter 4 a quasipotential $W(x)$ whose gradient represents an average force that generalizes the asymptotic properties of the potential function to non-gradient systems (Cameron 2012, Bressloff & Newby 2014). We build off the work in (Newby 2014) that used a WKB approximation to construct a quasipotential for the Morris-Lecar equation, an ODE whose evolution depends on a stochastically-changing number of open ion gates. We show how this method can be used to derive the form of the Hamilton-Jacobi equation for a system with both a switching state modeled by a general continuous-time Markov chain and diffusive noise from Brownian motion.

We demonstrate the emergence of various numerical challenges in the process not previously reported in the literature, and develop new algorithms to overcome them. Applying our method to an idealized model of three beads moving in two dimensions, which we show to replicate a similar "flexible clustering" behavior as we identify in Chapter 2, we show the ability of our method to identify most-probable transition paths along with the associated barrier, which we confirm makes asymptotic predictions of transition times consistent with Monte Carlo simulations.

1.4 Previous Work and Model

The work in this dissertation builds on experimental biology and computational modeling research performed by the Forest/Bloom group at UNC (Vasquez & Bloom 2014, Vasquez et al. 2016, Hult et al. 2017) as described in section 1.1. We will describe here the specific form of the model from (Hult et al. 2017) which is used in the research in Chapter 2. Note that in Chapter 3, we will describe a series of modifications to create an idealized model to study averaging of the switching forces in more detail, and this idealized model is used in Chapter 4. We employ the identical model and code in (Hult et al. 2017), resolving the entire yeast genome into 2803 total beads, each of which represents approximately 5k base pairs (bp). As such, each bead is interpreted as a chromosome domain or a tension blob, as explained in the polymer physics literature. The beads are arranged on 32 chromosome arms having lengths that reflect their experimentally identified lengths. Each arm is tethered at both ends to the nuclear wall: all emanating from the centromere at one end, with the other end tethered to one of six telomeres. Along each arm, there are entropic, nonlinear springs (the wormlike chain model is used here) that connect neighboring beads. The beads also experience Brownian, entropic, repulsive and hydrodynamic drag forces, and are physically confined to the nucleus. We simulate approximately 20 minutes of G1 during interphase, and each simulation is initialized with 32 chromosome arms tethered at the centromere and one of six telomeres on the nuclear wall, otherwise randomly located within the nucleus (idealized as a spherical domain).

1.4.1 Model

This model is a representation of the full interphase yeast genome, consisting of 16 chromosomes each with two arms, discretized as a polymer bead-spring model with 2803 beads. The centromeres of all arms are tethered to the nuclear wall at a single common location, and the telomeres of each arm are each tethered to one of six telomere sites.

Each bead, corresponding to a tension blob of approximately 5kbp of DNA, is represented as a point in three-dimensional space, and is affected by the following forces:

1. Spring force
2. Excluded Volume
3. Thermodynamic Noise

4. Nuclear Boundary force

5. Dynamic Crosslinking

We begin by making an assumption that the beads are moving in a highly viscous environment, and therefore are always moving at their terminal velocity in the fluid. The relationship between force F and terminal velocity V is given by

$$V = \frac{1}{\rho} F \tag{1.4}$$

where ρ is the effective drag coefficient, $\rho = 2.5 \times 10^{-3}$ pN s/nm. **For the rest of this work, "forces" will be described in terms of the terminal velocity that arises from this overdamped approximation.**

This leads to the equation of motion that is simulated to produce the bead dynamics:

$$\rho \dot{X}_i = F^C_{i+1,i} - F^C_{i,i-1} + \sum_{j \neq i} F^{EV}_{ij} + \sum_{j \neq i} F^D_{ij} I(\text{i and j are crosslinked}) + \sqrt{k_B T} dW \tag{1.5}$$

plus the boundary enforcement which is handled separately.

1.4.2 Polymer spring force

The force applied by bead i on bead $j = i+1$ (or vice versa) due to the polymer chain spring force follows a wormlike spring formula. If $R_{ij} = |X_i - X_j|$ is the distance between beads i and j, and $R_0 = 1700$ nm is the extensibility of the wormlike spring, the magnitude of the spring force is given by:

$$F^C_{ij} = 0.2176 \left(-1 + \frac{1}{\left(1 - \frac{R_{ij}}{R_0}\right)^2} + \frac{4R_{ij}}{R_0} \right) \text{ nm/s} \tag{1.6}$$

The total bead-chain force on bead i is given by the sum of the force towards bead $i - 1$ and the force towards bead $i + 1$, whose magnitudes follow eq. (1.6).

1.4.3 Excluded Volume

The excluded volume force repelling any pair of beads is given by an isotropic Gaussian potential function. The magnitude of the corresponding repulsive force between any pair of beads i and j is given as follows:

$$F_{ij}^{\mathrm{EV}} = cR_{ij} \exp\left(-aR_{ij}^2\right) \text{ nm/s} \tag{1.7}$$

using the parameters

$$c = 0.03322 \text{ nm}^{-1} \qquad a = 3.26797 \times 10^{-5} \text{ nm}^{-2}$$

1.4.4 Thermodynamic noise

The thermodynamic noise is parameterized by the temperature parameter $k_B T = 4.11$ and the effective viscosity $\rho = 2.5 \times 10^{-3}$. Based on these parameters, a Brownian motion term is derived, consisting of a scale term times unit Gaussians. Specifically, the noise term for the position adjustment in a timestep of length Δt is given by:

$$\Delta B = 57.2750\sqrt{\Delta t} \text{ nm s}^{-1/2} \tag{1.8}$$

1.4.5 Nuclear Boundary

The beads are confined within the nucleus, which is modeled as a sphere centered at the origin with radius 1000. If, after a bead position update, any beads end up outside of this sphere, these beads are each immediately moved to the nearest point within (on the boundary of) the sphere. In other words, if $\vec{R}_i > 1000$, then

$$\vec{R}_i \leftarrow \frac{1000}{|\vec{R}_i|}\vec{R}_i \tag{1.9}$$

1.4.6 Dynamic Crosslinking

We model the effect of SMC proteins by transiently crosslinking pairs of non-neighboring beads in the nucleolus, represented by a contiguous chain of 361 beads on chromosome XII. (We note in passing that (Hult et al. 2017) also studied, experimentally and computationally, when the nucleolus is split onto separate chromosome arms, showing this to have a negligible affect on the clustering and interaction results.) A transient crosslink is modeled by a pair of stochastic events — a binding and unbinding of two non-adjacent nucleolar beads — and the timescale of these events is tuned by a single parameter μ (measured in seconds).

Below is described the implementation of the dynamic crosslinking, split up into two sections: determining of which beads are bound when, and the special dynamics that are applied to bound beads, taking the form of an additional spring force.

Bonds are formed between pairs of active beads within a minimum distance of $90nm$. In the case of multiple conflicting possibilities, bonds are formed greedily starting with the closest pair. Once a bond forms, it remains until one of the two bound beads switches to the inactive state. Upon entering the active or inactive state, the bead remains in that state for a duration drawn from the respective waiting time distribution as given in eq. (1.10). Note that both the active and inactive distributions are defined by the single parameter μ (measured in seconds), corresponding to the mean active time. The beads spend $1/9$ as much time in the inactive state, and standard deviations are one fifth the mean in both cases. Due to this coupling, even as the parameter μ is varied, these ratios remain the same.

$$t_{\text{on}} \sim \mathcal{N}\left(\mu, \frac{\mu}{5}\right) \tag{1.10}$$

$$t_{\text{off}} \sim \mathcal{N}\left(\frac{\mu}{9}, \frac{\mu}{45}\right) \tag{1.11}$$

While a pair of beads is bound, they are affected by a further wormlike spring force of the form

$$F_{ij}^{D} = 10.88\left(-1 + \frac{1}{\left(1 - \frac{R_{ij}}{R_0}\right)^2} + \frac{4R_{ij}}{R_0}\right) \text{ nm/s} \tag{1.12}$$

Hence, crosslinks are established and broken stochastically in the nucleolus, and the single parameter μ dictates the kinetic timescales for crosslinking.

1.4.6.1 Dynamic Crosslinking Force

The dynamic crosslinking springs use a similar wormlike spring model as the bead-chain springs, albeit with a higher spring force coefficient and a shorter extensibility. This means that the magnitude of the spring force (that is applied to both bound beads, respectively towards the other) is given by:

$$F_{ij} = 10.8794\left(-1 + \frac{1}{(1-\tilde{r})^2} + 4\tilde{r}\right) \text{ nm/s} \tag{1.13}$$

where

$$\tilde{r} = \min\left(\frac{R_{ij}}{45 \text{ nm}}, 0.999\right)$$

Furthermore, the magnitude of the force is truncated such that it would not cause the distance between two beads to be decreased by more than half in a single timestep. Since this force is applied to both beads, this means that the magnitude of the force is truncated at a value of $\frac{R_{ij}}{4}$.

1.4.7 Bead Position Update

As the viscous approximation causes the system to be first-order, the code updates bead positions using a straightforward Euler-Maruyama update with a timestep of $\Delta t = 0.001$ s. Specifically, the deterministic drift terms (bead-chain springs, dynamic crosslinking springs, and excluded volume force) are multiplied by Δt and then added to the bead positions. The noise term is already scaled with Δt, and so is directly added to the positions in the form shown above. Finally, the special boundary enforcement update is applied at the end of the update step.

Simulations are performed for a total simulation time of $3700s$, and the first $2400s$ are discarded to allow the system to reach thermal equilibrium, as was done in (Hult et al. 2017). This leaves a total observation time of $1300s$. This number also roughly corresponds to the amount of time a yeast cell spends in the interphase portion of the cell cycle, and is thereby a reasonable representation of the sorts of changes in spatiotemporal structure that may be observed.

CHAPTER 2: EMERGENT CLUSTERING IN CHROMOSOME MODEL[1]

2.1 Introduction

As discussed in section 1.1, previous work in (Hult et al. 2017) investigated a model of the full interphase yeast genome with dynamic crosslinking in the nucleolus, and observed emergent clustering structure in the nucleolus depending on the dynamic timescale of the crosslinking μ.

(Hult et al. 2017) simulated this model for thee values of the parameter μ: $0.09, 0.9, 90$.For a small value of μ, $\mu = 0.09$s, the nucleolar beads condensed into large, stable clusters containing approximately 10-20 beads. However, for larger values of μ, these clusters are not observed. Based on these results, we begin a further investigation of this model, with two primary goals:

1. Investigate the relationship between μ and emergence of clustering on a finer scale

2. Develop and apply automated tools for identifying spatio-temporal clustering in this model

Our first goal is to consider a wider range of values of the dynamic timescale parameter μ than were originally used in (Hult et al. 2017). Instead of the three values $0.09, 0.9, 90$, we will use twenty points over this same interval. Due to the apparent similarity of the results from 0.9 and 90, we chose to place the first ten points spaced logarithmically over the range $[0.09, 0.9]$ and the remaining ten over the range $[0.9, 90]$.

Other than the variation of μ, parameters were unchanged from (Hult et al. 2017). We simulated a total of 3700 seconds of data, recording only the final 1300 seconds so as to allow the system time to reach thermal equilibrium prior to data recording. We performed such a simulation for each of the twenty values of μ being considered.

Based on the results of these finer-resolution simulations, we will observe that in between the fast value of $\mu = 0.09$ and the middle value of $\mu = 0.9$, there exists a regime in which clusters are clearly visible, but as opposed to the large, stable clusters observed at $\mu = 0.09$, the clusters contain fewer beads, and most notably,

[1] This chapter is adapted from work previously published in PLOS Computational Biology under the CC-BY license (`https://creativecommons.org/licenses/by/4.0/`). The original citation is: Walker, Benjamin, Dane Taylor, Josh Lawrimore, Caitlin Hult, David Adalsteinsson, Kerry Bloom & M Gregory Forest. 2019. “Transient crosslinking kinetics optimize gene cluster interactions.” *PLoS Computational Biology* 15(8):e1007124.

exhibit significantly higher frequencies of beads entering and exiting clusters. Based on this observation, we will propose the existence of a new regime, leading to a total of three:

1. Rigid clustering (fast crosslinking $\mu < 0.1$)
2. Flexible clustering (intermediate crosslinking $\mu \in [0.2, 1]$)
3. No clustering (slow crosslinking $\mu > 1.6$)

Note that the intervals do not touch as they are designed to represent areas where we are confident this structure occurs, leaving space in the middle for a "transition".

Our desire to understand the spatio-temporal structure on a deeper level than just by inspection also motivates us to create metrics, scalar values that we can measure (as a statistical average) at each value of μ, and then look at to understand how the behavior changes dependent on μ. Based on the notion of "interactions" inspired by Hi-C data , we propose four interaction metrics.

(A) The *interaction fraction* indicates the fraction of possible unique bead pairs that interact at least once during an interphase simulation.

(B) The *mean interaction number* indicates the number of simultaneous interactions (i.e., beads within distance d^*) for a nucleolus bead, averaged across time and across beads.

(C) The *mean waiting time* indicates for any two beads, selected at random, the average time that passes between their i-th and $(i+1)$-th interactions.

(D) The *mean interaction duration* indicates the amount of time beads enter and reside within the interaction distance.

These metrics are described in additional detail in section 2.3.4. Note that we leave open the definition of "interaction" - we will consider both a two-point distance based definition, and later will also apply our community detection algorithm to produce a higher-order spatio-temporal measure of interaction, and will present the values of the metrics both ways.

The second goal we pursue in this chapter is automated detection of spatio-temporal clusters. Given that (Hult et al. 2017) observed structure in the form of clusters, it is natural to seek further characterization of this structure, but detailed investigation requires understanding of what beads are in what clusters. Furthermore, the beads move with each timestep, and can enter and exit clusters, which means that this computation of bead membership in clusters must also understand the spatio-temporal structure of the data.

We will map the spatio-temporal data into a temporal network, allowing us to apply network community detection methods, namely the Louvain multi-layer modularity optimization algorithm. We extract spatio-temporal community partitions for each of the three identified regimes and analyze the resulting communities, looking at how cluster stability and lifetime vary both with the crosslinking timescale μ and with the size of the cluster for a particular value of μ.

2.2 Methods

The model used for simulations of the yeast genome in this chapter is the same model as previously studied in (Hult et al. 2017). Details of this model are given in section 1.4

2.2.1 Pairwise-Distance Maps for High-Throughput Chromosome Conformation Capture

Each simulation of the Rouse-like polymer model yields time-series data $\{\mathbf{x}^i(t)\} \in \mathbb{R}^3$ that defines the 3D location of each bead $i \in \{1, \dots, N\}$ at each discrete time step $t = 0, 1, \dots$. We establish a connection between our simulated data and the state-of-the-art in chromosome imaging—namely, high-throughput conformation capture (Hi-C) — by constructing and analyzing *pairwise-distance maps*. Hi-C "images" the conformation of chromosomes using a combination of proximity-based ligation and massively parallel sequencing, which yields a map that is correlated with the pairwise distances between gene segments. While the actual pairwise distances between gene segments cannot be directly measured, Hi-C implements spatially constrained ligation followed by a locus-specific polymerase chain reaction to obtain *pairwise count maps* that are correlated with spatial proximity: the count between two gene segments monotonically decreases as the physical 3-dimensional distance between them increases.

To provide an analogue to Hi-C imaging, we construct pairwise-distance maps for our simulated data $\{\mathbf{x}^i(t)\}$. Let

$$F : \{\mathbf{x}^i\}_{i=1}^{i=N} \mapsto \mathbb{R}^{N\times N} \tag{2.1}$$

define a *map* (used here in the mathematical sense) from a set of N points $\{\mathbf{x}^i\} \in \mathbb{R}^3$ to a matrix such that each entry (i, j) in the matrix gives the distance between point i and point j. Whereas Hi-C imaging aims to study the positioning of chromosomes using noisy measurements that are inversely correlated with pairwise distances, for our simulations we have access to the complete information about the chromosome positioning. We therefore define and study several variations for pairwise distance maps, which will allow us to also study artifacts that can arise under different preprocessing techniques, such as averaging the time series data across time windows and/or averaging across multiple simulations with different initial conditions. We define the following pairwise distance maps:

- An *instantaneous* pairwise distance map $X(t) = F(\{\mathbf{x}^i(t)\})$ encodes pairwise distances between beads (i.e., chromosome domains) at a particular timestep t.
- A *time-averaged* pairwise distance map $Y(\tau) = \frac{1}{|\tau|} \sum_{t\in\tau} X(t)$ encodes the pairwise distance between beads averaged across a set of timesteps τ.
- A *population-averaged* pairwise distance map $Z(t) = \langle X(t)\rangle_p$ encodes the pairwise distance at timestep t between beads, which are averaged across several simulations that have different initial conditions (which are chosen uniformly at random).

These pairwise distance maps represent the data that is sought after, but cannot be directly measured, by Hi-C imaging. Moreover, by defining several distance maps we are able to study "averaging" artifacts that can arise due to various limitations of Hi-C imaging. For example, Hi-C imaging obtains measurements that are typically averaged across a large heterogeneous distribution of cells that are subjected to nonidentical conditions and exist at nonidentical states in their cell cycles.

2.2.2 Microscope Image Acquisition, Processing and Analysis

2.2.2.1 Yeast Strains for Experiment

The budding yeast strains used in this study were obtained by transforming the yeast strain EMS219 (Mat alpha, his5 leu2-3,212 ura3-50 CAN1 asp5 gal2 (form I1 rDNA::leu2 URA3+)) with CDC14-GFP:KAN^{R}, to label the nucleolus, and SPC29-RFP:HYG^{R}, to label the spindle pole body, to generate the yeast strain DCY1021.1. DCY1021.1 was transformed to knock out FOB1 and HMO1 to generates DCY1055.1 and DCY1056.1 respectively.

2.2.2.2 Image Acquisition and Baseline Processing

Fluorescent image stacks of unbudded yeast cells were acquired using a Eclipse Ti wide-field inverted microscope (Nikon) with a 100× Apo TIRF 1.49 NA objective (Nikon) and Clara charge-coupled device camera (Andor) using Nikon NIS Elements imaging software (Nikon). Each image stack contained 7 Z-planes with 200 nm step-size.

Image stacks of experimental images were cropped to 7 Z-plane image stacks of single cells using ImageJ and saved as TIFF files. The cropped Z-stacks were read into MATLAB 2018b (MathWorks), converted into maximum intensity projections, and the projections of hmo1Δ and fob1Δ were cropped to 55×55 pixels, to match the dimensions of WT projections, using MATLAB function padarray with `replicate` option specified to extend outer edge of pixel values to ensure the center of all cropped images was the brightest pixel. The intensity values all projections were normalized by subtracting all intensity values by the minimum value and then dividing the resulting values by the maximum intensity value after subtraction. The normalized intensity values were stored with double point precision, preventing any loss in dynamic range.

2.2.2.3 Image Analysis

The areas of nucleolar signals were determined by setting all values below threshold, calculated using `multithresh` function, to NaN and then summing number of values that were not NaNs. That pixel count was converted to μm^2 by multiplying the sums by 0.0648^2, the area of each pixel in μm^2.

To calculate the standard deviation of the intensities of the nucleolar signal, the non-NaN values remaining after thresholding were re-normalized using the same method described above, and the standard deviation of those values was measured.

To count clusters within the nucleolar signal, the normalized images were deconvolved with 5×5 Gaussian structural element, using deconvblind function, and underwent two rounds of background subtraction by setting all intensity values below threshold value, calculated using default `multithresh` function, to NaN and then all NaN values to 0. Clusters were identified using the `imregionalmax` function and counted using `bwconncomp` function.

The simulated images generated from our simulations were analyzed as described above with the additional step of measuring the standard deviation of the each simulated maximum intensity projection. All

WT images were analyzed using the script wtExpIm.m. All hmo1Δ and fob1Δ images were analyzed using the script cropExpIm.m. All simulated images were analyzed using the script clusterCountLoop.m.

All MATLAB scripts have been made available at (Lawrimore N.d.).

2.2.3 Cluster Identification via Network Community Detection

2.2.3.1 Gene-Interaction Networks from Pairwise-Distance Data

Given a pairwise distance map $X \in \mathbb{R}^{N\times N}$ in which each entry X_{ij} gives the (possibly averaged) Euclidean distance between beads i and j as described in section 2.2.1, we construct a network model in which there are weighted edges (i.e., interactions) only between beads that are in close proximity to each other and for which each edge weight $A_{ij} \geq 0$ decreases monotonically with distance X_{ij}. We propose a model with two parameters, d^* and s, which represent a distance threshold and a decay rate, respectively. In particular, we define a network adjacency matrix A having entries

$$A_{ij} = \begin{cases} e^{-sX_{ij}}, & X_{ij} < d^* \\ 0, & X_{ij} \geq d^*. \end{cases} \tag{2.2}$$

Note that there exists an undirected edge between i and j (i.e., $A_{ij} > 0$) only when $X_{ij} < d^*$, and s controls the rate in which the edge weight A_{ij} decreases with increasing distance X_{ij}. Because the edge weights exponentially decrease with distance, the community detection algorithms we study are insensitive to the choice for d^*, provided that d^* is sufficiently large so that the network is connected. Our choice $d^* = 325$ in section 2.3.7 ensures there is an edge between all beads in the same cluster and was found to yield qualitatively similar results for other choices of d^*.

Equation (2.2) defines a map between a distance matrix and an affinity matrix that encodes a network. Note that for any such adjacency matrix A, we can equivalently define the network using the graph-theoretic formulation $G(\mathcal{V}, \mathcal{E})$. Here $\mathcal{V} = \{1, \dots, N\}$ denotes the set of nodes (i.e., the beads in the chromosome model) and $\mathcal{E} = \{(i, j, A_{ij}) : A_{ij} > 0\}$ denotes the set of weighted edges (i.e., a set encoding which beads are interacting as well as their interaction strengths).

In section 2.2.1 we defined several versions of pairwise distance maps—instantaneous, time-averaged, and population-averaged maps—and a network model can be constructed for any of these maps:

- An *instantaneous interaction network* refers to a network associated with an instantaneous pairwise distance map $X(t)$.
- A *time-averaged interaction network* refers to a network associated with a time-averaged pairwise distance map $Y(\tau)$. We point out that due to the nonlinearity of eq. (2.2), a network associated with a time-averaged distance map can in general differ from a temporal average of instantaneous interaction networks, averaged across the same time interval.
- A *population-averaged interaction network* refers to a network associated with a population-averaged distance map $\langle Z(t)\rangle_p$.

In addition to the above network models, we are particularly interested in constructing and studying *temporal interaction networks*, which we define as a sequence of time-averaged interaction networks encoded by a sequence of adjacency matrices $\{A_{ij}^s\}$. In particular, given a sequence of timesteps $s \in \{1, 2, \dots, T\}$, we partition time into a sequence of time windows $\tau_s = \{(s-1)\Delta+1, \dots, s\Delta\}$ for $s = 1, 2, \dots$ of width Δ. We then define a sequence of time-averaged networks $\{G(s)\}$ for $s = 1, 2, \dots$ associated with these distance maps, which are time-averaged across the non-overlapping time windows $\{\tau_s\}$. The result is a sequence of adjacency matrices so that each entry A_{ij}^s indicates the absence ($A_{ij}^s = 0$) or presence ($A_{ij}^s > 0$) of an edge between i and j during time window τ_s.

In practice, we choose the time window width Δ to be similar to—but slightly larger than— μ. Matching the time-scales of μ and Δ allows the temporal interaction networks to efficiently capture the dynamics of interactions. Specifically, if Δ is too short then the temporal network will be identical across many time steps, which is not an efficient use of computer memory. Moreover, if Δ is too large, then the temporal network data will be too coarse to identify interaction dynamics occurring at a faster time scale. We chose $\Delta = 10$ to aggregate the time-varying bead-location data (which was saved every 0.1 second) into 1-second intervals. We studied 1000 such time windows to produce fig. 2.8.

2.2.3.2 Spatiotemporal Gene Clusters Revealed by Community Detection in Temporal Networks

We analyze spatiotemporal clustering of chromosomes using community detection methodology for temporal interaction networks, particularly an approach based on multi-layer-modularity optimization. Given a sequence of adjacency matrices $\{A^s\}$ for $s \in \{1, 2, \dots, T\}$, we study the *multi-layer modularity measure*

(Mucha et al. 2010)

$$Q = \frac{1}{2\mu} \sum_{i,j,s,r} \left[\left(A_{ij}^s - \gamma \frac{k_i^s k_j^s}{2m_s} \right) \delta(s,r) + \omega \delta(i,j) C_{sr} \right] \delta(c_{is}, c_{jr}), \tag{2.3}$$

where A_{ij}^s denotes an entry in the adjacency matrix for network layer s (i.e., that associated with time window τ_s), γ is again a tunable "resolution parameter," $k_i^s = \sum_j A_{ij}^s$ is the weighted node degree for node i in layer s, $2m_s = \sum_{ij} A_{ij}^s$ is twice the total number of undirected edges in layer s, $\delta(m,n)$ is again a Dirac delta function, $C_{sr} = \delta(s, r-1) + \delta(s, r+1)$ defines the coupling between consecutive (time) layers and $C_{sr} = 1$ if only if $r = s \pm 1$ (otherwise $C_{sr} = 0$), and $\{c_{is}\}$ are the integer indices that indicate the community for each node i in each layer s. If one wished to analyze just a single network (e.g., a time-averaged or population-averaged network), then one can simply set $C_{sr} = 0$ so that the second term in the square brackets is discarded.

Letting $i \in \{1, \dots, N\}$ enumerate the nodes and $s \in \{1, \dots, T\}$ enumerate the network layers, the goal is to assign a community label c_{is} to each node-layer pair (i, s) to maximize Q (Mucha et al. 2010). Here, $c_{is} = c$ indicates that node i is in community c during time window τ_s. There are many techniques to solve such an optimization problem, and we identify partitions that optimize Q using a variational approach commonly referred to as the Louvain algorithm (Blondel, Guillaume, Lambiotte & Lefebvre 2008, Jutla, Jeub & Mucha 2011-2019). To provide some intuition into this optimization problem, we briefly comment on how the different terms in Eq. (2.3) contribute to this optimization problem.

Consider the first term, $(A_{ij}^s - \gamma \frac{k_i^s k_j^s}{2m_s})$, which is a slight generalization of Newman's original definition of the modularity measure (Newman 2006) (which assumed $\gamma = 1$). The part $\frac{k_i^s k_j^s}{2m_s}$ is the expected probability of an edge between i and j in layer s according to the configuration null-model for networks (Newman 2006), and the effect on Q of this null-model comparison is scaled by γ. Thus, as whole, the first term is largest when there exists an edge between i and j in time window τ_s and when the expected probability of such an edge is smallest. Effectively, this term influences optimal partitions to give (i, s) and (j, s) the same community label if there is an edge between nodes i and j in time window τ_s. We next consider the second term, $\omega \delta(i,j) C_{sr}$, which is nonnegative since $\omega > 0$ and $C_{sr} \in \{0, 1\}$. Since $C_{sr} = 1$ only when τ_s and τ_r are consecutive time windows (i.e., $|s - r| = 1$), this term influences the community labels c_{is} and c_{ir} to be the same from one time window to the next.

2.2.4 Calibration with CHAMP

Note that multi-layer modularity involves two parameters γ and ω, which are "tuning knobs" to identify clusters in which their size and temporal coherence are appropriate. In practice, we explore a wide range of parameter values $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ and $\omega \in [\omega_{\min}, \omega_{\max}]$ to study the multiscale organization of clusters. This approach efficiently explores clustering phenomena at multiple spatial and temporal scales, identifying at which scales clustering is most prevalent and at which scales clustering is nonexistent. To identify appropriate values for γ and ω, we used the CHAMP algorithm (Weir et al. 2017, Weir, Gibson & Mucha 2017) (which utilizes fast algorithms that detect convex hulls in the (γ, ω) parameter space) and comparisons to other community-detection algorithms including the study of connected-components.

2.3 Results

2.3.1 Transient Crosslinking Timescale Influences Nucleolus Clustering

We first focus in the relatively short crosslink timescale regime, extending the simulations of (Hult et al. 2017) at discrete values $\mu = 0.09, 0.9, 90$. These will establish a basic understanding of how the kinetic timescale μ for crosslinking sensitively affects the organization of the nucleolus and the dynamics of the architecture. From our refined simulations across the above four decades, the essence of the story can be told with results for three selected values $\mu \in \{0.09, 0.19, 1.6\}$. In fig. 2.1(A–C), we present visualizations, i.e., "snapshots," of the beads' 3D positions during the simulations. The nucleolus on Chromosome XII is highlighted in blue and all remaining chromosome arms are colored gray. In fig. 2.1(D–F), we show only the nucleolar beads, which are colored according to the network community detection analyses that we describe in the following sections.

Based on fig. 2.1(A)–(F) and the videos, we identify three qualitative regimes for nucleolus clustering:

1. *rigid clustering* whereby strong, stable clusters arise, e.g., with $\mu = 0.09$.

2. *flexible clustering, or cluster plasticity* with slightly weaker clustering whereby communities of genes form and persist, however the clusters frequently interact, merge, and divide, swapping genes per interaction, e.g., with $\mu = 0.19$.

3. *non-clustering* with a lack of robust communities in which the beads act as lone units, pairs or triplets, e.g., with $\mu = 1.6$.

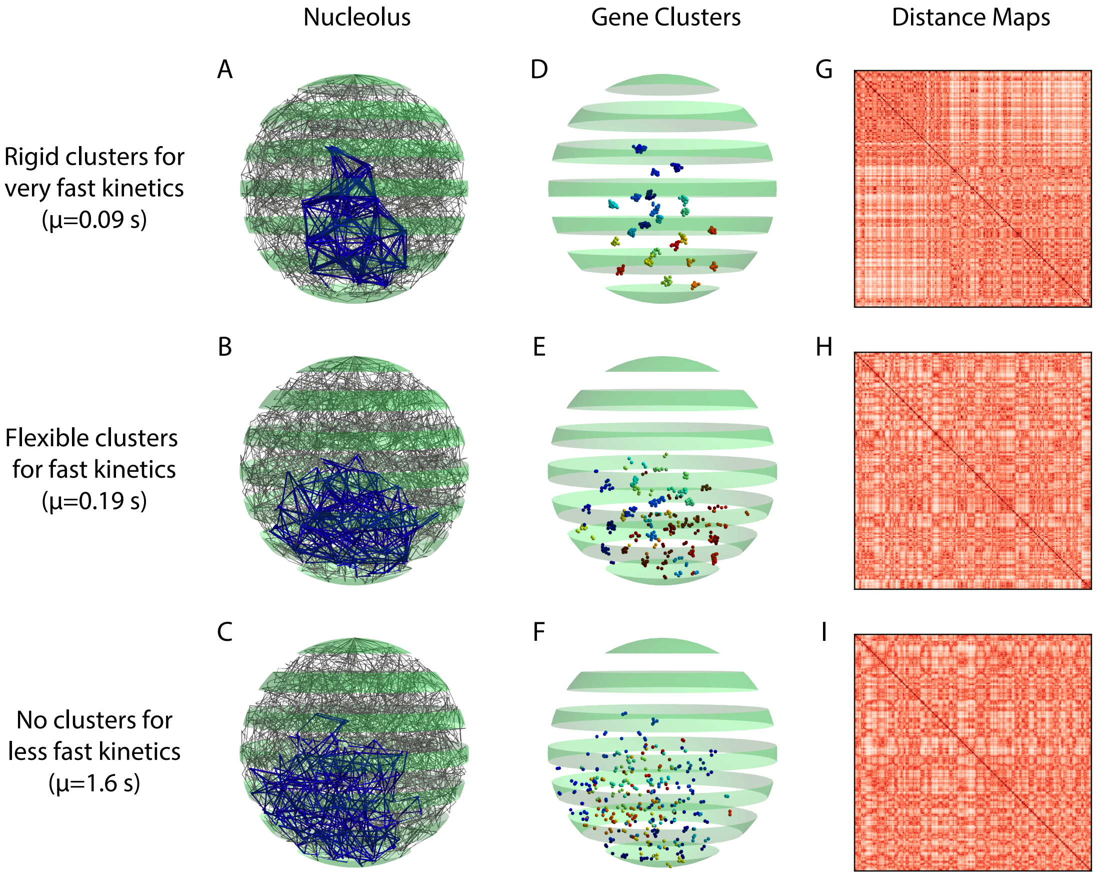

Figure 2.1: Interphase yeast genome (A-C: Full, D-I: Nucleolus) using the polymer bead-spring model. (see section 1.4) that implements transient SMC-protein-mediated crosslinking of 5kbp domains within the nucleolus on Chromosome XII. The top, middle, and bottom rows depict chromosome conformations for three values of the kinetic timescale parameter μ: (top, $\mu = 0.09$) induces rigid clusters; (center, $\mu = 0.19$) induces flexible clusters; (bottom, $\mu = 1.6$) induces no clusters. We identify the timescales with the intra-nucleolar clustering behavior they induce. (A)–(C) 3D "snapshots" of all 16 yeast chromosomes during interphase. Blue beads and edges highlight the nucleolus. (D)–(F) Visualization of nucleolar beads ($5kbp$ chromosome domains) that self-organize into clusters. The beads' positions are identical to those in (A)–(C) and their colors indicate their cluster labels, identified using modularity optimization as described in Methods Section: section 2.2.3.2. (G)–(I) Heatmaps of the pairwise distances between beads in the nucleolus from one snapshot of the 4D time series, which provide an analogue of Hi-C bead-bead proximity data (see section 2.2.1). Note that it is difficult to predict the absence/presence of clusters from heat maps. Note that it is difficult to predict the absence/presence of clusters from heat maps.

We will continue to use this terminology when referring to these three clustering regimes. We note that (Hult et al. 2017) discovered the two extreme regimes: robust clusters for $\mu = 0.09$, and the lack of clusters for $\mu = 90$. As they did not finely sample the decades of timescales in between, they did not discover that the transition from robust to no clustering is in fact non-monotone with respect to gene-gene interactions, nor that the transition is essentially complete already at $\mu = 1.6$, and that the most biologically interesting and relevant regime occurs at $\mu = 0.19$. Furthermore, without automated structure detection algorithms, they would not have been able to detect and dynamically track clusters of genes and their interactions that explain the peak in gene-gene interactions at $\mu = 0.19$. This transition behavior and the optimal properties that arise will be the focus of several sections to follow.

In fig. 2.1(G)–(I), we show heatmaps of the bead-bead distances associated with the bead positions of the snapshots in (D)–(F), identical to those in (A)–(C); construction of heatmaps is described in the section 2.2.1. Heatmaps are widely used in Hi-C to depict population averages of pairwise gene-gene proximity data (Eser, Chandler-Brown, Ay, Straight, Duan, Noble & Skotheim 2017, Schalbetter, Goloborodko, Fudenberg, Belton, Miles, Yu, Dekker, Mirny & Baxter 2017, Fudenberg, Imakaev, Lu, Goloborodko, Abdennur & Mirny 2016, Dekker, Marti-Renom & Mirny 2013, Imakaev, Fudenberg, McCord, Naumova, Goloborodko, Lajoie, Dekker & Mirny 2012, Lieberman-Aiden et al. 2009) and in simulated data from polymer bead-spring models, both from 3D snapshots and time averages (Verdaasdonk et al. 2013, Vasquez et al. 2016, Hult et al. 2017). Comparing the second and third columns of Fig. fig. 2.1, we note the difficulty (false negatives and false positives) in detecting the presence of structure and sub-organization in column 2 from visual examination of heatmaps in column 3.

As shown in (Hult et al. 2017), the time average of 4D simulated datasets, even in the strong clustering regime, wipes out the sub-structure of snapshots when averaging over the entire G1 phase. An alternative approach has been to use polymer modeling to generate chromosome conformations, and to select those conformations that best match Hi-C data, so-called restraint-based polymer modeling (Dekker et al. 2017). Simultaneously, there have been efforts to develop methodologies to identify gene clusters in a rigorous and automated way from Hi-C data (Rajapakse et al. 2010, Rajapakse & Groudine 2011, Rajapakse, Groudine & Mesbahi 2011). Our conclusion is that there is a need for a more reliable and objective method to study the clustering of chromosome domains in the nucleolus, especially spatio-temporal methods that take into account how bead positions and sub-organization change with time, weighing both spatial proximity and temporal coherence in the detection method. In the following sections, we present a scalable and automated technique to identify and track the dynamics of clusters. First, however, we will present new experiments that provide empirical evidence for clustering in the nucleolus.

2.3.2 Evidence of Nucleolus Clustering in Experimental and Simulated Microscopy Images

We conducted experiments to qualitatively compare image-based cluster analysis between our model and empirical measurements obtained from live cell microscopy and demonstrate the effect that SMC protein mutation can have on clustering in the nucleolus, extending the results previously reported in (Hult et al. 2017). Here, we study three yeast strains: wild-type (WT), fob1 and hmo1. Importantly, fob1Δ and hmo1Δ are mutations that lack key proteins reported to crosslink or loop segments of rDNA within the nucleolus. Fob1Δ is required for maintenance of the rDNA copy number and regulates the association of condensin with rDNA repeats (Johzuka & Horiuchi 2009, Johzuka, Terasawa, Ogawa, Ogawa & Horiuchi 2006). The replication fork barrier within the rDNA is a binding site for Fob1Δ that, together with several other components (Tof1, Csm1 and Lrs4), are responsible for the concentration of condensin within the nucleolus (Johzuka & Horiuchi 2009). Strain hmo1 is an abundant high mobility group protein that localizes to the nucleolus and has been proposed to share functions with UBF1, which is involved in rDNA transcriptional regulation within the nucleolus (Prieto & McStay 2007, Albert, Colleran, Leger-Silvestre, Berger, Dez, Normand, Perez-Fernandez, McStay & Gadal 2013). Fob1 and Hmo1 are non-essential genes and were deleted from the genome to allow us to study their effect on nucleolus morphology due to functional modifications of crosslinking. See section 2.2.2.1 for further details.

In fig. 2.2, we present images and analyses of nucleoli of these strains using fluorescent, live-cell microscopy. To visualize nucleoli, we fused Cdc14 protein phosphatase to green fluorescent protein (GFP) (Hult et al. 2017). Nucleolar protein fusions occupy a distinct region of the nucleus that is adjacent to the nuclear envelope and (typically) opposed to the spindle pole body. We describe the image acquisition and processing steps in section 2.2.2.2, and we highlight a few details here. Following image acquisition, we construct *maximum intensity projections* (MIP) centered on the nucleolus. See top row of fig. 2.2(A). Due to potential variation in CDC14-GFP protein copy number and nucleolar/rDNA size from cell to cell, we normalized the nucleolar CDC14-GFP signal after excluding all intensity values below an intensity threshold. To this end, we first selected a threshold using Otsu's method (Otsu 1979), which we implemented using the MATLAB function multithresh. One can interpret the threshold as a binary mask, as shown in the second row of fig. 2.2(A). After applying the mask, we normalized the nucleolar signal by subtracting all intensities by the minimum value and then dividing them by the new maximum intensity that is obtained after subtraction. The third row of fig. 2.2(A) depicts normalized images. Mutations of Hmo1 and Fob1 were found to alter the

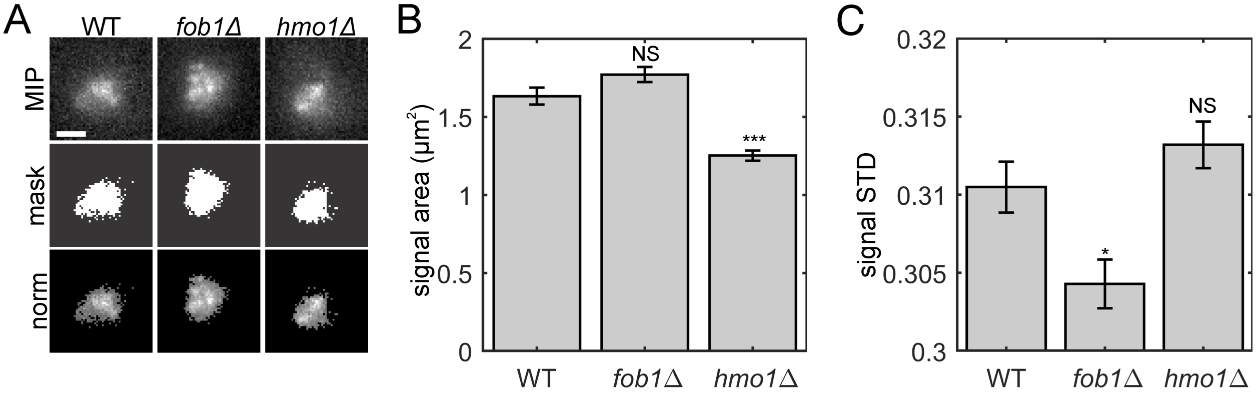

Figure 2.2: Experimental images of the nucleolus. Algorithmic thresholding of CDC14-GFP reveals alterations in nucleolar area and variance of signal intensity. (A) Top row depicts images of maximum intensity projection of WT, fob1Δ, and hmo1Δ yeast cells containing CDC14-GFP to label the nucleolus. Scale bar is 1 μm. The middle row are images of the binary mask generated by thresholding the projections using Otsu's method. The bottom row are images of the projections where the intensities of pixels below the threshold were set to zero. See section 2.2.2.2 for more detail. (B) Bar chart of the nucleolar signal are calculated by the area of the binary mask. (C) Bar chart of the standard deviation of normalized nucleolar signal. In panels (B) and (C), the non-significant changes are labeled 'NS' (see text), the bars represent an average value across n cells, and error bars indicate standard error.

area and signal intensity of nucleoli labeled with CDC14-GFP across a range of intensity thresholds, which we surmise is due to alterations in the architecture of, i.e., clustering within, the nucleolus.

In fig. 2.2(B)-(C), we provide results for an analysis of nucleolar morphology: (B) the area of nucleolar signal; and (C) the standard deviation of the normalized signal. This analysis was implemented using the numerical algorithms presented in (Hult et al. 2017), which we further describe in section 2.2.2.3.

As shown in fig. 2.2(B), null mutations of hmo1Δ significantly altered the area of the nucleolar signal, whereas null mutations of fob1Δ did not. This was assessed by a Student's two-tailed T-test, which yielded $p = 3 \times 10^{-8}$ for the former and $p = 0.07$ for the latter. As shown in fig. 2.2(C), the standard deviations of the normalized images were significantly lowered for the fob1Δ null mutation, but this did not occur for the hmo1Δ null mutation $p = 0.01$ versus $p = 0.2$). The non-significant changes are labeled 'NS' in the figure. The error bars indicate standard errors across n cells, where $n = 84$, 70 and 77 for the WT, fob1Δ and hmo1Δ strains, respectively.

We note that (Hult et al. 2017) also studied the area and variance of the nucleolus using experimental and simulated images. They found, for example, that the distribution of areas occupied by the nucleolus displays a lognormal distribution for WT cells in G1. Also, recall that we implemented thresholding based on Otsu's method; in contrast, (Hult et al. 2017) explored a range of threshold values and found qualitatively similar results to be consistent across a range of threshold values. They did not, however, explore the area and variance for simulated images for a wide range of μ, which is the focus of our next experiment.

To explore whether varying the kinetic timescale μ for our simulations yields similar changes as those arising under the fob1Δ and hmo1Δ mutations, we applied the microscope simulator of (Hult et al. 2017) to our 4D simulated data and analyzed the images using the same image analyses as described in fig. 2.2. First, we converted our 4D simulated data into a timelapse sequence with 22 time points, i.e., snapshots. Each nucleolus bead was convolved with a point spread function and a maximum intensity projection was created for each timepoint. We depict 11 such images in fig. 2.3(A). In panel (B), we plot the area of the nucleolar signal (computed using Otsu's threshold) versus μ. Note that the nucleolus area increases as μ increases. In panel (C), we plot the standard deviation of nucleolar signal versus μ, which has the opposite trend. In fig. 2.3(D), we plot the standard deviation of images obtained after a normalization step that is identical to that implemented for the experimental images (see discussion for fig. 2.2. Interestingly, the dependence on μ of the signal's standard deviation drastically changes depending on whether or not it is normalized. Given that normalization is required to control for cell-to-cell differences in CDC14-GFP and in nucleolar/rDNA size, we sought develop a metric to measure clustering in the CDC14-GFP signal that was independent of the absolute values of the intensities.

Our final experiment studies cluster formation in the nucleolus and compares clustering observed in the experimental and simulated microscopy images. We developed a cluster detection algorithm written with MATLAB (see section 2.2.2.3) and applied it to both the experimental and simulated images. We have made the code available at (Lawrimore N.d.). In fig. 2.4(A), we depict images of maximum intensity projections for WT, fob1Δ and hmo1Δ strains with (top row) and without (bottom row) visualizations of detected clusters, which are represented by green circles. In panel (B), we depict identical information as in panel (A) except we show simulated images for three values of μ. In fig. 2.4(C) and (D), we show the number of clusters for the experimental and simulated images, respectively. For the experimental images, we give results for WT, fob1Δ and hmo1Δ, whereas for the simulated images we present results for $\mu \in [.09, 90]$. We observe that the number of clusters was significantly decreased in the hmo1Δ null mutation, but not the fob1Δ null mutation ($p = 0.04$ for hmo1Δ versus $p = 0.3$ for fob1Δ). We also observe that increasing μ yielded a general trend in which there were fewer clusters. Taken together, these data suggest that gene clustering can directly impact the size and shape of the nucleolus. This underscores the need for robust and objective tools for identifying gene clusters.

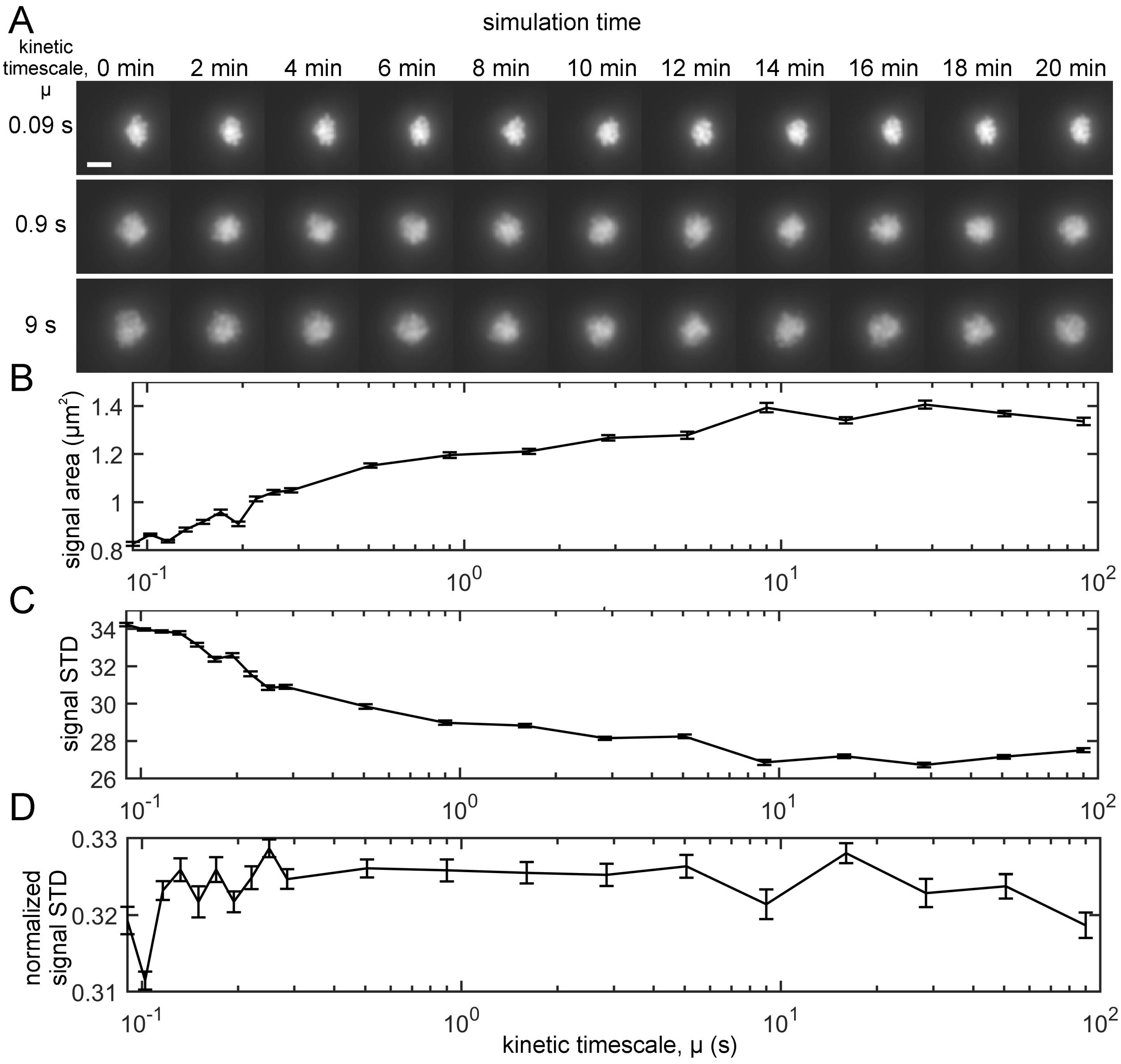

Figure 2.3: Simulated microscope images for polymer bead-spring model with transient crosslinking and varying kinetic timescale μ. Varying μ alters the area and variance of the intensity for the nucleolar signal, which models the affect of the fob1Δ and hmo1Δ mutations. (A) Timelapse montage of simulated microscope images for $\mu \in \{0.09, 0.9, 0\}$ (seconds). Scale bar is μm. (B) Area of nucleolus signal (μm^2) as a function of $\mu \in [10^{-1}, 10^2]$. Areas were calculated by measuring the area of the binary mask generated by applying Otsu's threshold to each image. (C) Standard deviation of the (non-normalized) nucleolus signal versus μ. (D) Standard deviation of the normalized nucleolus signals versus μ; these we normalized identically to the normalization of the experimental microscope images. In panels (B)-(D), error bars indicate standard errors observed using 22 time points for each value of μ.

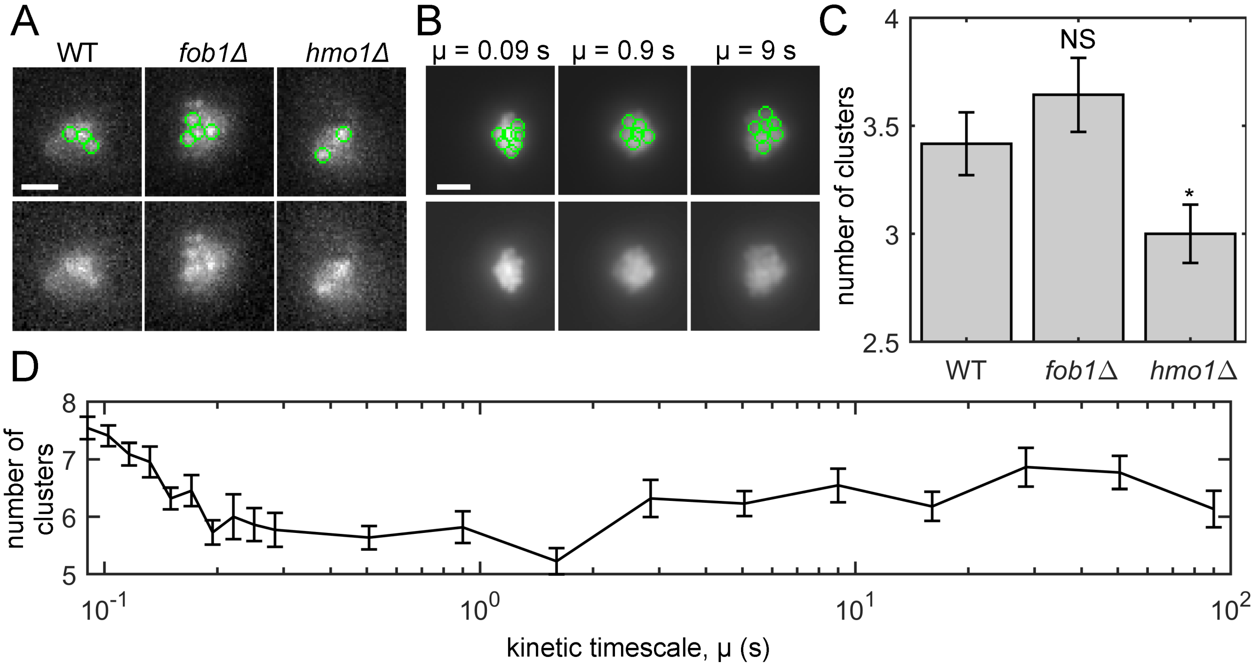

Figure 2.4: Comparison of experimental and simulated microscope images. Varying μ in the polymer bead-spring model captures the effect on nucleolus clustering incurred by the crosslink-altering mutations. (A) Maximum intensity projections of CDC14-GFP in WT, fob1Δ, and hmo1Δ cells with identified clusters marked by green circles. The lower row depicts the same images as the top row, except the circles are removed. (Images are the same as fig. 2.2(A).) (B) Same information as panel (A), except the results depict simulated images based on 4D simulation data for three values of μ. Scale bars in (A) and (B) are 1 μm. (C) Bar chart showing the number of clusters for each strain. The bars and error bars indicate the average and standard error across n cells, where $n = 84$, 70, and 77 for WT, fob1Δ, and hmo1Δ, respectively. Significance was assessed via a Student's two-tailed T-test: $p = 0.3$ for WT versus fob1Δ and $p = 0.04$ for WT versus hmo1Δ. (D) Average number of clusters for simulated images as a function of μ. The averages and standard errors (see error bars) were calculated using 22 time points for each μ.

2.3.3 Histograms of 2-Point Pairwise Distances between Nucleolar Beads

A simple and previously used method for analyzing distances between beads is to create a histogram of all bead-bead pairwise distances. As explored in (Hult et al. 2017), this two-point statistic can provide evidence of clustering and can be used to query simple properties such as whether or not the clusters change over time. In this section, we repeat this analysis on our data and extend it by showing what effects averaging over time and averaging over populations has on the results. We show that averaging one cell over time prevents observing the flexible clustering through pairwise distances, and averaging over populations prevents observing any sort of clustering. We provide further details on the computation of these distances in section 2.2.1.

In fig. 2.5(A), we plot the distribution of all pairwise distances $\{d_{ij}^{(t)}\}$ at a single time t for three kinetic timescales, given by the same values $\mu = 0.09, 0.19, 1.6$ as shown in fig. 2.1. For $\mu = 0.09$, the pairwise distance distribution is clearly a multimodal distribution (Hult et al. 2017). The peak near $d \approx 50$ represents a large number of very short pairwise distances between beads in the same cluster. For slightly larger d, the density drops to zeros, indicating a separation distance between clusters. Interestingly, we observe two more peaks near $d \approx 300$ and $d \approx 600$. The clarity of these peaks suggests that the clusters themselves are regularly spaced from one another, reminiscent of a lattice structure. This shows the three layers of the multiscale structure of the nucleolus for $\mu = 0.09$: its existence as a dense, secluded section of the nucleus, the self-organization of intra-nucleolar clusters, and the individual beads within each cluster. For $\mu = 0.19$, one can also observe in fig. 2.5(A) three peaks in the empirical probability density for bead-bead distances, but these peaks are much less pronounced. This shows a gradual transition in the degree of clustering as we increase μ. There is also a smaller gap between peaks. Together, these observations recapitulate our observations in fig. 2.1(E), wherein the clusters can be observed to be less compact. Finally, for $\mu = 1.6$, there is no multimodal structure in the bead-bead distance plot. This is consistent with our expectation that there is no clustering structure present for this range of μ.

The rigid and flexible clustering cases differ not only in how strong the clustering is at any given time, but also in how stable the structure is in time. We investigate this by considering how averaging pairwise-distances either across across time (fig. 2.5(B)) or over multiple simulations (fig. 2.5(C)) influences pairwise-distance probability densities.

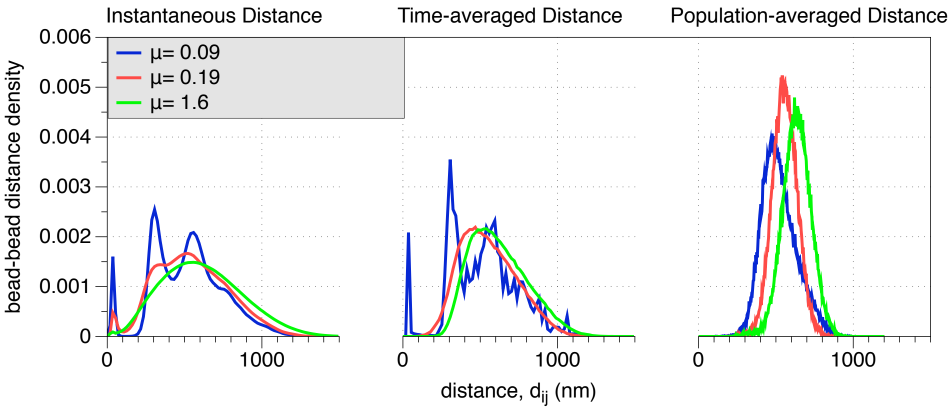

Figure 2.5: (A) Instantaneous distances, (B) time-average distances, and (C) population-averaged distances. In each panel, we show results for the three kinetic regimes illustrated in fig. 2.1: $\mu = 0.09$, $\mu = 0.19$), and $\mu = 1.6$. Multimodal histograms are a "signal" for the presence of clusters. This signal is strongest for $\mu = 0.09$, is nonexistent for $\mu = 1.6$, and we observe a new regime for $\mu = 0.19$ (flexible clustering or cluster plasticity), whereby 'soft' clusters form but deform over time through cluster interactions.

In fig. 2.5(B), we plot the empirical probability densities for pairwise distances averaged across our 20 minute simulations. Note for $\mu = 0.19$ that the density is no longer multimodal, implying that aggregating the data across a large time range inhibits the detection of flexible clusters, which by definition change with time. Note that the rigid clusters, which are very stable across time, remain discernible as the pairwise probability density remains multimodal. Unsurprisingly, the slow crosslinking appears qualitatively very similar in the long time average, as there was no apparent structure in the first place.

In fig. 2.5(C), we plot the empirical probability densities for pairwise distances at a single time but averaged across 10 simulations with different random initial conditions. Note for all μ that there is no longer any multimodal structure for these densities, highlighting that averaging across heterogeneous cell populations obscures the detection of clusters.

2.3.4 Flexible Clustering Regime Maximizes Bead Mixing

Next, we study how the kinetic time scale μ (i.e., and thus the presence of clusters) affects the properties of pairwise *gene interactions*. A pair of beads is said to be interacting if they are in very close proximity and the distance between them drops below d^*. As discussed in fig. 2.5, we choose $d^* = 100nm$ unless otherwise noted. In the following experiment, we show that increasing μ not only inhibits the formation of clusters, but that there exists a particular range of μ that optimizes *gene mixing*, or the overall interaction frequency of all pairs of genes. These experiments illustrate how clustering, which inherently describes multi-way

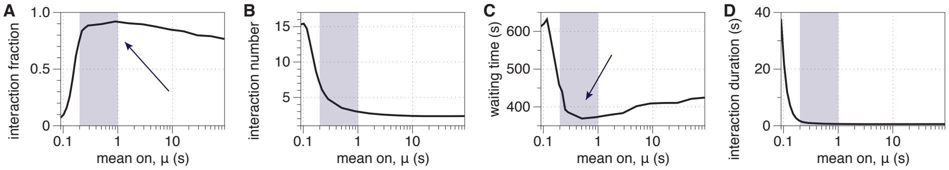

Figure 2.6: Mixing statistics for gene interactions: (A) mixing fraction; (B) mean interaction number; (C) mean waiting time; and (D) mean interaction duration, which we plot as versus μ. The shaded regions indicate the regime of flexible clustering, $\mu \in (0.19, 1)$. The arrows in panels (A) and (C) highlight that the interaction fraction and mean waiting time are both optimized for this range of μ. In contrast, this regime is associated with sharp transitions for the mean interaction number and duration, which both monotonically decrease with μ.

relationships, can be studied through pairwise distances, which inherently describe two-way relationships, and how there remain important open problems related to the time series signal processing of 4D chromosome conformation datasets.

We study the following summary statistics for gene mixing:

(A) The *interaction fraction* indicates the fraction of possible unique bead pairs that interact at least once during an interphase simulation.

(B) The *mean interaction number* indicates the number of simultaneous interactions (i.e., beads within distance d^*) for a nucleolus bead, averaged across time and across beads.

(C) The *mean waiting time* indicates for any two beads, selected at random, the average time that passes between their i-th and $(i + 1)$-th interactions.

(D) The *mean interaction duration* indicates the amount of time beads enter and reside within the interaction distance.

In fig. 2.6(A–D), we plot these summary statistics across a wide range of μ. We identify three regimes that optimize different attributes. $\mu \approx 0.1$ yields a self-organized structure that maximizes the number and the duration of gene-gene interactions (see panels (B) and (D)). Recall from Video 1 that $\mu = 0.09$ yields many large clusters that are stable (i.e., do not change) over time. This is reflected in a high number of interactions with beads in the same cluster and low number of interactions with beads not in the same cluster.

With $0.15 \lessapprox \mu \lessapprox 1$, we see flexible clustering behavior from Video 2. Notably, we find here that this flexible clustering has interesting properties beyond simply being a weaker version of the strong clustering from the rigid clustering regime. Namely, fig. 2.6(A) shows that these μ values maximize the fraction of pairs of beads that interact at least once over the simulation, and fig. 2.6(C) shows that these values minimize the waiting times between subsequent interactions. Thus, we can say that *flexible clustering promotes the number of both simultaneous and overall distinct pairwise gene interactions in the nucleolus*. This behavior arises from a balance between the number of intra-cluster gene-gene interactions, which is still elevated due to the moderate clustering as shown in (B), and the ability for genes to frequently switch between clusters during cluster interactions, as indicated by the reduced waiting time in (C). SMC proteins with such crosslinking timescales will thereby promote collective interactions among all active genes. These circumstances could accelerate a homology search, for example, to facilitate DNA repair, if the sister chromosomes were suddenly activated by a family of SMC proteins whose binding affinity was near this "sweet spot".

Finally, $\mu \gtrapprox 1.5$ is associated with a non-clustering regime, as shown in Video 3. The lack of clustering is reflected by a low number of gene-gene interactions, and the freely diffusing nature of the beads is reflected by short interaction duration and high interaction fraction.

We observe that a rapid change in the behavior of the system begins at approximately $\mu = 0.01$, most clearly visible in the interaction fraction and waiting time statistics. This indicates that the strong clustering behavior identified in (Hult et al. 2017) was in fact very close to a transition point in the system, although this was not known at the time. The fact that the behavior changes so drastically in this short range underscores the need for simulations with different values of μ varying on a fine resolution, and for metrics that reveal the way in which system behavior changes and allow for clear visualization.

2.3.5 Identifying Cluster Membership with Network Community Detection Algorithms

Having found that flexible clustering maximizes interesting properties of gene interaction, we seek to develop tools to identify and label the spatiotemporal clusters. In the rigid clustering regime, the clusters are so well-defined that any reasonable algorithm will detect them, but this is not the case for the flexible clustering. To detect and track flexible clustering, we utilize both spatial and temporal information to identify and track clusters.

While we have access to 4D bead position time series data, we begin by transforming this into a multilayer network problem as described in section 2.2.3.1. This is motivated by the fact that the most similar data available in biology, the Hi-C dataset, does not measure true distances between genome regions, but rather a notion of similarity based on average proximity (Dekker et al. 2002). The result of this transformation is a time sequence of weighted, undirected networks whose edge weights represent how near two beads tend to be to each other at that point in time. We refer to this sequence of networks as a temporal network.

Given a gene-interaction network, we identify communities using an approach based on multi-layer modularity (Mucha et al. 2010). See (Chen, Hero III & Rajapakse 2016, Cabreros, Abbe & Tsirigos 2016) for examples where community detection was applied to network models derived from Hi-C data. We present the algorithm in detail in section 2.2.3.2, and we briefly describe it here.

The *modularity measure* was originally introduced (Newman 2006) to detect communities in a single, non-temporal network; it is a scalar that quantifies—as compared to a null-model lacking communities—the extent to which a network's nodes can be partitioned into disjoint sets (i.e. communities) so that there is a prevalence of edges between nodes in the same community and relatively few edges between nodes in different communities. By searching over different possible ways to partition nodes into communities, one seeks to find an optimal partition that maximizes the modularity score (Blondel et al. 2008). Because each community contains a prevalence of edges, and edges only exist between pairs of genes that are in close proximity, a modularity-optimizing partition equivalently assigns genes into disjoint clusters so that each gene is nearer to genes in its cluster than to genes in other clusters.

Our analysis is primarily based on an extended version of modularity that allows one to detect time-varying communities in temporal networks and is called the multi-layer modularity measure (Mucha et al. 2010). In contrast to a community in a time-independent network (which is defined by a set of nodes), to specify a time-varying community one must also identify for each node the time-steps for which it is in the community. Using a variational technique (Jutla, Jeub & Mucha 2011-2019), we optimize the multi-layer modularity measure to simultaneously assign every node to a community at every time step. Each time-varying community in the network corresponds to a time-varying gene cluster, which is a set of genes that remain in close proximity for some duration.

A key feature of the multi-layer-modularity approach for community detection is that the framework involves two parameters, γ and ω, which provide "tuning knobs" (Mucha et al. 2010, Weir et al. 2017) to identify, respectively, the appropriate sizes and temporal coherence of communities/clusters. Parameter γ is a *resolution parameter* (Reichardt & Bornholdt 2006) and allows one to select whether modularity-optimizing partitions involve many small communities or just a few, very large communities. Similarly, ω is a *coupling parameter* and allows one to choose if the communities can change drastically from one time step to the next or if they are restricted to changing slowly over time. We explored a range of choices to select appropriate values. Finally, we highlight that this approach significantly contrasts traditional clustering algorithms such as k-means clustering (Kanungo, Mount, Netanyahu, Piatko, Silverman & Wu 2002), which specifies the number of clusters *a priori* (i.e., rather than allow the appropriate resolution to be dictated by the data) and which does not naturally extend to time-varying data.

We present now videos[2] of fast, intermediate, and slow crosslinking, in which we color the beads according to the community labels that we detect. These new videos provided qualitative evidence supporting the ability of our algorithm to find clusters that have appropriate spatial and temporal scales. Looking at the rigid clustering, we see the coloring strongly agrees with our visual perception in the clusters. A similar but less decisive conclusion can be made from observing the flexible clustering.

These videos indicate good agreement between visual perception of clusters and the clusters that are detected by the multi-layer modularity algorithm - when beads visually appear to be clumped together, they tend to also be the same color in the videos, which reaffirms the validity of our choice of clustering algorithm. However, especially when looking at the video depicting the non-clustering regime with slow crosslinking, we identify a key and common issue with clustering algorithms - they typically identify the “best” clusters, even when no clusters actually exist.

2.3.6 Gene Mixing at the Community Level Further Supports Flexible Clustering as the Mechanism for Optimality

In this section, we provide further evidence that flexible clustering is the mechanism that is responsible for the optimality observed in fig. 2.6. To this end, we will revisit and modify our definitions for *gene interactions* and *gene mixing*, which were defined at the “bead level” (i.e., for pairs of beads). We now define similar, but slightly different, concepts that are defined at the “community-level” in that they reflect only community-membership information and do not require the precise bead locations. We say that two beads are “communicating” if they are in the same community. That is, all beads in the same cluster are communicating with each other, and beads in different clusters are not communicating. With this modified definition in hand, we define summary statistics for gene mixing at the community level that are analogous to the 2-point

[2] Resources available at — Fast: `https://github.com/bwalker1/chromosome-videos/blob/master/Dataset0_color_finer_realtime_altView.mp4`, Intermediate: `https://github.com/bwalker1/chromosome-videos/blob/master/Dataset6_color_finer_realtime_altView.mp4`, Slow: `https://github.com/bwalker1/chromosome-videos/blob/master/Dataset12_color_finer_realtime_altView.mp4`

summary statistics for pairwise gene interactions that we previously defined in section 2.3.4. Analogous to gene mixing at the bead level, we now define "cross communication" at the community level.

(A) The *communicating fraction* indicates the fraction of bead pairs that are in the same community at least once during a simulation.

(B) The *average beads per community* indicates, for a nucleolus bead, the average number of beads in the same community at the same time, averaged across time and across beads.

(C) The *mean waiting time* indicates for any two beads, selected at random, the average time that passes between when they are no longer in the same community and when they are next in the same community.

(D) The *mean interaction duration* indicates the amount of time between beads when they are first in the same community and when they are no longer in the same community.

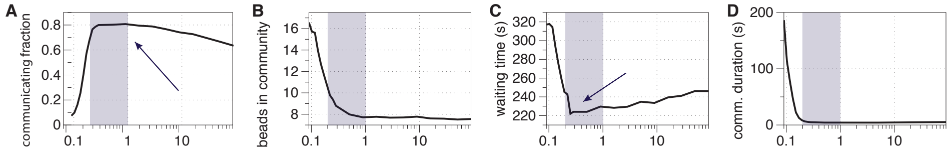

Figure 2.7: Cross communication describes the dynamics of community memberships of beads through community-level mixing. We study cross-communication for a large range of μ by plotting four summary statistics: (A) communicating fraction; (B) mean interaction number; (C) mean waiting time; and (D) mean interaction duration (see text). The shaded regions indicate the regime of flexible clustering, $\mu \in (0.19, 1)$, and the arrows in panels (A) and (C) highlight that the interaction fraction and mean waiting time are both optimized for this range of μ. These results are qualitatively identical to the results in fig. 2.6 for gene mixing, illustrating that cluster formation and the exchanges of beads between clusters determines the timescale of mixing.

In fig. 2.7, we present summary statistics for cross communication that are analogous to our results in fig. 2.6 for pairwise gene interactions. Note that the results in fig. 2.7 are qualitatively identical to those in fig. 2.6, supporting our hypothesis that gene-level mixing is determined by community-level cross communication. That is, the formation of gene clusters and exchanges of genes between them governs the timescale at which nucleolar domains come in close proximity of one another.

2.3.7 Temporal Stability of Clusters

Using temporal community detection, we are now finally able to quantitatively support our first observations made in section 2.3.1 that there are three distinct clustering regimes: rigid clustering, flexible clustering, and no clustering. We support these observations by studying the properties of the detected clusters.

In fig. 2.8(A)–(C), we plot the average lifetime of clusters as a function of their average size (averaged over time). Panels (A)–(C) indicate the three clustering regimes with $\mu \in \{0.09, 0.19, 1.6\}$, respectively.

For the rigid clustering regime, in fig. 2.8(A), we see that most of the clusters are large, with an average size of 10 or more beads, and also have a long lifetime of over 100 seconds. This is consistent with our prior observations (e.g. Video 1) that showed large clusters that appeared very stable in time. We also see that the large clusters survive for much longer than the small clusters.

For the flexible clustering regime, in fig. 2.8(B), we can see the same general trend that larger clusters tend to have a longer lifetime than smaller clusters, but there is a much wider spread of cluster sizes, with only a moderate number of large clusters.

For the non-clustering regime, there appears to be little relationship between cluster size and stability beyond an average size of approximately 3 beads. The clusters also tend to be much smaller, with almost no clusters with an average size over 10 beads.

In fig. 2.8(D)–(E), we plot the probability that a bead remains in the same community upon the next timestep, again as a function of cluster size. Panels (D)–(E) indicate results for $\mu \in \{0.09, 0.19, 1.6\}$, respectively. In agreement with panels (A)–(C), one can observe that larger clusters are more stable. Note also that the communities exhibit more plasticity for $\mu = 0.19$ than for $\mu = 0.09$ since beads have a higher average probability for changing the community to which they belong.

2.3.8 Comparison of Community Detection Algorithms

We begin by comparing the two single-layer community detection algorithms: the connected components algorithm and the single-layer modularity optimization algorithm. We are interested in observing how the resolution parameter γ in the modularity optimization affects the results. Recall that γ controls the scale of the null-model penalty in the modularity equation, eq. (2.3), and that a higher value of γ corresponds to an increased penalty for putting poorly-connected beads in the same community, and thereby fewer communities.

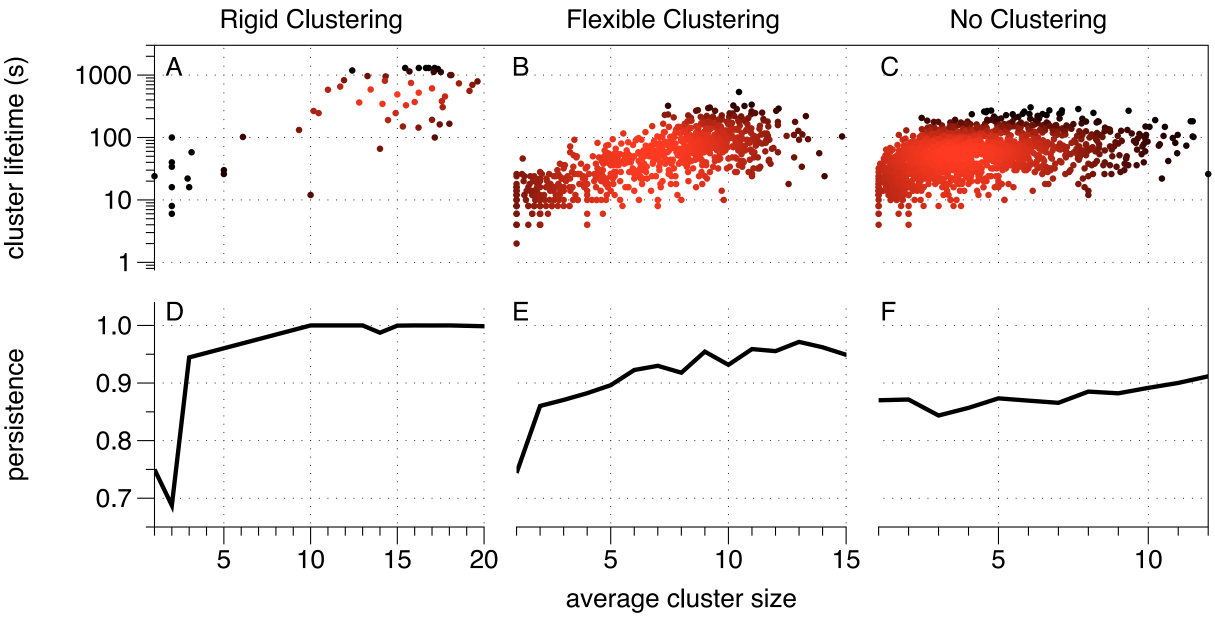

Figure 2.8: The birth and death of gene clusters identified using modularity-based community detection in temporal networks. Panels (A), (B) and (C) depict the lifetime (i.e., duration) of each gene cluster versus the average cluster size for $\mu = 0.09$, 0.19, and 1.6, respectively. The points' colors have been chosen to highlight the density of points (with red indicating where there are many points close to one another). (D)–(F): The persistence (i.e., temporal coherence) of clusters is indicated by the probability that a randomly selected bead remains in the same clusters in the next time window, again plotted versus the average number of beads in that clusters. Results reflect $d^* = 325$, $\gamma = 10$, $\omega = 1$.

There is, however, no *a priori* value for γ - its value should be chosen based on the problem, with a particular thought given to sensitivity analysis: we expect that a meaningful community partition should not be highly sensitive to variations in γ. To this end, we compute partitions for a large range of values of γ, and use the CHAMP algorithm (Weir et al. 2017) to select those resulting partitions that are optimal in terms of modularity for some range of γ. Because the resolution parameter affects the size of communities, or equivalently the number of communities, we can measure the sensitivity by looking at variation in the number of communities as we change γ. In a region in which the clustering is more robust, we expect to see the number of communities not significantly increase even as we increase γ. We also measure the similarity between the partition from the single-layer modularity and that from connected components using Adjusted Mutual Information (AMI), a standard metric for measuring similarity of partitions. An AMI value of 1 indicates perfect agreement between partitions (modulo permutation of labels), whereas a value of 0 indicates no better agreement than would be expected by chance.

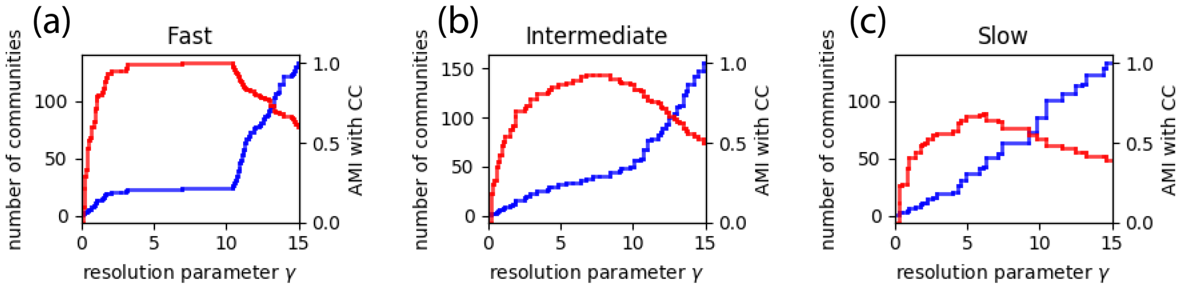

Figure 2.9: Plot of number of communities (blue) and AMI with connected components partition (red) for the single-layer modularity community detection algorithm. (a) Fast crosslinking, showing wide range of $\gamma \in [3, 10]$ in which the number of communities is fixed and the AMI is near 1. (b) Intermediate crosslinking, with higher AMI in the same region but no clear region of robustness. (c) Slow crosslinking, showing no region of robustness and minimal agreement between the connected components and modularity.

Figure 2.9 shows the number of communities and AMI with connected components as a function of gamma for the three crosslinking regimes. In fig. 2.9(a), we see the results in the fast regime, where we expect the best performance, as we have seen the clustering to be the most clear. Indeed, in this case, we see the existence of a large range of approximately $\gamma \in [3, 10]$ in which the number of communities is almost constant, indicating a very high level of robustness. We also see that the AMI is very close to 1, indicating near-total agreement between the two algorithms. We can conclude that there is very clear clustering in the fast crosslinking data, as we had previously seen by eye, and that both the connected components algorithm and the single-slice modularity for $\gamma \in [3, 10]$ recover this partition.

However, this very clear signature is not preserved in the intermediate or slow crosslinking cases. Considering the intermediate crosslinking in fig. 2.9(b), we do not see a clear flat region in the blue curve showing the number of communities. There is a clearly visible change around $\gamma = 10$, similar to the fast crosslinking, at which the number of crosslinking starts increasing much more rapidly. Visually, the region $\gamma \in [5, 10]$ appears to be the closest to flat, at is also the most similar to the connected components, nearly identical around $\gamma = 7$. We note, however, that in this case the connected components algorithm is no longer expected to produce a reliable baseline as it did in the fast crosslinking case, as there is no longer a clear separation of distance between clusters. Visually, we have seen the presence of clusters in the intermediate crosslinking, so we would like to have an effect algorithm for identifying them. This reinforces the drive to incorporate additional information through the use of the multi-layer modularity algorithm.

We will also briefly consider the analysis on the slow crosslinking, although we do not see clusters forming in this parameter regime. Consist with this expectation, we see a steady increase in the number of clusters as we change γ, with no clear region of any level of robustness. This is consistent with expectations associated with applying the algorithm to data with no underlying clustering structure.

2.3.9 Calibration of Multi-layer Modularity

In the previous section, we demonstrated the use of the CHAMP algorithm to demonstrate that the single-layer modularity community detection algorithm was producing robust, meaningful partitions for the range $\gamma \in [3, 10]$ in the fast crosslinking. However, this stability was not preserved in the intermediate and slow crosslinking case.

We now move on to considering the multi-layer community detection algorithm, which generates a partition simultaneously over all layers in the multi-layer network, seeking to optimize the multi-layer modularity criterion eq. (2.3). In addition to the resolution parameter γ, this introduces a new parameter, the coupling parameter ω, which describes the penalty associated with a bead switching community from one layer to the next. We will now apply the two-dimensional extension of the CHAMP algorithm, by computing a large number of partitions with varying values of both γ and ω and seeking those partitions that are optimal in the modularity criterion over some region of the combined $[\gamma, \omega]$ space. We will use this to justify our choice of the values of γ and ω used in our previous analysis.

We consider overall a range of parameters $\gamma \in [0, 15]$, $\omega \in [0, 5]$, although with the expectation that each of these parameters should in fact be positive. Our visualization of the output of the CHAMP algorithm is presented by showing the regions each optimal partition is optimal for in this region, colored based on their average AMI with neighbors. This means that a part of the space in which regions are large and colored purple (indicating high agreement with neighbors) corresponds to low sensitivity, and therefore high robustness, of the partition with respect to γ and ω, and areas where partitions are small and colored blue indicate much higher sensitivity, and correspondingly lower robustness, with respect to these parameters. The results for the fast crosslinking are presented in fig. 2.10. Because we have already established that this parameter regime produces an essentially trivial community detection problem, we expect there to be a wide range of parameter values producing essentially the optimal result. Indeed, we observe that the region of the space approximately described by $\gamma \in [3, 13]$ and for almost any value of ω is covered by a small number of partitions that have a very high mutual AMI and are therefore very similar. Intuitively we can understand that these partitions are in agreement on the actual clusters, with only small variation most likely focused on the occasional isolated beads. As long as we have ω somewhat positive to encourage alignment of the partition between layers, we also see negligible dependence on its particular value. This also makes intuitive sense, as there is so little temporal variation in the clusters that this penalty will rarely come into effect.

We can move on to considering the intermediate crosslinking case, for which the results are shown in fig. 2.11. We know from inspection that the clusters are significantly less stable and less separated, and so we expect there to not be the same degree of a single “correct” partition leading to near-total agreement over a wide range of the parameter space. However, we do still observe in the bottom-right of the image, an area in which partitions cover a fairly large range of space with significantly higher pairwise AMI than in the other parts of the space. We note that for sufficiently large γ, the algorithm will produce partitions in which every node is in its own “community,” which is robust in γ but uninformative. However, the observed robustness for values of $\gamma \approx 10$ can be observed to correspond to nontrivial clustering, and these two facts in combination indicate extraction of significant clustering.

Finally, we present the equivalent results for the slow crosslinking case in fig. 2.12. We see, unsurprisingly, that there is significant variation in the partitions based on the parameters, as we do not expect there is any real underlying community structure. We also note that in this slow crosslinking case, there seems to be increased sensitivity to ω relative to the other two cases. This makes sense under the perspective that in the slow crosslinking case, the lack of stable clusters would mean much higher levels of community switching in a given partition, so as ω was increased and the algorithm was forced to keep beads together, it would lead to a greater change than in the fast and intermediate crosslinking cases, in which the presence of clusters meant more beads were remaining in the same community over time in the first place. Based on these plots, we select values of $\gamma = 10$, $\omega = 1$, which seem to lie in the most stable regions of the plots for fast and intermediate crosslinking.

2.4 Discussion

The dynamic self-organization of the eukaryote genome is fundamental to the understanding of life at the cellular level. The last quarter century has witnessed remarkable technological advances that provide massive datasets of both the spatial conformation of chromosomal DNA from cell populations (3C and Hi-C generalizations) and the dynamic motion of domains in living cells (GFP tagging and tracking of specific DNA sequences), from the yeast to the human genome. Data mining of this massive data has likewise witnessed remarkable advances in understanding the hierarchical packaging mechanisms of DNA that act on

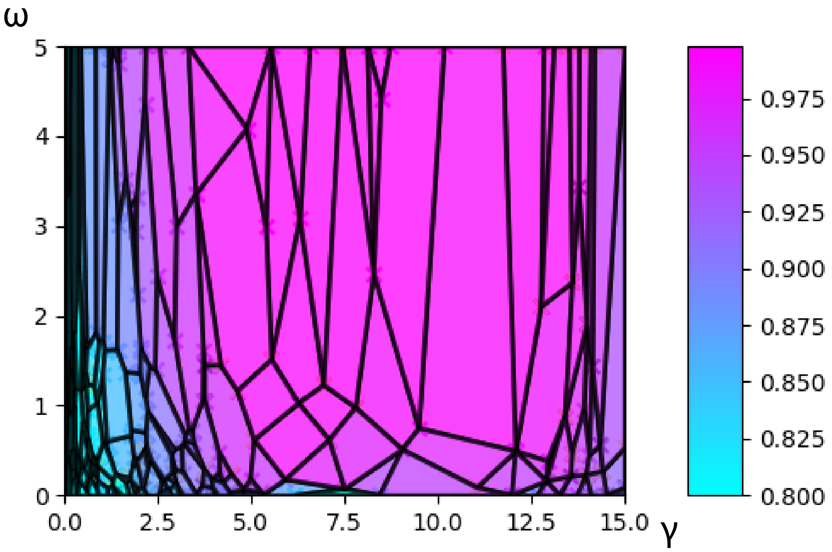

Figure 2.10: 2D CHAMP plot for fast crosslinking data, showing very widespread agreement between partitions for a large fraction of the space, where $\gamma > 3$ and for essentially any values of ω.

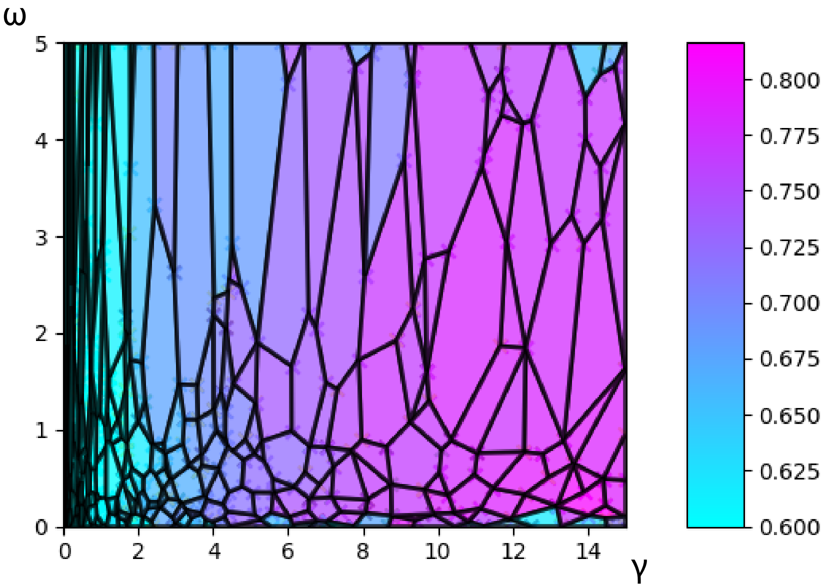

Figure 2.11: 2D CHAMP plot for intermediate crosslinking. Compared to the fast crosslinking, there is much more variability in the partition as the parameters change, but the most robust area appears to be for $\gamma > 10$.

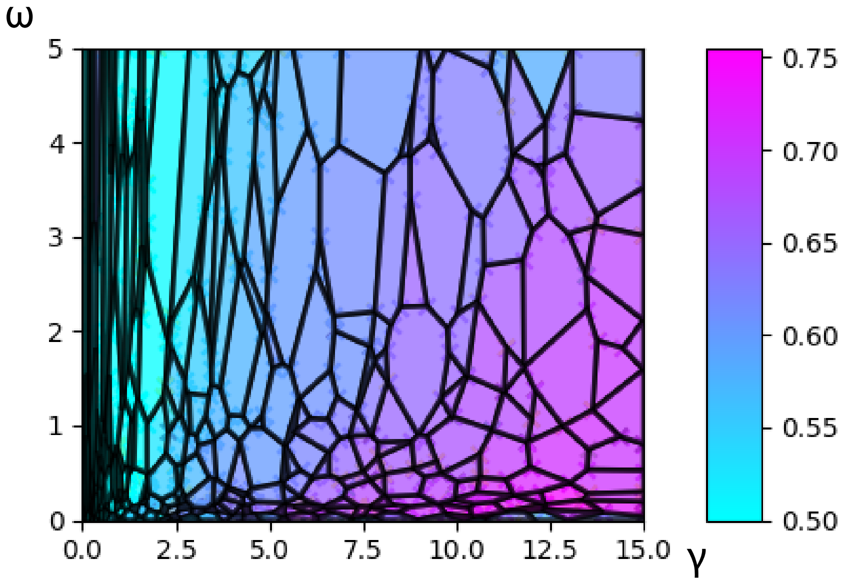

Figure 2.12: 2D CHAMP plot for slow crosslinking. Compared to the fast and intermediate crosslinking, it is much more sensitive to the coupling parameter ω, which agrees with the intuition that the lack of clustering means beads will not stay in the same community across layers if the ω penalty doesn't force it.

top of the genome, e.g., histones and structural maintenance of chromosome (SMC) proteins, the topology of individual chromosome fibers, their topologically associated domains, and the territories they occupy in the nucleus. The third wave of advances has come from 4D modeling of chromosomes based on stochastic models of entropic, confined polymers, and the coupling of SMC proteins that either bind and crosslink genes on chromosomes or generate loops on individual chromosomes. As these three approaches continue to mature and inform one another, at an ever-increasing pace, insights into the structure and dynamics of the genome continue to deepen.

The motivation for the work in the chapter lies in the information that can be inferred from these massive datasets, from Hi-C, live cell imaging experiments, and polymer physics modeling. In previous studies, cf.(Hult et al. 2017) and references therein, we showed that heterogeneity in experimental images derive from substructures that are formed within the nucleolus. We decided to use the array of tools from network community detection analysis, modeling, and associated fast algorithms, to automate the search for dynamic architectures in the 4D datasets generated by polymer modeling. We note a similar network analysis approach has been applied to Hi-C datasets (Rajapakse et al. 2010, Rajapakse & Groudine 2011, Rajapakse, Groudine & Mesbahi 2011), whereas our datasets have the added feature of highly resolved temporal information. Our aim was to infer organization beyond 2-point, time-averaged or population-averaged, gene-gene proximity statistics and heat maps generated from the statistics, and to remove the bias of an individual's visual determination of structure. To do so, we used the advances in network-based models, their temporal generalization, data analysis, and algorithms, and applied this arsenal of tools to 4D datasets across four decades of crosslinking timescales to: (i) robustly identify clusters, or communities, of genes (5k bp domains, or beads, in our model), i.e., to directly detect gene sub-organization at the scale it exists rather than attempt to seek larger scale organization from gene-gene statistics and heat maps; (ii) determine the size distribution (number of genes) in such sub-structures; (iii) determine the persistence times of communities; and, (iv) determine the interaction frequency of communities and corresponding gene exchanges, which are the drivers of gene-gene interaction statistics. In this way, network algorithms automate gene community detection and persistence, with robustness built in by enforcing insensitivity to algorithm tuning parameters.

We elected to build and implement these network tools on the 4D datasets generated in house, from simulations of our recent polymer modeling of interphase budding yeast (Hult et al. 2017). In this model, a pool of SMC proteins transiently and indiscriminately crosslink 5k bp domains within the nucleolus on Chromosome XII. The kinetics of the cross-linking anchors relative to the substrate is a major driver of sub-nuclear organization. If the crosslinkers bind and release more rapidly than the chains can relocate, the non-intuitive consequence is that the chains explore less space. When dense clusters of crosslinker/binding sites arise, they persist for extended time periods when the crosslinking kinetics is sufficiently fast.

We previously showed in (Hult et al. 2017) that very short-lived ($\mu = 0.09sec$) binding kinetics provided closer agreement with experimental results (highest degree of compaction of the nucleolus into a crescent shape against the nuclear wall). It was also shown via visualization of the simulated 4D datasets that this timescale induces a decomposition of the nucleolus into a large number of clusters each consisting of many 5k bp domains, and these clusters were persistent over time. On the other hand, with long-lived crosslinks ($\mu = 90sec$) the clusters disappeared. These results reveal that the timescales of the crosslinkers relative to entropic fluctuations of the chromosome polymer chains are a fundamental contributor to genome organization.

Here, the sample set of binding kinetics in (Hult et al. 2017) was expanded to 4D datasets of interphase, sampling over four decades of crosslinking duration timescales. We applied standard, distance-based, 2-point statistical metrics and visualization tools, and then analyzed the full range of 4D datasets with the fast, automated network models and tools. The network algorithms search for and detect time-varying communities (clusters, sub-structures) at the scales they exist, not at a prescribed scale of two or more genes; the spatial and temporal scales are identified with the criteria that they are robust to the algorithm parameters. We then use this information to label and color-code communities, using community-level description to understand the persistence and interactions (merger and division marked by gene exchanges) between communities.

With the above community-scale information and statistics, we generalize standard gene-gene interaction statistics across the four decades of bond duration timescale. As a generalization of waiting times for 2 distant genes beads to come within a specific distance of one another, we calculate waiting times for genes in the same community to leave and then re-enter another common community, and calculate the fraction of all genes that were in the same community at least once during interphase, which we call the community cross-communication fraction.

From these analyses, we discovered a novel dynamic self-organization regime, wherein the rigid, persistent communities at relatively short-lived crosslink timescales ($\mu = .09 sec$) transition at slightly longer-lived crosslink timescales ($\mu = 0.19 sec$) to more mobile (literally, the clusters diffuse faster) communities that interact far more frequently, each interaction corresponding to merger, subsequent division, and an exchange of genes. We refer to this regime as flexible community structure with enhanced cross-communication. Furthermore, we discovered non-monotonicity in the dynamic self-organization behavior: the community cross-communication fraction is maximized, coincident with a minimum waiting time between genes departing and returning to common communities, with a crosslink timescale of $\mu = 0.19 sec$. Both properties fall off, albeit in different ways, for shorter and longer timescales.

We emphasize that these network tools and fast algorithms are amenable to any 4D dataset from polymer models. While we restricted the analysis in this study to the nucleolus during G1 of budding yeast where SMC proteins are allowed to transiently crosslink 5k bp domains, the same analyses can be applied to data with tandem SMC crosslinking and loop generation, for any cell type and for any phase of the cell cycle. Moreover, given the growing interest in network-based analyses for Hi-C data, network modeling is well-positioned to provide a fruitful direction for data assimilation efforts aimed at connecting simulated and empirical 4D chromosome conformation data. An important challenge facing this pursuit is the development of improved data pre-processing and community-detection methodology for temporal and multimodal network datasets (Taylor, Shai, Stanley & Mucha 2016, Taylor, Caceres & Mucha 2017).

Having established the the existence of complex dynamic emergent clustering behavior in our model, we now set our sights on building a deeper mathematical understanding of the effective force created by this on/off switching crosslinking force and how that affects the resulting dynamics. We would like to pursue a principled average that is able to capture both the rigid clustering and the flexible clustering that emerges at slightly longer values of μ. As μ does not affect the magnitude of the force or the fraction of time it is active, this already suggests the need for a deeper consideration than simply averaging out the force based on the fraction of time the pairs of beads are bound. Additionally, several computation aspects of the model, such as the normally distributed waiting times in the active and inactive states, and the finite singularity in the spring force function, would pose difficulties in averaging. Motivated by this, with a goal of understanding the way in which switching forces produce emergent behavior. In Chapter 3, we will address these problems by constructing an alternative version of the model that will maintain the behavior of interest identified in this chapter, the combination of rigid clustering at fast timescale with flexible clustering maximizing mixing at intermediate timescale, verifying the equivalence using the same statistics introduced here.

CHAPTER 3: MODEL REDUCTION

3.1 Introduction

The previous chapter introduced the model of the interphase yeast genome from (Hult et al. 2017, Walker, Taylor, Lawrimore, Hult, Adalsteinsson, Bloom & Forest 2019) and demonstrated the presence of complex spatio-temporal clustering behaviors modulated by a dynamic timescale parameter μ. This behavior is strongly reminiscent of particles moving on an energy landscape – at low energy, the particles condense into the energy wells, as observed in the fast crosslinking, and at high energy, they move around freely, as observed in the slow crosslinking. This leads us to seek an equivalent representation of our model in terms of an energy landscape. This would then allow us to measure predict existence and size of clusters by viewing them as energy wells, as well as understand their stability by analyzing the depth of the well.

We would like to apply the theory and methods discussed in section 1.3 to this system, but as the model that we investigated in Chapter 2 is a complex stochastic model that was constructed with a goal of maximizing similarity to the actual dynamics of the DNA using a bead-chain polymer with a resolution of 5kbp per bead, we will need to make several modifications first. The primary goal of this chapter will be to introduce these modifications, demonstrate the purpose they serve in facilitating the desired averaging, compare with the original model used in Chapter 2, and show that these reductions maintain the identified behavior of interest. The modifications we will make come in two forms:

1. Simplification of the model by eliminating parts that are not required for the emergence of the spatio-temporal clustering behavior
2. Reformulation of parts of the model to create functional forms amenable to the desired analysis

Towards the former, we will eliminate the chain force and all non-nucleolar beads from the model. Towards the latter, we will change the nuclear boundary from a hard boundary to a soft boundary, and reformulate the random process by which bonds form and break from Gaussian waiting times to exponential waiting times, such that they follow a continuous-time Markov chain (CTMC) process, which is both more physically inspired and suitable for future averaging work.

We will present our reduced model in two forms: first, a general form with free parameters designed to create a framework for investigation of the relationship between parameters and behavior, and then a specific set of such parameters that are designed and shown to produce agreement with the previous chapter under the reduced model.

3.2 New Model

Based on the reductions described above, we will describe here a general form for a minimal model that capture the dynamics observed in (Walker et al. 2019). To serve as reference, the dynamics of bead i are given by

$$\dot{x}_i = -cx_i + \sum_{j \neq i} \alpha k b_{ij}(x_j - x_i) + \sum_{j \neq i} c_{ev} \exp\left(-\frac{(x_j - x_i)^2}{a_{ev}}\right) \tag{3.1}$$

for $x_i \in \mathbb{R}^d$, where b_{ij} is 1 if i and j are currently bound, and otherwise 0. Unlike the model from Chapter 2 in which beads switched between active and inactive states with normally distributed waiting times, the binding/unbinding is defined by exponential waiting times in a manner that constitutes a continuous time Markov Chain (as will be used in Chapter 4).

Details and justification are shown in the section 3.3, and the parameters are investigated in section 3.5 based on comparison of metrics applied to simulations of each model.

3.3 Changes to Model

In this section we will describe all modifications that are made to the model from (Walker et al. 2019) to arrive at the model shown in section 3.2. The modifications are introduced piece by piece, and subsequently in section 3.5 the marginal effect of each modification will be investigated.

These modifications come in two stages: in section 3.3.1, we strip out extraneous dynamics that are not necessary to create the clustering and mixing behavior we are interested in; in section 3.3.2, we introduce the new formulation of the stochastic switching for the crosslinking.

3.3.1 Reduced Model

We begin by removing elements of the model that are not directly relevant to the formation of clusters by pairwise dynamic crosslinking. The results of interest in (Walker et al. 2019) were in the 361 beads comprising the nucleolus — the non-nucleolar beads did not significantly contribute to these dynamics, except in that they pushed the nucleolus into the nuclear wall through the excluded volume force. For this reason, we will simply remove these beads from the model. However, we must ensure that the remaining nucleolar beads are still confined to a similar amount of volume. While in the original model, they occupied a volume between the non-nucleolar beads and the hard nuclear wall, we will in the reduced model impose an artificial "confinement" force, designed to confine the nucleolar beads to a desired volume without using a hard boundary that is numerically challenging. This takes the form of a harmonic energy well. We will show that a spring constant of 0.002 leads to agreement with the original model in section 3.5.1.1.

The SDE that governed the motion of the beads in the original model contained the following terms: polymer chain bonds, excluded volume, dynamic crosslinks, and thermodynamic noise. We remove the polymer chain bonds entirely, as the magnitude of the force is negligible compared to the stronger springs representing the dynamic crosslinking.

We will also simplify the dynamic crosslinking force by replacing the wormlike springs, for which the force function exhibits a singularity at finite length, with harmonic springs.

The original wormlike spring formula is of the form

$$F_C = \alpha \left(-1 + \frac{1}{\left(1 - \frac{r}{45}\right)^2} + \frac{4r}{45} \right) \tag{3.2}$$

where r is the spring displacement and α is a constant determining dynamic crosslinking spring strength. In the model, under standard model units, its value is 10.8794. section in the parameter choice bit or something

We will replace these wormlike springs with linear springs, introducing a new spring scale parameter k, relative to the original scale parameter α, leading to a new spring force of the form

$$F = -\alpha k r \tag{3.3}$$

From a Taylor expansion of eq. (3.2), we can observe that for the linear spring force in eq. (3.3) to agree with the slope of the wormlike spring at $r = 0$, we would take a spring constant of $k = \frac{2}{15}$. However, because the spring is typically extended some fraction of the maximum extensibility of $45nm$, we will consider larger values of k, which can be understood as choosing the linear spring force to agree with the wormlike spring force at some particular value of r. We investigate the effect of the choice of k on the dynamics in section 3.5.1.2.

3.3.2 New Crosslinking

3.3.2.1 Crosslinking as a Continuous-Time Markov process

As an alternative to the formulation described in section 3.3.1, we consider a modified version that is hopefully more physically motivated and also more amenable to the mathematical analysis that we want to perform. This new model is based on a continuous-time Markov chain formulation, and is described as follows:

Each bead can be either unbound, or bound to a single other bead. Bonds are symmetric. If a bond is formed between two beads, assume that the lifetime of the bond is an exponentially distributed random variable with rate α, meaning it has an expected lifetime of $\frac{1}{\alpha}$. To simulate it, one can simply draw such an exponentially distributed random variable and use it as the lifetime.

Unlike the original model which greedily formed bonds between pairs of beads in order of increasing distance, we assume here that there is an "affinity function" that represents the preference of bonds to form between beads that are closer. Specifically, the function $a(r)$ gives the (exponential process) rate at which bonds form between two currently unbound beads that are separated by a distance r. For all possible pairs of beads i, j with positions $\vec{x}_i, \vec{x}_j$, a bond will form at a rate $a(|\vec{x}_i - \vec{x}_j|)$. At a particular instant in time, let the matrix

$$M_{ij} = a(|\vec{x}_i - \vec{x}_j|) \tag{3.4}$$

represent the transition rates into the bound states (while still unbound). Note that as soon as a bond forms, the bead is no longer in an unbound state and so these rates no longer apply. Recall again the rate α representing bonds breaking. For the sake of mathematical convenience, we will describe the unbound state as the bead being bound to itself.

We introduce also the matrix A_{ij}, whose entries represent the state of each bead. In other words, $A_{ij} = 1$ iff beads i and j are bound (note: $A_{ii} = 1$ indicates bead i is unbound). Clearly each row of A is entirely zeroes except for a single one, making it a permutation matrix. Also, it is symmetric.

3.4 Method of Comparison

In order to quantify what effects the changes we make to the model have on the resulting dynamics, we consider here two of the statistics from section 2.3.4, the average nearby statistic and the mixing coefficient.

These two statistics are chosen because they represent the two key behaviors observed in (Walker et al. 2019) – the average nearby statistic exhibits a monotonic decrease as the key mean on parameter μ increases, whereas the mixing coefficient is maximized in the intermediate crosslinking/flexible clustering regime. In the following results section, we will show that this same behavior is preserved through the modifications made to the model, demonstrating that we are not losing the fundamentally important aspects of interest shown in Chapter 2.

3.5 Results

3.5.1 Parameter Selection

In this section, we discuss the choice of new parameters that are introduced through the simplifications. When converting the hard nuclear boundary to a soft well, we must choose the strength of that well, which is discussed in section 3.5.1.1. When converting the wormlike springs to linear springs we must choose the corresponding spring coefficient, which is discussed in section 3.5.1.2.

3.5.1.1 Confinement Parameter

As described in section 3.3.1, we replace the hard boundary nuclear wall with a soft boundary through a cubic spring force. The spring constant associated with this force will control the overall volume occupied by the beads, and so we aim to choose it so that the scale of the volume is similar in the reduced model as in the original model.

To determine an appropriate value of the confinement parameter c, we alter it and observe the effect on the value of the average nearby statistic for both fast and slow crosslinking, shown in fig. 3.1. From the original model, we expect a value of approximately 3 in the slow crosslinking and 15 in the fast crosslinking. We can observe that in this reduced model, for any particular value of the confinement coefficient, the slow crosslinking has a higher value relative to the fast crosslinking than would be expected.

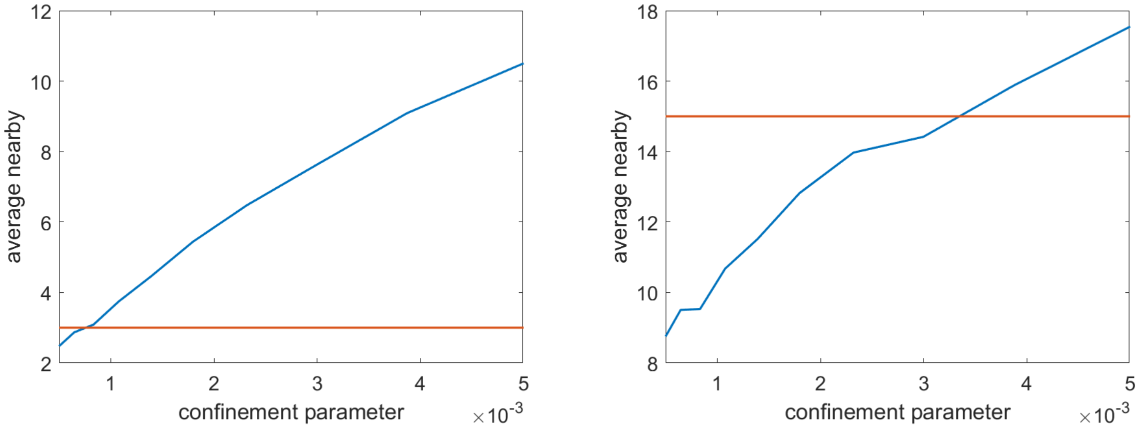

Figure 3.1: Comparison of average nearby statistic (proxy for density) across various values of the confinement strength parameter for non-clustering regime (left) and clustered regime (right).

The density agrees in the slow crosslinking at a value for the confinement parameter of approximately 1×10^{-3}, whereas agreement for the fast crosslinking occurs around 3×10^{-3}. This relationship may be grounded in the different mechanics of the new crosslinking, which we will discuss further in section 3.5.2. However, we also note that both statistics simply vary monotonically with the parameter, so there does not appear to be significant sensitivity of the qualitative dynamics to the choice of this parameter. For the remainder of this chapter we will use a value of 2×10^{-3} for the confinement parameter, in the middle of this range.

3.5.1.2 Linear Spring Constant

In this section we investigate the effect that the choice of the spring constant for the crosslinking force has on the dynamics. Recall that in eq. (3.3), we proposed a parameter α that scales the linear spring relative to the physical parameter k defining the strength of the wormlike spring in section 1.4. We will investigate how the choice of this parameter α affects the resulting dynamics of the system.

Because the wormlike spring used in section 1.4 diverges to infinity at a finite maximum extensibility, it is always the case that the linear spring will be weaker than the wormlike spring for sufficiently long extensibility. However, due to the observation in simulations that nearby beads in clusters are often separated by distances of 30nm, we would intuitively expect that we would want to choose the linear spring force to be similar in strength at roughly this length scale. This would mean being stronger than the wormlike spring at lesser displacements, and weaker at greater displacements. We choose a value of $\alpha = 1$ as it just follows the plateau in the average nearby statistic.

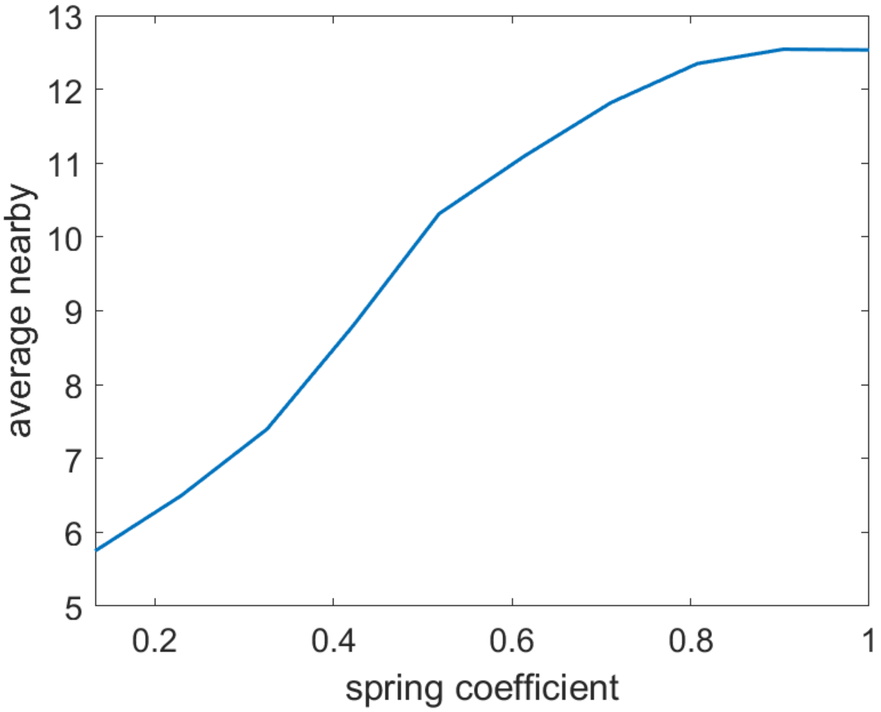

Figure 3.2: Average nearby statistic (representing size of clusters) for a range of values of the spring strength parameter α from eq. (3.3), with $\mu = 0.09$. We see that as α increases to values near 1, the average nearby statistic levels out at approximately 12.5, which is similar to the values seen in the fast crosslinking in Chapter 2. This suggests that $\alpha = 1$ is a reasonable choice for our reduced model.

3.5.2 Fraction of time bound by bead

A key difference in the new crosslinking scheme is that the rate at which an unbound bead forms a new bond depends on how many other beads are nearby. This is due to the CTMC formulation in which the bead has an independent chance at each point in time (or equivalently, independent rates), to form a bond with every other bead nearby, scaled by the value of the affinity function. This contrasts with the original crosslinking model, where the bead would immediately form a bond with the nearest other active bead within range once it became active again. This means that a bead on the outside of a cluster will take longer to form bonds than beads in the interior of a cluster, as they are further away from the beads in the cluster.

This is compounded with the fact that the interior beads are more likely to bond with each other, thereby reducing the number of available partners for exterior beads. With these two factors combined, we expect

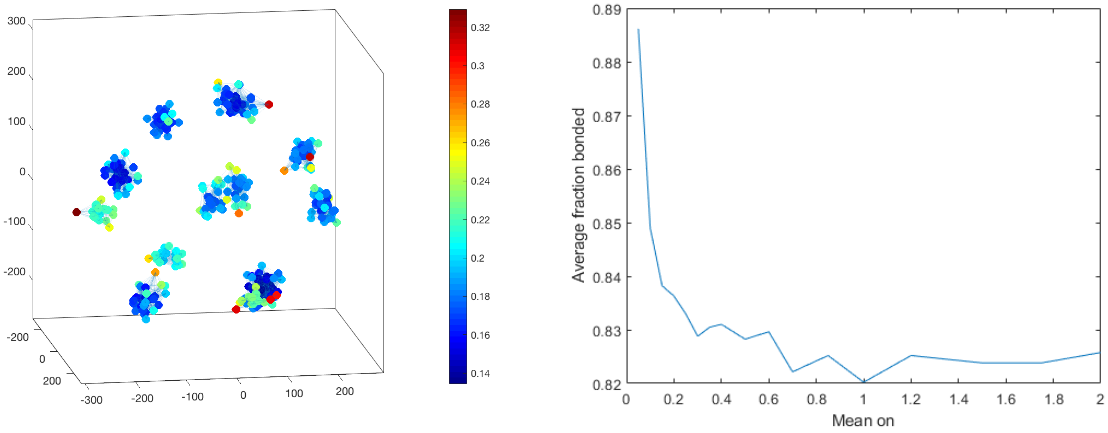

Figure 3.3: (A) Plot of fraction of time unbound for beads in a connection network (represented by color). Note that beads on the extremity of a cluster spend a significantly lower fraction of time bound than those in the middle. (B) Fraction of time spent bound as a function of the mean on parameter μ.

to see a significant difference in the fraction of time spent bound for beads depending on their position in a cluster. To test this, we take a set of clustered bead positions from a simulation and then freeze the beads in space. With the bead positions unchanging, we then compute the expected fraction of time that each bead would spend unbound on average, as shown in fig. 3.3(A). Exactly as expected, we can see that the beads on the extremities of clusters are colored red, corresponding to a high probability of being unbound, as high as approximately one third. This contrasts with beads in the center of large clusters, for which the value can be as low as 0.14. thus we can see that even for two beads that are clearly in the same cluster, beads towards the outside may spend twice as much time unbound as those in the center.

In line with our goals of identifying parameter values that maximize agreement with the original model, we also ask about the overall fraction of time beads spend bound. In the original model, we fixed the fraction of time that beads are active at 90% by instituting the 9/1 ratio between mean on and mean off times. For the same reason as outlined above, we would expect that as the crosslinking slows and the density lowers, we would see a corresponding lowering in the fraction of time that beads spend bound, event for the same affinity function.

The results, shown in fig. 3.3(B), agree with this expectation. This is indicative of a fundamental difference in the CTMC binding scheme, that beads will form crosslinks more quickly when surrounded by more other beads. We do observe that the overall range of variation is still within the range $(0.82, 0.88)$, and therefore doesn't appear to be a huge change, given that we are still able to observe the expected clustering behavior, and the difference primarily arises in the non-clustering regime which we are not as directly interested in.

3.6 Discussion

In this chapter, we discussed a series of changes made to the model from (Hult et al. 2017, Walker et al. 2019) designed to maximize the mathematical simplicity while maintaining the particular behavior of interest. This signals a change in perspective, from one of understanding the biology through a model designed to agree with the underlying mechanics, to one of trying to understand the mathematics of the resulting emergent structure.

We set out in this chapter motivated by the work in Chapter 2, in which we showed our model to exhibit complex emergent spatio-temporal clustering behavior mediated by the dynamic timescale parameter μ, noting in particular non-monotonic dependence of the bead-bead mixing on μ. Combined with observations that μ seemed to behave as an "effective temperature," with beads condensing into clusters at low μ but spreading out over the domain at high μ, we were motivated to pursue a further investigation into determining an averaged, "effective force" representing the stochastically switching on/off crosslinking, and with it a corresponding effective energy landscape.

In this chapter, we made a series of modifications to the model from Chapter 2, seeking to follow the two goals laid out in section 3.1, both reducing complexity and changing mathematical forms to similar ones more amenable to our planned averaging work, but without changing the fundamental dynamic clustering behavior that we had observed previously. We did this by making a series of incremental changes, and performing each time a comparison between old and new using the statistics that we created in Chapter 2.

After removing all non-nucleolar beads from the simulation and replacing the hard nuclear wall with a harmonic energy well, we compared the *average nearby* statistic with the confinement parameter, giving the strength of the energy well. The use of the average nearby statistic was motivated by the observation that it is a representation of local density, and the strength of the confinement force controls the size of the space that the beads occupy. We observed that for any particular value of the confinement parameter, this statistic was comparatively lower for the fast crosslinking and higher for the slow crosslinking. Perfect agreement was not expected, including for the reason that the confinement force is isotropic and therefore the beads will be distributed approximately radially symmetrically, but the nucleolus in the original model took the form of a flattened body against the nucleolar wall. Furthermore, in the original model the equivalent of the confinement force was the combination of a hard boundary on one side and a soft boundary on the other. Combined with the fact that this statistic varies monotonically with the confinement parameter and there does not seem to be a significant qualitative effect on the dynamics, we were able to choose a value of 2×10^{-3} and continue with our reductions.

We next considered the replacement of the non-linear wormlike spring with a harmonic spring, due to the challenges of averaging out a force with a finite singularity. In doing so, we needed to select a spring constant that leads to a qualitatively similar force. Due to the non-linear increase in the force, for any particular value of the spring constant, the harmonic spring will be stronger at smaller extensions, but weaker at larger extensions, and we can understand the value of the spring constant in terms of the value at which it is equal in force to the original non-linear spring.

The most significant change that we made to the model was the manner in which the stochastic crosslinks form and break. While the model in (Hult et al. 2017) used normally-distributed waiting times with a standard deviation set to one-fifth of the mean, in this chapter we replaced all waiting time distributions with exponential distributions of the same mean, and further changed waiting times from being a bead-level active/inactive setting, where active beads would then essentially instantly bond with the nearest available other active bead, with bond-level waiting times, where unbound beads would have a certain rate at which they would form a bond with any particular other nearby unbound bead, which can be expressed as a CTMC process. The main change to be understood on an intuitive level is the fact that this means that beads will, after breaking a bond, reform a new bond more quickly if they are surrounded by more other beads. Consequently, faster crosslinking leads to the emergence of clusters, which leads to positive feedback as beads in these clusters then reform bonds even more quickly. This leads to an apparent case of larger clusters being able to form in this model than in the previous model. However, we have shown that our statistics lead to similar overall behavior and the qualitative reproduction of the behaviors of interest in the model, so it does not appear that this change is significant. There is additionally an argument to be made that the CTMC form of the crosslinking is a more physically fundamental formulation, and therefore should be automatically favored over the previous normal distribution approach.

With these changes made to the model, we are now ready to build the mathematical understanding of how to perform principled averaging on a system of this form, with pairwise crosslinking forces stochastically switching according to a CTMC process. In the next chapter, we will use the general model framework with a small number of beads and develop the theory and methods required to compute an effective energy landscape through the framework of quasipotentials.

CHAPTER 4: QUASIPOTENTIAL ANALYSIS

4.1 Introduction

Now that we have altered the model originally described in section 1.4 to the reduced form described in section 3.2, we will develop the tools required to understand transitions between minima in this type of model, and compute quasipotential barrier heights and corresponding asymptotic transition times.

As introduced in section 1.3, energy landscapes are a powerful tool for understanding stability and transitions in gradient systems. However, stochastic crosslinking of the form considered in the previous chapters produces a non-gradient system. However, many of the same analyses can still be performed by replacing the potential function that arises from a gradient force with a so-called quasipotential function. In this chapter, we will extend previous literature on computing quasipotentials to systems of a form that includes the model presented in Chapter 3.

Depending on the form of the system, a number of different approaches for deriving a quasipotential may be used, but in the case of stochastic switching between gradient forces, we can apply the WKB ansatz approach, which is demonstrated in (Newby 2014). This approach involves splitting the Fokker-Planck equation based on the state of the switching force and then assuming a WKB-form steady state solution (see details in section 4.2.2). Ultimately, we produce an equation of the form

$$H(x, \nabla W) = 0$$

where x is the collective system position vector, that implicitly defines the gradient of the quasipotential.

However, in contrast to previous work, we now consider a more general system in which particles are subject not only to a stochastically switching force, but also to noise through Brownian motion. Furthermore, we show how the Hamilton-Jacobi equation can be expressed in terms of a general force switching according to a continuous-time Markov chain. While previous work on finding quasipotentials for non-gradient systems generally looked at one- to three-dimensional problems (Zhou, Aliyu, Aurell & Huang 2012, Newby 2014, Brzeźniak, Cerrai & Freidlin 2015, Moore, Stieha, Nolting, Cameron & Abbott 2015, Bressloff &

Faugeras 2017, Newby, Schiller, Wessler, Edelstein, Forest & Lai 2017, Yang, Potter & Cameron 2019, Chen, Zhu & Liu 2019), our goals here will require application to higher-dimensional systems with multiple beads each moving in multiple dimensions. We thus will investigate how the addition of Brownian motion, the different form of switching, and the increase in dimensions affect our ability to compute quasipotentials for the system. In Chapter 3, we presented a new model section 3.2 whose form falls under the framework for which we can derive the Hamilton-Jacobi equation.

We will present an idealized model using three beads moving in two dimensions that is designed to exhibit a similar flexible clustering/mixing behavior as was observed in the intermediate crosslinking in Chapter 2, and then demonstrate that applying the methods standard in the literature to the equations we derived defining our quasipotential leads to numerical problems that were not previously reported. This numerical instability arises in both the implicit solver for the gradient of the quasipotential and the solver for the most-probable path. We find that Newton's method does not generally converge from the starting guesses we can make, and so we use a modified version of Newton's method with an additional fallback designed to ensure convergence. This also leads to high sensitivity in the Hessian matrix, which is used in the GMAM method for finding the most-probable path (Heymann & Vanden-Eijnden 2008, Newby 2014), and so we use instead the string method, which does not require the Hessian matrix. Still, we see significant high-frequency noise in the string method update, arising if the number of images in the string not small. We present a set of numerical methods, along with code, that overcomes all of these challenges, and demonstrate validity of the quasipotential by showing agreement between quasipotential barriers and Monte Carlo escape time asymptotics.

In section 4.2, we derive the form of the Hamiltonian for our problem and describe numerical methods we apply to solve for the gradient of the quasipotential and most-probable transition paths. In section 4.3 we validate the formulation of the Hamiltonian on a 1d problem in which the string descent is not necessary. In section 4.4 we then apply the full method to compute transition paths and asymptotic escape times for a system of three particles moving in two dimensions. Finally, in section 4.5 we review our contributions and note future directions of research.

4.2 Formulation

In this section we detail the steps required to compute the gradient of the quasipotential ∇W along transition paths.

In section 4.2.2 we will derive the Hamiltonian and the associated Hamilton-Jacobi equation that define the gradient of the quasipotential ∇W for our problem. In section 4.2.3 we will describe the deterministic dynamics which arise when the amount of noise in the system is taken to 0, and show that the quasipotential also vanishes at fixed points of the deterministic dynamics. In section 4.2.4, we will describe the algorithm that we use to address the additional numerical challenges in this version of the problem and solve the Hamiltonian-Jacobi equation for ∇W. In section 4.2.5 we describe the numerical procedure for finding most-probable transition paths which minimize the action, and in section 4.2.6 we describe how we compute asymptotic escape times using Monte Carlo methods to validate the quasipotential values.

4.2.1 Model Framework

Recall the general form of an overdamped gradient system under Brownian motion:

$$dX = v(X)dt + \sqrt{2k_BT}dW \tag{4.1}$$

. Inspired by our work in Chapters 2 and 3, we want to consider a framework in which the forces affecting each particle are subject to some form of random switching. We modify the force term $v(x)$ to switch between different states, with states indexed by random variable $s_t \in 1, 2, ..., n$. In preparation for the asymptotic analysis, we also introduce a parameter $\epsilon = k_BT$ to arrive at the general mathematical form of the equation for the dynamics

$$dX_t = v(X_t; s_t)dt + \sqrt{2\epsilon}dW \tag{4.2}$$

.

The configuration of the system is represented as the combination of a position $X_t \in \mathbb{R}^m$ and a switching state index s_t, for some spatial dimensionality m and number of possible switching states n.

Equation (4.2) can be thought of as an overdamped Langevin-style equation where the switching state $s_t \in \{1, ..., n\}$ affects the drift term. In order to perform the analysis we are aiming for, we will need to put some constraint on the evolution of the discrete random variable s_t. In Chapter 3, we showed that a modification to the crosslinking formulation to use exponential waiting times, where the formation rate depends on pairwise distance between beads, allows the evolution of the crosslinking state to be formulated as a continuous-time Markov chain (CTMC). We will assume that s_t follows a continuous-time Markov Chain (CTMC) process whose transition rate matrix $\frac{1}{\epsilon}S(X_t)$ depends on the position variable X_t. In this way the processes governing position and switching state are coupled. We multiply by the same factor $\epsilon = k_B T$ to couple the noise from Brownian motion and CTMC switching and prepare to take a distinguished limit of small noise.

4.2.2 Deriving the Hamilton-Jacobi Equation

Let the force (drift) on the i-th position coordinate under switching configuration s be represented as v_i^s, and these values also represented in the matrix $V_{si} = v_i^s$. Let the transition rate into state j from state k be represented as S_{jk}, such that S is a CTMC process transition matrix. Finally, let $p_s(x,t)$ represent the joint probability function between the discrete variable s and the continuous position variables of each bead,

$$\rho_s(x,t) = \rho(x,t|s_t = s)\mathrm{P}(s_t = s) \quad \text{for } s = 1, 2, \ldots n \tag{4.3}$$

where $\rho(x,t|s_t = s)$ is the conditional density for the process X at time t given that the force state is currently in state s.

Each individual momentum variable p_s will follow a Fokker-Planck equation associated with that state's drift term from eq. (4.2), with an additional coupling term to represent transitions between states. These coupled Fokker-Planck equations take the form

$$\frac{\partial p_s}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left[v_i^s p_s\right] + \epsilon \sum_i \frac{\partial^2}{\partial x_i^2}\left[p_s\right] + \frac{1}{\epsilon}\left[S\vec{p}\right]_s \quad \text{for } s = 1, 2, \ldots n \tag{4.4}$$

with the associated steady state equation,

$$0 = -\sum_i \frac{\partial}{\partial x_i}\left[v_i^s p_s\right] + \epsilon \sum_i \frac{\partial^2}{\partial x_i^2}\left[p_s\right] + \frac{1}{\epsilon}\left[S\vec{p}\right]_s \quad \text{for } s = 1, 2, \ldots n \tag{4.5}$$

If the drift term v did not depend on the switching process s_t, the Langevin process for X_t would have a potential function constructed from a path integral of v. However, in our formulation in which the drift function v exhibits random switching, it is no longer the gradient of a potential function, and so such a U

cannot be found. To allow for understanding of transitions between states of the system, we will develop in this work the theory behind the quasipotential for overdamped Langevin systems with stochastically switching drift, by applying an equivalent WKB ansatz as in (Newby 2014).

We now consider a small-noise asymptotic as $\epsilon \to 0$ in which the magnitude of the Brownian noise goes to 0 as the rate of the stochastic switching goes to infinity - in this limit, there is no diffusion and the forces v exist in a superposition of the v^s according to the steady state distribution of s. We form an asymptotic expansion as $\epsilon \to 0$ by making the assumption that the steady state distribution $p_s(x)$ follows a WKB ansatz of the form

$$p_s(x) = r_s(x) \exp\left(-\frac{1}{\epsilon} W(x)\right). \tag{4.6}$$

We see here that $W(x)$ takes the place of the potential $U(x)$ in 1.2; $W(x)$ is the quasipotential. We then plug eq. (4.6) into eq. (4.5), seeking equations for $r_s(x)$ and $W(x)$ given by the order $\frac{1}{\epsilon}$ terms, which are the lowest order in ϵ. Note the addition of a pre-exponential term r_s depending on the state s. In this way, it differs from a typical WKB expansion in which there would be no pre-exponential term present at lowest order. This term only encapsulates the relationship between states – the full pre-exponential term only emerges at higher order in ϵ.

The resulting order $\frac{1}{\epsilon}$ equation can be written in terms of a matrix M, which depends on the particle positions x and the gradient of the quasipotential ∇W, left-multiplying the vector of state probabilities $\vec{r} = (r_1, r_2, \ldots, r_n)^T$ as

$$M(x, \nabla W)\vec{r} = \vec{0} \tag{4.7}$$

Details are shown in section 4.2.2. The matrix M can be written as the sum of three matrices corresponding to the three terms in eq. (4.5): D, the diffusion matrix; A, the advection matrix; and S, the switching matrix:

$$M(x, \nabla W) = D(\nabla W) + A(x, \nabla W) + S(x) \tag{4.8}$$

Note that S is unchanged from its original definition as the CTMC transition rate matrix and serves the purpose of coupling the different states, and D and A are diagonal matrices representing the diffusion and drift respectively within each state of the switching force.

We can observe that in order to solve the system eq. (4.7) we must have $\det M(x, \nabla W) = 0$. Choosing the Hamiltonian as $H(x, p) = \max\lambda$ s.t. $M(x, p)v = \lambda v$, the greatest eigenvalue, we can see that the lowest order equation arising from WKB having a solution becomes equivalent to the Hamilton-Jacobi equation, $H(x, p) = 0$. Note that we have introduced a new variable p as the second argument, by analogy to the typical form of the Hamilton-Jacobi equation. This means that our solution for ∇W will be given by the value of p that solves the Hamilton-Jacobi equation.

Combining the above with an additional curl-zero constraint, we obtain a pair of equations that uniquely define the gradient of the quasipotential:

$$H(x, \nabla W(x)) = 0 \tag{4.9}$$

$$\nabla \times \nabla W(x) = 0 \tag{4.10}$$

In this work we will compute the quasipotential along most-probable paths (MPPs) parameterized as $\phi(s) : [0, 1] \to \mathbb{R}^m$, in which case we can replace the second equation with one evaluated only on the path:

$$\left.\frac{dH(x, p)}{dp}\right|_{p=\nabla W(x)} \parallel \frac{d\phi}{ds} \tag{4.11}$$

Add here: connection to gradient system, MPP everywhere parallel to the gradient, here, its everywhere parallel to the quasipotential. But this is only true at convergence, we don't know grad W until know the path, so cannot directly use things like the string method to find the path, and we need to know the path to find W. So we talk about these numerical issues next.

4.2.3 Deterministic Dynamics

The unknown MPP along which we seek to determine the quasipotential is "anchored" at its end points, where we can determine *a priori* the value of ∇W. These points act like boundary conditions and are the fixed points of the "deterministic dynamics," which are defined by the limit $\epsilon \to 0$ of eq. (4.2). Recall that in this limit, the Brownian noise term vanishes, and the CTMC switching rates go to infinity such that the system always exists in a superposition of states consistent with the steady state of the switching matrix $S(\vec{x})$

at its current position. The steady state dynamics can be expressed as

$$\begin{aligned}\frac{d\vec{x}}{dt} &= F_{\text{det}}(\vec{x}) \\ \vec{F}_{\text{det}}(\vec{x}) &= \vec{r}\,V(\vec{x}) \\ \vec{r} &= \text{null}S \qquad \sum_i r_i = 1\end{aligned} \tag{4.12}$$

where $V(\vec{x})$ is the matrix defining the force on each bead in each possible switching state and $\vec{r}$ is null-vector of $S(\vec{x})$ appropriately normalized so that it represents the steady-state distribution.

Points at which $\vec{F}_{\text{det}}(\vec{x}) = \vec{0}$ are referred to as fixed points of the deterministic dynamics. Such points have the property that $\nabla W = 0$, as shown in section 4.2.7.2. When computing the quasipotential along a most-probable path $\phi(s)$, if each end of the path is located at a fixed point of the deterministic dynamics, then we know that

$$\nabla W(\phi(0)) = \nabla W(\phi(1)) = 0 \tag{4.13}$$

This equation, combined with eqs. (4.9) and (4.11) will provide all of the information required for a numerical procedure to find most-probable paths and the quasipotential along them in a system.

4.2.4 Solving for ∇W

We will now describe the numerical optimization procedure by which we obtain ∇W, the gradient of the quasipotential, through which we obtain the final quasipotential by numerical integration. This will require formulating the simultaneous solution of eqs. (4.9) and (4.11) as an optimization problem using Lagrange multipliers and then plugging in the Hamiltonian constructed in section 4.2.2.

Our approach for computing ∇W is based on the Newton's method equations in (Newby 2014), which are presented for an arbitrary $\mathcal{H}$, and also require the gradient and Hessian of $\mathcal{H}$ with respect to the momentum variables p. Using the Hamiltonian derived in section 4.2.7.1 along with the differentiation formulas shown in section 4.2.7.3, we are able to apply these equations to our problem.

In practice, we have observed that simple application of these Newton's method equations often fails to converge for the problems we have attempted. To address this, we added a fallback scheme that is designed to slightly improve the degree to which eq. (4.11) is satisfied while maintaining eq. (4.9) to numerical tolerance. If the Newton's step fails to improve the quality of the solution, this fallback is used instead, which is guaranteed to produce a better solution. In practice, the fallback is able to quickly bring the guess close enough to the true solution for Newton's method to begin to converge quadratically. Additional details of this routine are written in section 4.2.8.1.

4.2.5 Computing Minimum Action Paths

Previous work on computing quasipotentials has either computed the quasipotential along MPPs, or on a grid using an upwind scheme (Newby 2014). As we are interested in transition asymptotics, we restrict this work to considering computing quasipotentials along MPPs, and so the routine for computing the quasipotential must simultaneously search for a MPP.

MPPs are computed by representing a path connecting two points x_a, x_b in the position space of the system using a path $\phi(s)$ such that $\phi(0) = x_a$, $\phi(1) = x_b$. We discretize the path ϕ as a sequence of "images" representing states along the transition path: $\phi_1, \phi_2, \phi_3, ..., \phi_N$. These images are chosen so that the arc length is constant between images, e.g. $||\phi_k - \phi_{k-1}|| =$ const.. To compute the MPP, we make an initial guess of the transition path (typically linearly interpolated) and then apply the string method (Weinan, Ren & Vanden-Eijnden 2002), in which each image along the path is moved a small distance opposite the direction of ∇W. Compared to typical applications of the string method, there is an additional step of applying the implicit solver to find the value of ∇W at each image, which itself depends on the current direction of the string.

Because we are looking for escape trajectories from a well, we allow the final image of the string to move according to the climbing string method (Ren & Vanden-Eijnden 2013). The final image moves in a different direction: letting ∇W_f refer to the gradient of the quasipotential at the final image, and $d\phi_f$ refer to the direction of the string at the final image, the direction is given by

$$\nabla W_f - (1 + \alpha)\text{proj}_{d\phi_f} \nabla W_f \tag{4.14}$$

This can be interpreted as descending in the directions orthogonal to $d\phi_f$, but climbing in the direction parallel to $d\phi_f$, as the projection term inverts that component.

Details of the algorithm are presented in section 4.2.8.2.

4.2.6 Monte Carlo Simulation

To validate our quasipotential, we compare it to Monte Carlo simulations of the system. These simulations are taken by numerically evaluating eq. (4.36) using a modified Euler-Maruyama method, to incorporate both the SDE and the stochastic switching. We track the current switching state s in addition to the position x, and update the position in each step as

$$x_{\text{new}} = x + \Delta t F(x; s) + \sqrt{2\epsilon}\Delta W \tag{4.15}$$

where $F(x; s)$ is the drift function at position x in state s, and $\Delta W \sim \mathcal{N}(0, \Delta t)$ is the increment of Brownian noise. Then, we compute any changes to the state variable s. The specifics of this update depend on the nature of the switching, but in general involve checking whether any changes occur within the timestep Δt by drawing waiting times from exponential distributions. Due to the memoryless property of exponential random variables, we may safely redraw in future time steps any waiting times that do not correspond to a transition within the current timestep, in case the transition rate changes between timesteps.

To empirically estimate the mean escape time associated with a particular ϵ, we initialize a Monte Carlo simulation in a state corresponding to a stable equilibrium of the deterministic dynamics, as obtained through descent of the deterministic force. We proceed to simulate eq. (4.15) until a termination condition is reached indicating escape from the basin of attraction that the simulation began in. The escape time is then recorded as τ_k. The simulation may also reach maximum time T without exiting the original basin of attraction, in which we take $\tau_k = T$. We assume escape times follow an exponential distribution with mean μ and use the maximum likelihood estimator (MLE) for the mean of an exponential distribution given samples capped at a maximum time , given by [1]

$$\hat{\mu} = \frac{\sum_k \tau_k}{\sum_k I\left(\tau_k < T\right)} \tag{4.16}$$

(Note that as $T \to \infty$, this reduces to the simple mean of the samples τ_k.

We then repeat the above process for a sequence of values of ϵ, computing a relationship $\hat{\mu}(\epsilon)$. We expect a linear relationship between $\frac{1}{\epsilon}$ and $\log \hat{\mu}$

$$\log \hat{\mu} = a + \frac{b}{\epsilon} \tag{4.17}$$

[1] `https://www.itl.nist.gov/div898/handbook/apr/section4/apr412.htm`

where the value b is corresponds theoretically with the quasipotential barrier height, and a would be given by the higher-order pre-exponential term.

4.2.7 Derivation Details

4.2.7.1 Derivation of the Hamiltonian

Here we will present the computation arising from plugging in the WKB ansatz eq. (4.6)

$$p_s(x) = r_s \exp\left(\frac{1}{\epsilon} W(x)\right)$$

to the steady state equation eq. (4.5)

$$0 = -\sum_i \frac{\partial}{\partial x_i}\left[v_i^s p_s\right] + \epsilon \sum_i \frac{\partial^2}{\partial x_i^2}\left[p_s\right] + \frac{1}{\epsilon}\left[S\vec{p}\right]_s$$

(both reproduced for clarity) and collecting lowest order terms in ϵ, which will be the $\frac{1}{\epsilon}$ term.

We will need to take spatial derivatives of the WKB ansatz:

$$\frac{\partial}{\partial x_i} \exp\left(\frac{1}{\epsilon} W(x)\right) = \exp\left(\frac{1}{\epsilon} W(x)\right)\left[-\frac{1}{\epsilon}\frac{\partial W}{\partial x_i}\right] \tag{4.18}$$

For the drift term, we obtain a $\frac{1}{\epsilon}$ term by differentiating once the exponential, thus giving us

$$-\sum_i \frac{\partial}{\partial x_i}\left[v_i^s p_s\right] \sim \frac{1}{\epsilon} r_s \exp\left(\frac{1}{\epsilon} W(x)\right) \sum_i v_i^s \frac{\partial W}{\partial x_i} \tag{4.19}$$

For the diffusion term, we obtain a $\frac{1}{\epsilon}$ term by differentiating the exponential twice and combining with the ϵ pre-factor, giving us

$$\epsilon \sum_i \frac{\partial^2}{\partial x_i^2}\left[p_s\right] \sim \frac{1}{\epsilon} r_s \exp\left(\frac{1}{\epsilon} W(x)\right) \sum_i \left(\frac{\partial W}{\partial x_i}\right)^2 \tag{4.20}$$

Finally, directly plugging in to the switching term, without discarding anything, produces

$$\frac{1}{\epsilon}\left[S\vec{p}\right]_s = \frac{1}{\epsilon} \exp\left(\frac{1}{\epsilon} W(x)\right)\left[S\vec{r}\right]_s$$

Thus, eq. (4.5) reduces to, at lowest order in ϵ,

$$0 = \frac{1}{\epsilon} r_s \exp\left(\frac{1}{\epsilon} W(x)\right) \sum_i v_i^s \frac{\partial W}{\partial x_i} + \frac{1}{\epsilon} r_s \exp\left(\frac{1}{\epsilon} W(x)\right) \sum_i \left(\frac{\partial W}{\partial x_i}\right)^2 + \frac{1}{\epsilon} \exp\left(\frac{1}{\epsilon} W(x)\right) [S\vec{r}]_s \quad (4.21)$$

We can cancel out the common term of $\frac{1}{\epsilon}\exp\left(\frac{1}{\epsilon}W(x)\right)$, and then rewrite the resulting equation as a matrix equation by viewing r_s as a vector and turning the advection (drift), diffusion, and switching terms into the matrices A, D, and S respectively:

$$[A + D + S]\,\vec{r} = M\vec{r} = \vec{0} \quad (4.22)$$

with the diffusion matrix D defined as

$$D = \left(\sum_i \left(\frac{\partial W}{\partial x_i}\right)^2\right) I \quad (4.23)$$

, the advection matrix A defined as

$$A = \text{diag}\left(V \nabla W\right) \quad (4.24)$$

(note: diag here indicates the mapping of a vector to the square matrix with it on the diagonal), and the matrix S unchanged from its original definition. Note also that D and A are diagonal matrices, so the only off-diagonal contributions come from S, which is not dependent on W.

4.2.7.2 Relationship between Deterministic Dynamics and Quasipotential

In this section we will demonstrate that fixed points of the deterministic dynamics correspond to points where $\nabla W = 0$.

Let x^d be a fixed point of the deterministic dynamics such that $F_{\text{det}}(x^d) = 0$, and let r^d be the probability vector across the states such that

$$\begin{aligned} S(x^d)r^d &= \vec{0} && \text{because } r^d \text{ is a nullvector} \\ r^d V(x^d) &= \vec{0} && \text{because } x^d \text{ is a fixed point} \end{aligned}$$

Recall the definition $M(x, \nabla W) = A(x, \nabla W) + D(\nabla W) + S(x)$. One can see from eqs. (4.23) and (4.24) that A, D vanish when $\nabla W = 0$, so that $M(x, 0) = S(x)$. It follows that $M(x, 0)r = S(x)r = 0$, so 0 is an eigenvalue of $M(x, 0)$ and therefore $H(x, 0) = 0$ for any x.

We apply eq. (4.25) to compute $\frac{dH}{dp_i}$, which we will show to be 0 for all i.

$$\frac{dH}{dp_i} = \frac{v_0^T dM u_0}{v_0^T u_0}$$

where u_0, v_0 are the right and left eigenvectors, respectively, of $M(x_d, 0) = S(x_d)$, corresponding to an eigenvalue $\lambda = 0$. Recalling that S is a CTMC transition rate matrix, the right eigenvector u_0 will naturally be the steady-state distribution r_d. Similarly, as a transition rate matrix its columns always sum to 0, and so the ones vector will be an eigenvector with eigenvalue 0, and thus $v_0 = [1, 1, 1 \ldots 1]^T$.

Noting that

$$\frac{dA}{dp_i} = \text{diag}(V_i)$$
$$\frac{dD}{dp_i} = 2p_i I = 0 \quad (p_i = 0)$$

we get that

$$dM(x_d, 0)u_0 = \text{diag}(V_i)r_d = v_0^T V_i^T r_d = \frac{d}{dt}x_d = 0$$

as x_d is a fixed point of the deterministic dynamics.

4.2.7.3 Derivatives of the Hamiltonian

In order to apply Newton's method for finding critical points, we will need to compute the gradient and Hessian of $H(x, p)$ with respect to the momentum variable p. To do so, we will use the formulas of (Magnus 1985) in differentiating eigenvalues of a matrix with respect to the entries of that matrix. Consider a real, square matrix $A = A_0 + dA$ such that $A_0 u_0 = \lambda_0 u_0$ and $v_0 A_0 = \lambda_0 v_0$, and then consider the function

$\lambda(dA)$ s.t. $\lambda(0) = \lambda_0$. Using the superscript + to refer to the Moore-Penrose inverse, we have

$$\begin{aligned} d\lambda &= \frac{v_0^T dA u_0}{v_0^T u_0} \\ d^2\lambda &= \frac{2v_0^T (dA) K_0 (\lambda_0 I - A_0)^+ K_0 (dA) u_0}{v_0^T u_0} \\ K_0 &= I - \frac{u_0 v_0^T}{v_0^T u_0} \end{aligned}$$

We now consider the case that A is in fact a function of the momentum variable $\vec{p}$, centered at a value $\vec{p}_0$. In this case we have $A(\vec{p}_0) = A_0$. Applying the chain rule, we get the formulas

$$\frac{d\lambda}{dp_i} = \frac{v_0^T A_{p_i} u_0}{v_0^T u_0} \tag{4.25}$$

$$\frac{d^2\lambda}{dp_i dp_j} = \frac{2v_0^T A_{p_i} K_0 (\lambda_0 I - A_0)^+ K_0 A_{p_j} u_0}{v_0^T u_0} + \delta_{ij} \frac{v_0^T A_{p_i, p_i} u_0}{v_0^T u_0} \tag{4.26}$$

Because $H(x, p)$ is simply the largest eigenvalue of the matrix M, we can compute the gradient and Hessian H_p and H_{pp} with direct application of eqs. (4.25) and (4.26).

4.2.8 Algorithm Implementation Details

4.2.8.1 Quasipotential Implicit Solver

The core Newton update from (Newby 2014) are given by the iterative update equations

$$\begin{aligned} p_{n+1} &= p_n + \mathcal{H}_{pp}^{-1} \left[\lambda_n \frac{d\phi}{ds} - \mathcal{H}_p \right] \\ \lambda_n &= \sqrt{\frac{\mathcal{H}_p \mathcal{H}_{pp}^{-1} \mathcal{H}_p - 2\mathcal{H}}{d\phi^T \mathcal{H}_{pp}^{-1} d\phi}} \end{aligned}$$

starting from an initial guess for p_0, where $\mathcal{H}, \mathcal{H}_p, \mathcal{H}_{pp}$ are evaluated at x, p_n, and the derivatives w.r.t p are computed shown in section 4.2.7.3. If the starting guess is sufficiently close to the correct value, p_n will converge to the value that satisfies eqs. (4.9) and (4.11), which is the value of ∇W. However, in practice, we find that for our problems the Newton method frequently fails to converge, and so we add a modification based on the structure of the optimization problem to guarantee convergence.

The solution to eq. (4.11) under the constraint eq. (4.9) can be equivalently posed as the unique solution of

$$\arg\max_p \left[p \cdot \frac{d\phi}{ds}\right] \qquad \text{s.t. } H(x,p) = 0 \tag{4.27}$$

as the surface $\mathcal{H}(x,p) = 0$ is convex. In other words, eq. (4.11) can be viewed as a maximization problem. Correspondingly, we require that each iteration increase the objective quantity $p \cdot \frac{d\phi}{ds}$. If the Newton's step fails to do so, we move a small distance κ in the direction of $\frac{d\phi}{ds}$ projected onto the normal to the $H(x,p) = 0$ surface:

$$p_n^* = p_n + \kappa \, \text{proj}_{\perp \mathcal{H}_p} \left(\frac{d\phi}{ds}\right)$$

We then use the root-finding version of Newton's method to find a p_{n+1} such that $H(x, p_{n+1}) = 0$, starting from p_n^*, until convergence:

$$p_n^* = p_n^* - \frac{H(x, p_n^*)}{H_p(x, p_n^*)} \tag{4.28}$$

If this did not increase the objective quantity $p \cdot \frac{d\phi}{ds}$, we reduce κ by a factor of 2 and try again. Because the surface is convex and we are moving in the objective direction projected onto the surface normal, we are guaranteed an improvement for sufficiently small κ. However, it is desirable that κ not be too much smaller than necessary to reduce the number of iterations until Newton's method begins to converge.

We find that with this modification we always reach a point where Newton's method begins to converge quadratically to the true solution.

4.2.8.2 String Method

We apply the climbing string method as described in (Ren & Vanden-Eijnden 2013). The only modification is the replacement of the gradient of the potential ∇U with the gradient of the quasipotential ∇W. These differ in that whereas ∇U evaluated at an image ϕ_i would depend only on the value of ϕ_i, ∇W depends on both ϕ_i and the direction of the string $\frac{d\phi}{ds}|_{\phi_i}$. We estimate this using a centered finite difference approximation,

$$\frac{d\phi}{ds}|_{\phi_i} \parallel \phi_{i+1} - \phi_{i-1} \tag{4.29}$$

Note that ∇W depends only on the direction, not the magnitude, so there is no need for the denominator of the finite difference formula. This however means that the update order affects the algorithm. In our case, we compute the directions $\frac{d\phi}{ds}$ at all images first, and then perform updates simultaneously.

4.3 1-Dimensional Case

To verify our formulation of the Hamiltonian (section 4.2.7.1) and correspondingly the quasipotential, we begin by applying our method to the one-dimensional case of a single bead subject to a constant excluded volume force and an on-off switching attractive force pulling it towards the origin. We note that the presence of only a single spatial dimension means the transition path is trivially known; this frees us from the additional optimization step of computing the most-probable transition path out of the minimum.

In this case, there are only two states of the switching force corresponding to whether the attractive force is switched off or on – we label these forces v^1 and v^2, respectively, in line with the notation introduced in section 4.2.2. They are given by

$$\begin{aligned} v^1(x) &= a_{ev} x \exp\left(-\frac{x^2}{c_{ev}}\right) \\ v^2(x) &= a_{ev} x \exp\left(-\frac{x^2}{c_{ev}}\right) - kx. \end{aligned}$$

In this example, we use parameters $k = 5$, $a_{ev} = 3$, $c_{ev} = 0.5$. We note that the method is independent to the choice of these parameters, and the choice is simply motivated to create a well with a basin of attraction extending to $|x| \approx 1$.

Correspondingly, the transition rate matrix takes the form

$$S = \begin{pmatrix} -a(x) & c \\ a(x) & -c \end{pmatrix} \tag{4.30}$$

with $a(x)$ being some decreasing function of distance from the origin , and c a constant that is here taken to be 0.5.

We consider three different choices of a, chosen to create quasipotential barriers of different heights, to allow us to show agreement in three different cases.

$$a_1(x) = 2e^{-3|x|^2} \tag{4.31}$$

$$a_2(x) = \frac{2}{1 + e^{20(|x|-0.75)}} \tag{4.32}$$

$$a_3(x) = \frac{4}{1 + e^{20(|x|-0.75)}} \tag{4.33}$$

Because the most-probable transition path can only move in the single dimension of the problem, we simply compute the quasipotential along some interval $(0, x_0)$ such that x_0 is to the right of the "saddle point" (which in one dimension is in fact a maximum). For the above choices, the maximum is in the vicinity of $x = 1$, so we compute values through $x_0 = 2$.

We simulate escape times as described in section 4.2.6, using the modified Euler-Maruyama method. When the particle is in the first (unbound) state, we sample an exponential waiting time from the binding rate

$$\tau_{\text{bind}} \sim \text{Exp}\left(\frac{\epsilon}{a(x)}\right) \tag{4.34}$$

and if $\tau < \Delta t$ the bead switches into the second (bound) state. Otherwise, the value of τ_{bind} is discarded. When the bond is formed, a second waiting time is drawn from the unbinding rate

$$\tau_{\text{break}} \sim \text{Exp}\left(\frac{\epsilon}{c}\right) \tag{4.35}$$

from which we compute the breaking time $t_{\text{break}} = t + \tau_{\text{break}}$. Thereafter, once $t \geq t_{\text{break}}$, the bead reverts back to the unbound state. Note that the binding rate depends on the changing variable x and so is resampled each timestep, whereas the unbinding rate c is constant and so the waiting time can be preserved until it is reached.

Figure 4.1(A) shows the results of computing the quasipotential for each of the three affinity functions, along with the deterministic energy computed by numerically integrating the deterministic force from section 4.2.3. This shows that while the two agree on the location of the highest point of the barrier, they disagree significantly on its height. Figure 4.1(B) shows the average Monte Carlo escape times for different values of ϵ, compared to lines whose slope is given by the height of the quasipotential barrier. The

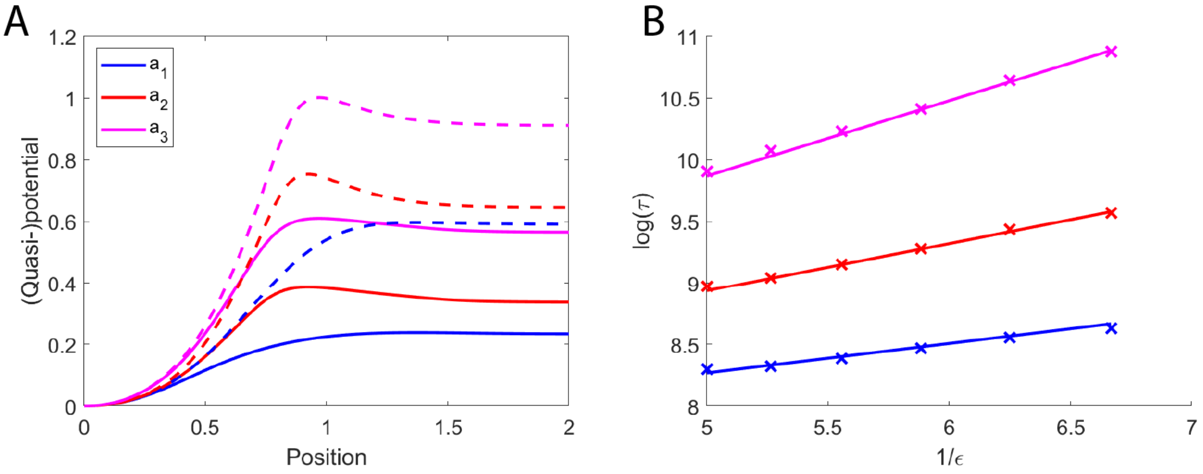

Figure 4.1: Comparison of Quasipotential and Escape Time Asymptotics for three different affinity functions. (A) Comparison of quasipotential along the string with the deterministic average, illustrating that the deterministic average would not agree with Monte Carlo statistics. (B) Average Monte Carlo escape times (points) showing linear behavior whose slope is predicted by the quasipotential string.

quasipotential barrier height clearly predicts well the slope followed by the escape times, confirming that our formulation of the quasipotential is consistent with the theory and preparing us to add the string descent element in higher-dimensional problems.

4.4 2-Dimensional Case

In this section, we detail the model of three beads moving in two dimensions that exhibits the same mixing behavior that originally inspired this work. We characterize the four stable states of the system, and show that the system spends most of its time in the state in which a pair of the beads are bound together. We compute a most-probable escape path from this state, and show that it lines up with asymptotic times for transitions between bound states.

4.4.1 Model

Our interest in stochastically switching Langevin systems is originally motivated by prior work on modeling dynamics of chromosomes inside a yeast nucleus (Walker et al. 2019). We consider here a toy model using the same functional forms of forces, while noting that the methods apply in general to any type of stochastically switching force between particles.

We consider three beads, affected by three forces: bonding, repulsion (excluded volume) and a global confinement force. These are collectively sufficient to produce multiple stable equilibria in the system.

These forces have the following forms:

$$\begin{aligned}
&\text{Confinement:} && f_{\text{c}}^i(x_i) = -\eta x_i \\
&\text{Excluded Volume:} && f_{\text{EV}}^i(\{x_j\}_j) = \sum_{j\neq i} a_{ev}\,(x_i - x_j)\exp\left(-\frac{(x_i-x_j)^2}{c_{ev}}\right) \\
&\text{Attraction:} && f_{\text{bond}}^i(\{x_j\}_j) = \sum_{j\neq i} kb_{ij}\,(x_j - x_i)
\end{aligned}$$

where $b_{ij} = 1$ if beads i and j are presently bonded and zero otherwise, and we use parameter values $k = 5$, $a_{ev} = 2$, $c_{ev} = 0.5$, $\eta = 1$.

This gives the following ODE for the position of bead i:

$$dX_i = \left(f_c^i + f_{\text{EV}}^i + f_{\text{bond}}^i\right) dt + \sqrt{2\epsilon} dW \tag{4.36}$$

where ϵ is a small positive parameter that controls the amount of stochasticity in the system, and also appears in the switching, as described below.

One can observe that the stochastic switching in the system occurs only in the bonding term - the other two forces are deterministic.

The stochastic switching of the bonding term models the crosslinking proteins that bind two nearby beads. Each bead can be either unbound, or bound to a single other bead. Bonds are symmetric. If a bond is formed between two beads, assume that the lifetime of the bond is an exponentially distributed random variable with rate c, meaning it has an expected lifetime of $\frac{1}{c}$. To simulate it, one can simply draw such an exponentially distributed random variable and use it as the lifetime.

It is natural to think that crosslinking proteins would be more likely to bind beads that are closer together. Therefore, by analogy to the form of eq. (4.30) we include an "affinity function" $a(r)$ dependent on pairwise distances for the binding rate. Specifically, $a(r)$ gives the (exponential process) rate at which a bond forms between two currently unbound beads i, j with positions $\vec{x}_i, \vec{x}_j$ that are separated by a distance $r = |\vec{x}_i - \vec{x}_j|$. We note that this rate is only meaningful until a bond forms – a rate of $a(r)$ is equivalent to stating that in an infinitesimally short time dt, there is a probability $\frac{dt}{a(r)}$ that a bond forms. However, note that as the system moves in time, these probabilities will change accordingly. In this section, we will use $a = a_2$, given in eq. (4.32), which is

$$a(x) = \frac{2}{1 + e^{20(|x|-0.75)}}.$$

Following the framework of section 4.2.1, the CTMC switches between the different binding configurations. These binding configurations corresponding to the states of the Markov chain are enumerated as: all three beads unbound ($s = 1$), bead 1 bound to 2 ($s = 2$), bead 1 bound to 3 ($s = 3$), bead 2 bound to 3 ($s = 4$). The corresponding transition rate matrix S takes the form

$$S = \begin{pmatrix} A & c & c & c \\ a(x_1 - x_2) & -c & 0 & 0 \\ a(x_1 - x_3) & 0 & -c & 0 \\ a(x_2 - x_3) & 0 & 0 & -c \end{pmatrix} \tag{4.37}$$

with $A = -a(x_1 - x_2) - a(x_1 - x_3) - a(x_2 - x_3)$. Recall again the fixed rate $c = 0.5$ describes bonds breaking.

To illustrate the qualitative behavior of this constructed model system, fig. 4.2 demonstrates a sample simulated trajectory. In particular, we can see that this model replicates the "mixing" property that motivated this research, with rapid switching between which pair of beads (analagous to a cluster) is currently bound. This motivates the further investigation into the behavior of this system, and the stability of the bound state.

4.4.2 Computing Most-Probable Escape Paths

We now proceed to demonstrating the process by which we compute a most-probable transition path out of a basin of attraction using the quasipotential climbing string method.

Compared to previous work on string descent, our problem exhibits a greater level of numerical instability. We begin by initializing a climbing string to search for an escape path out of the bound state minimum. Based on preliminary observations that emergence of high-frequency error in images along the string prevents convergence, we consider the most natural way to reduce such error, which is by reducing the number of images. In fig. 4.3, we plot the total change over all images since the previous iteration (A), and the height of the quasipotential barrier (B), over the course of iteration. We observe that while a string with 10 images

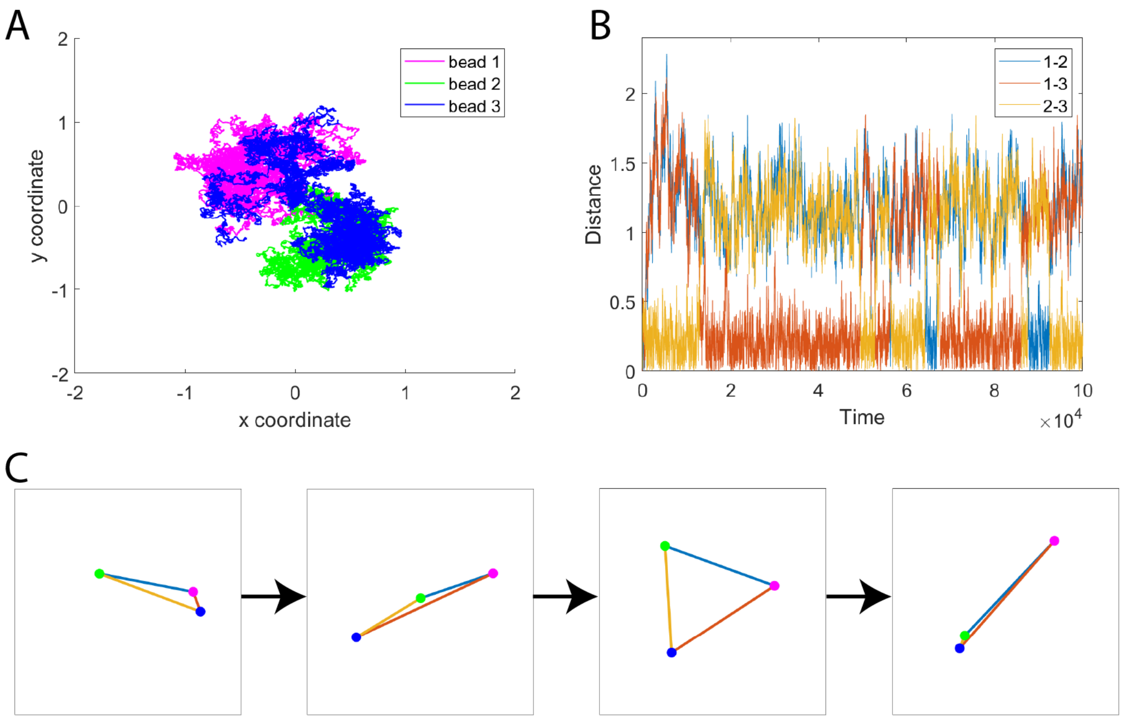

Figure 4.2: (A) Trajectories taken by three beads over the first 20000 timesteps of a simulation. (B) Plot of pairwise distances between the three beads over time. Observe that at any time, one distance is small (between the two currently bound beads) and the other two distances are large, with which pair is bound rapidly switching. (C) Illustration of a transition between two bound states. The bound beads separate, arranging into a line. The line then morphs into a triangle. Finally, two beads again approach and enter a bound state.

exhibits a decline in change indicating convergence, strings with more images do not, with the change instead settling into a high-frequency oscillation.

Based on these results, we will use our string with 10 images to determine the asymptotic transition time between bound states. This path is shown in additional detail fig. 4.4. Panel A shows the two-dimensional trajectory, which lies entirely along the y-axis, consistent with the observation in fig. 4.2(C,D,E) in which the transition passes through a line. Panel B shows the values of the coordinates along the string, demonstrating that they do not change in a direct linear fashion. With this string, we can now compute a quasipotential barrier and compare to the Monte Carlo asymptotics.

4.4.3 Monte Carlo Escape Statistics

To validate our quasipotential string descent method, we use it to predict escape time asymptotics from the bound state. As mentioned in section 4.4.1, our test system exhibits switching between various

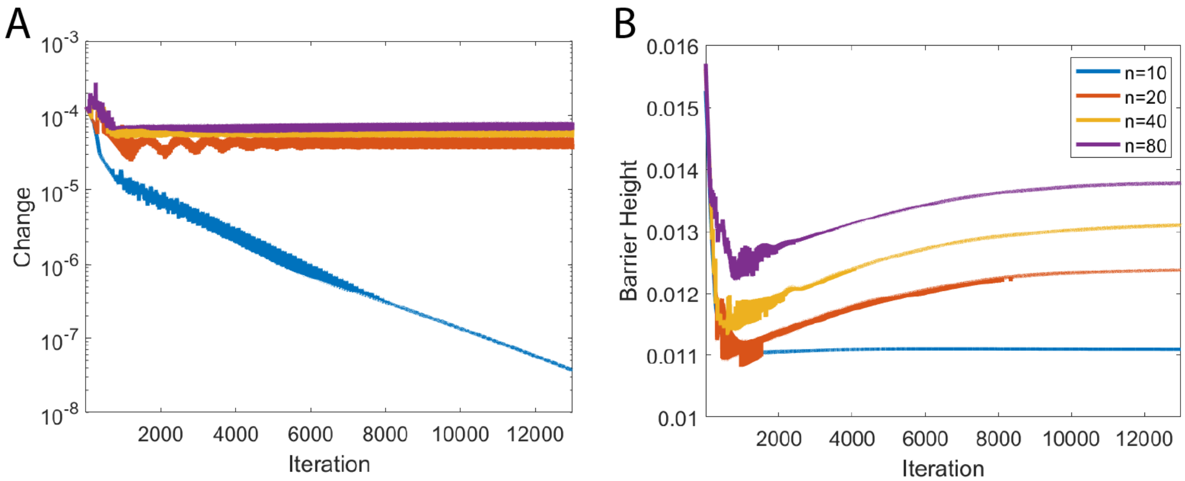

Figure 4.3: (A) Maximum change over all images from previous iteration, showing convergence only in the case of 10 images. (B) Quasipotential barrier height over iterations, showing that the $n = 10$ converges to a barrier height of approximately 0.011.

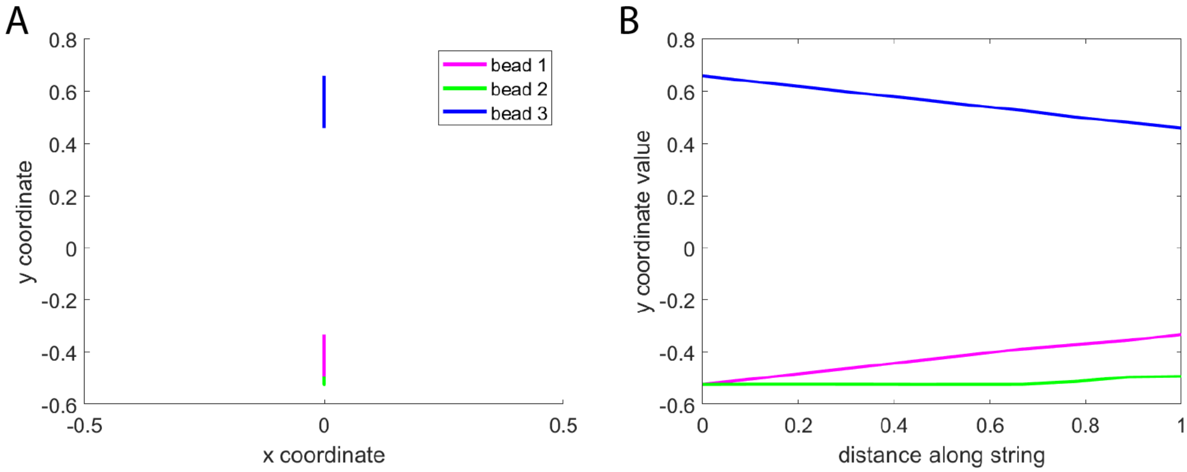

Figure 4.4: Visualization of escape path computed using quasipotential string descent, connecting bound state (fig. 4.2C) to saddle leading to line (fig. 4.2D). Note that this transition happens entirely on the y-axis. (A) Trajectory shown in 2 dimensions. (B) Plot of just the y-coordinate varying along the string. Note that the behavior is not simply linear, especially visible in the case of bead 2 (green).

permutations of the stable bound state, spending only minimal time in other stable configurations. For this reason, we compute escape times by initializing the simulation with beads 1 and 2 bound together, and measure time until a different pair of beads becomes bound, by a criterion that the newly bound beads should be separated by a distance under 0.3 and the original bound pair is separated by a distance of at least 1.

We run this simulations for a collection of values of $\epsilon \in [0.5, 0.35]$, and compute mean escape time $\hat{\mu}$ as described in section 4.2.6. In fig. 4.5(A), we see that for values of $1/\epsilon \geq 100$ ($\epsilon < 0.01$) the escape times are linear in the log-log plot, as predicted by the asymptotic relation, with a slope of approximately 0.118. We then compare this to the quasipotential barrier computed by integrating ∇W along the most-probable escape path computed in section 4.4.2, which is shown in fig. 4.5(B) to be approximately 0.11. This lines up quite closely with the value computed from Monte Carlo statistics. On the other hand, the value computed using the deterministic average, equal to approximately 0.96, does not.

4.5 Discussion

In this chapter, we extended the theory of quasipotentials to systems of Langevin equations with a force term that switches stochastically according to a continuous-time Markov process. We derived the Hamiltonian for this problem, and demonstrated the numerical problems that arise when using the standard methods in the literature on our problem. We developed modifications to these methods so that we are able to compute most-probable transition paths and quasipotential barriers along these paths. We also provide code available on Github to perform these computations on arbitrary systems within our framework.

We have observed that for this particular problem, computation of the quasipotential is much more sensitive than was reported in prior literature, and a large part of this paper is addressed to overcoming this numerical sensitivity. Through a combination of a backward Euler scheme for the string descent update and a low number of images along the transition path, our strings descended stably. However, in future work, we would like to develop improved methods that maintain stability on a higher number of images without excessively reducing the timestep.

Our idealized model contrasts with previous work in which the model randomly switches between deterministic differential equations (Newby 2014, Bressloff & Faugeras 2017), as our model includes noise both from random switching and from Brownian noise, and our switching takes the form of a general CTMC

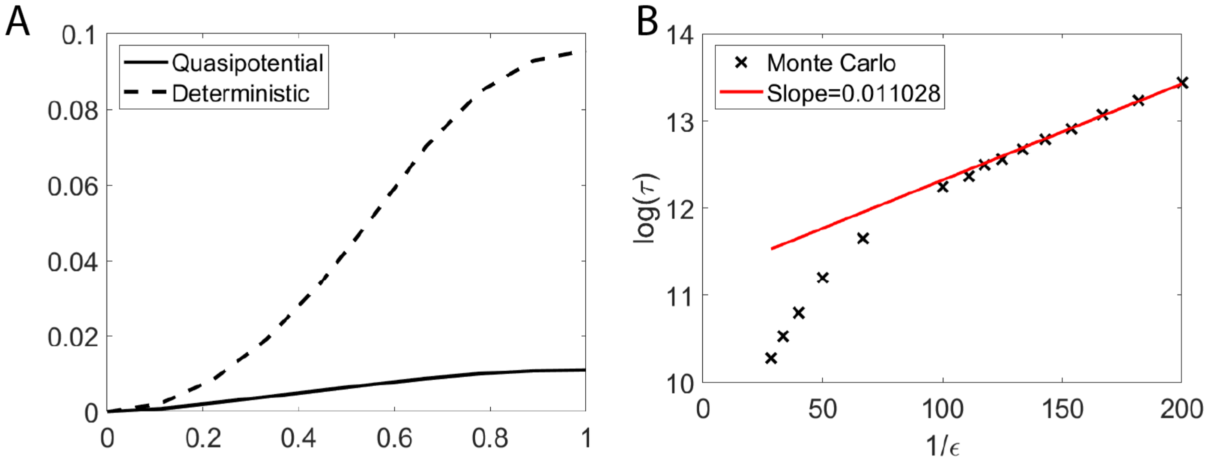

Figure 4.5: (A) Quasipotential (solid) and deterministic energy (dotted) along transition path from fig. 4.4. Quasipotential barrier height is approximately 0.011. (B) Comparison of asymptotic escape times computed via Monte Carlo simulation to slope taken from quasipotential barrier height.

process. We demonstrate that obtaining the effective energy landscape for such a system requires considering both forms of noise simultaneously, which can be accomplished through a WKB ansatz approach.

We found numerical issues arising in two separate places in our algorithm: the implicit solve step for ∇W, as discussed in section 4.2.4, and the optimization method for finding the most probable transition path, as discussed in section 4.4.2. In each case, we showed how to modify the algorithm to overcome these issues. As solving for ∇W is a convex optimization problem, it is expected that there should exist a good solver, and our modified Newton's method seems to fit this role, with guaranteed convergence under a short number of iterations from practical considerations. However, there is room for future work in further understanding the error in the MPP computation step, as the current remedy of reducing the number of images used leaves desired the ability to compute paths on a finer resolution. We reiterate that the most commonly observed alternative to the string method, the Geometric Minimum Action Method (GMAM), does not alleviate this problems as it runs into instability due to high sensitivity in the Hessian matrix. However, it is conceivable that continued work on the method used for this optimization step could produce a more stable algorithm in higher numbers of images.

Applying our methods to our idealized model, we showed how a stochastically switching pairwise crosslinking force could create effective bound states, with the system exhibiting switching between which pair of beads is in the effective bound state. We note that this is not simply switching between which pair of beads are currently bound by the switching crosslinking force - when beads are in this so-called effective bound state, they are switching back and forth between being bound and unbound, but on average are bound enough of the time to hold together until a rare event through a combination of Brownian noise and switching of the force allows them to separate and a new pair of beads to enter the effective bound state. We showed the ability to accurately predict the asymptotic timescales on which this escape occurs, showing strong agreement between the slope of Monte Carlo simulations of the timescale and the height of the quasipotential barrier, in line with the Arrhenius law of escape times. This validates our methods and code and sets the stage for use in further applications.

The WKB ansatz approach also allows for the computation of additional terms by following through the asymptotic expansion to higher order. In particular, the next term in the series would give a pre-exponential term that would allow for computation of the intercept of the escape time asymptotics as shown in figs. 4.1 and 4.5. While the current work focused on developing the machinery to compute most-probable transition paths under our framework, future work could expand on this and compute this pre-exponential coefficient.

CHAPTER 5: CONCLUSIONS

In this dissertation, we discussed a multi-faceted mathematical approach to better understanding the mechanisms by which stochastic pairwise crosslinking on a sufficiently fast timescale can lead to complex emergent structure. In this chapter, we will review the three objectives set out in Chapter 1, reviewing the contributions made to each, and then discussing implications along with ongoing and future work.

5.1 Goal One

We begin by considering the first goal presented in Chapter 1, to investigate the role of the dynamic timescale parameter μ in the formation of the emergent spatio-temporal clustering structure found in (Hult et al. 2017). In the original paper, simulations were performed for $\mu \in 0.09, 0.9, 90$ (seconds), a set of three values spanning four orders of magnitude of timescale. Observations of the results showed the presence of strong clustering in the $\mu = 0.09$ case, but not in the longer timescales. This suggests the presence of a "phase transition" in the system, where it switches from the clustering behavior to an essentially homogeneous distribution of beads over the nucleolus.

In Chapter 2, we showed the results of performing a series of simulations at 20 values of μ spaced over the same interval $[0.09, 90]$, with 10 points spaced logarithmically over $[0.09, 0.9]$ and 10 spaced logarithmically over $(0.09, 90]$. The original expectation was that this would show in greater detail the dissolution of clusters as the system switched from the clustering to the non-clustering regime. Indeed, we observed that as μ increased, the clusters because less rigid, with beads more and more frequently observed to enter and leave clusters, and the apparent size of the clusters as determined by inspection decreasing accordingly. As μ continues to increase, by the value of $\mu \approx 1$ as shown in (Hult et al. 2017), there are no longer clusters clearly visible to the human eye.

However, we sought more detailed insight into the nature of this behavior and this transition. To this end, and inspired by analogy to the notion of "interaction" in Hi-C data, where sections of DNA that are sufficiently nearby (at the time of processing) are counted as an interaction, we defined interactions as times when beads were within a set interaction distance. In this case, we chose a value of $100nm$, although our core results did not appear qualitatively sensitive to variation in this parameter. Having introduced this notion of interaction, we then computed and plotted them across our range of μ values, and in doing so revealed that in addition to the previously known behavior in which the clusters dissolved as μ increased, measures of the mixing in the system showed a maximum value in the intermediate range $\mu \in [0.2, 0.8]$, as the clusters dissolved.

Based on these observations, we identified three characteristic values of μ representing these three different clustering behaviors – the fast crosslinking/rigid clustering $\mu = 0.09$, the intermediate crosslinking/flexible clustering $\mu = 0.19$, and the slow crosslinking/non-clustering $\mu = 1.6$. The choice of the fast and slow values are intended to sandwich the transition as narrowly as reasonably possible. The choice of $\mu = 0.19$ for the intermediate regime is on the fast side of the intermediate regime; this is motivated by the observation that the mixing rate has completed its sharp rise by this value, but by keeping μ as small as possible, the ability to see the clusters with the naked eye in image and video representations of the simulation output is preserved.

We were overall able to make a series of additional statements about the behavior and the model, and its dependence on the timescale parameter μ, consistent with our first goal. We showed in fine detail the transition in clustering that occurs over roughly $\mu \in [0.09, 1]$, most visible by looking at the local density measured in our average nearby statistic. However, beyond simply showing this transition, we identified that in an intermediate regime of approximately $\mu \in [0.2, 0.8]$, we see other statistics that describe overall rates of mixing in the beads maximized. Mixing is a property of known biological importance, especially for nuclear processes such as homology search. Consequently, this discovery indicates that the switching on/off crosslinking behavior is, when performed on a specific timescale, able to contribute to a second function.

5.2 Goal Two

We then moved on to our second goal, algorithmically identifying spatio-temporal clusters in our data. While we initially had only intended to extract the rigid clusters that were shown in (Hult et al. 2017), we now have the additional task of identifying clusters in the intermediate timescale, flexible clustering regime, which we have shown to be of significant interest due to the maximization of bead-bead cross-communication.

Again inspired by similarity to the methods used in analysis of Hi-C data, we mapped our data to a similarity matrix using an exponential weighting function based on the pairwise spatial distances, averaged over short time windows. This created a temporal multi-layer network, with one layer for each such time window.

We incorporated the CHAMP algorithm of (Weir et al. 2017) which builds a map of how the best partition identified by modularity optimization changes with respect to the resolution parameter γ and linking parameter ω. This allows qualitative identification of regions of greater stability in the parameter space, which reveals the parameter choices corresponding to robust output of the algorithm. We note that we performed all analyses for all three regimes, including the slow crosslinking regime in which we were confident of the absence of any meaningful clustering to be detected by our community detection algorithm, in order to check whether the results were consistent with this expectation.

We then performed analysis of cluster lifetimes and persistence, two measures that reflect the stability of the clusters. We already knew from inspection that the clusters were significantly more temporally stable in the fast crosslinking than in the intermediate crosslinking. However, this analysis allowed us to describe this mathematically, and we saw that in the intermediate case the communities exhibited a lower persistence, indicating that beads were more likely to leave the community in a given timestep, and overall shorter lifetimes, indicating a shorter duration between a significant enough change happened to the community that the modularity optimization algorithm no longer found it advantageous to maintain that community into the next timestep at all.

Furthermore, we performed this analysis relative to the size of the cluster in question. This allowed us to observe a trend that larger clusters were more stable, across both the fast and intermediate crosslinking. Interestingly, this does not appear to remain to a significant degree in the slow crosslinking regime, in which persistence is essentially even across all but the smallest of communities, which further reinforces our belief that the communities we have found in the intermediate crosslinking are significant whereas those in the slow crosslinking are not. In the case of the fast crosslinking, this analysis is not as informative because almost all of the clusters are both large and extremely stable, so there is not much of a relationship to be observed. However, in the case of the intermediate crosslinking, we can see a clear trend that larger clusters are temporally more stable. This suggests that the larger clusters detected by the algorithm likely correspond to the clusters that are visible when looking at videos of the data, whereas many of the smaller clusters are likely small associations of beads in the area between the larger clusters.

One important observation we made in this analysis was the fact that in the flexible clustering regime where the clustering was often nontrivial to detect, the multi-layer modularity could detect temporal clusters that were not clearly detectable to the single-layer algorithm. This makes sense, as beads in a temporal cluster should be close together not just at one point in time, but over an extended period of time, and so factoring in this extra information can allow for better discrimination of community structure. This is also an important consideration in light of the earlier comments made, that the act of time- or population-averaging the data obscures structure that is present when looking instantaneously, as it shows that looking only at instantaneous data can also obscure structure visible over time. The multi-layer modularity algorithm, however, does not average over time, such as is done in the time averages shown in fig. 2.5, but rather incorporates the temporal information. This indicates that future work using polymer models to reveal instantaneous behavior that typical experimental observations should similarly take care not to miss structure that is visible when using temporal data even if it is not clearly detectable at any single instant in time.

5.3 Goal Three

While a variety of approaches exist for computing quasipotentials for non-gradient systems, depending on the specific nature of the non-gradient system, the particular approach most natural for a system that exhibits stochastic switching between gradient forces is the WKB ansatz approach, as described in Chapter 4. While the specific aspects of combining stochastic switching with Brownian noise, or the general CTMC form in which we considered the switching, were not reflected in the literature, we were able to derive the Hamilton-Jacobi equation for our system by fairly straightforward application of established methods. However, we found a series of new challenges arose when trying to apply these methods to numerically compute quasipotentials and transition paths for an idealized system of three beads in two dimensions that was designed to exhibit switching behavior reminiscent of the results from Chapter 2.

Recall that our investigation into mathematical averaging of stochastic crosslinking was motivated by the observation of beads associating into clusters as though on an effective energy landscape, even though the stochastically switching forces meant the system was never truly in equilibrium. We would like to apply the methods presented here in future work to analyze the clustering states observed in this model as minima of the deterministic dynamics, and then compute the quasipotential barrier to escape from these minima and use this barrier size to understand stability of the clustering state in terms of system parameters.

One of the challenges facing scaling up this method to more particles is the size of the matrix M, which increases factorially with the number of particles. Our previous work in (Walker et al. 2019) showed a collection of approximately 380 beads condensing into dynamic clusters of approximately 5-10 beads in suitable parameter regimes. In future work, we would like to apply our methods to larger numbers of beads using forces directly taken from biological modeling. The largest challenge to this will be addressing the state space of the switching process, which increases exponentially with the number of beads. The algorithm does not use the matrix $M(x, p)$ in full, only the derived (scalar) Hamiltonian $H(x, p)$ arising from the greatest eigenvalue. Since x and p only scale up in dimension linearly with number of particles, this suggests an alternative approach in which we construct an approximation either the map $(x, p) \to H(x, p)$, which is used to implicitly solve for p, or even directly the map $x \to p$ (noting that the direction $\frac{d\phi}{ds}$ of the string would also need to be included). Recently machine learning methods have been used to develop approximations to challenging-to-compute mathematical functions, and so this is an approach we would like to incorporate into our method in future work. Of particular interest would be methods that leverage the nature of the problem in which the gradient of the quasipotential at a particular bead depend primarily on the other nearby beads. This would encourage an approach using graph-structured machine learning data, such as the graph convolutional networks (Kipf & Welling 2016) or the more sophisticated graph attention network (Veličković, Cucurull, Casanova, Romero, Lio & Bengio 2017).

5.4 Ongoing Work

Another area in which the work described here could be extended to further applications would be through consideration of other random mechanisms besides pairwise crosslinking. In particular, recent experiments (Ganji, Shaltiel, Bisht, Kim, Kalichava, Haering & Dekker 2018, Terakawa, Bisht, Eeftens, Dekker, Haering & Greene 2017) and polymer modeling results (He, Lawrimore, Cook, Van Gorder, De Larimat, Adalsteinsson, Forest & Bloom 2020) have shown the presence of Z-loops in DNA. (He et al. 2020), which represented a plasmid of 10kbp of DNA in yeast using approximately 380 beads, showed in particular that a stochastic tension-dependent loop extrusion mechanism, hypothesized to be performed by condensin, could lead to the emergence of such Z-loops.

Ongoing work in the Forest/Bloom lab is building off of the model used in (He et al. 2020) to develop a flexible fine-resolution model of yeast chromatin with a particular focus on further understanding the development of Z-loops from a mechanistic perspective. This model differs significantly in scale from the model used in Chapter 2, in which a single bead represented approximately 5kbp of DNA. As a result, the spatio-temporal dynamics of the DNA are being investigated at a much different scale. Due to this much smaller scale, the dynamics affecting the polymer model are also different – the model of protein actions in (He et al. 2020) added on top of the DNA a model of histone loops, a critical aspect of the chromatin structure, and a tension-based loop extrusion model of the SMC protein condensin. Compared to the model from (Hult et al. 2017, Walker et al. 2019) used in Chapter 2, this smaller scale also exhibits emergent behavior from stochastically switching forces, but due to the different scale and different types of protein actions modeled, the type of emergent behavior of interest also differs. While clusters also emerge in this model, the Z-loop is added to the list of structures to study. However, as this is ultimately emergent structure from random pairwise actions, much of the work presented in this dissertation will be applicable in research on this model.

The primary objective of ongoing work is to extend this model from its original specific focus on a 10kb plasmid ring structure to account for more general geometries and scales, along with allowing for flexible configuration of simulation dynamics, to produce a scientific modeling tool that allows for a deeper investigation of the predictions that polymer models make of the behavior of the chromatin under these protein actions, especially relating to Z-loops. Inspired by the ring-based model of cohesin from (Lawrimore, Aicher, Hahn, Fulp, Kompa, Vicci, Falvo, Taylor & Bloom 2016), we have added cohesin rings to this model. Cohesins are modeled as rings of 9 beads connected by the same springs as the DNA beads. These rings are placed around the chromatin, and are then free to move under the effects of Brownian motion, with the excluded volume force keeping the rings around the chromatin. Nearby cohesin rings can form dynamic crosslinks, in a similar manner to the crosslinks in Chapter 2.

An interesting aspect of the methods involves how to identify Z-loops in the data. Given the full access to spatio-temporal position data that is obtained from simulations, we need to ask how to define a Z-loop, and how to define automatic tools for identifying such loops in the data. In a sense, this can be thought of as a generalization of the community detection work done in Chapter 2 – now, instead of looking for clusters of beads, we are also interested in a different type of structure. However, the lack of prior study on Z-loops means that there is not a clearly defined standard for what exactly a Z-loop should look like. A simple approach would be to look at just the connections of the polymers representing the chromatin combined with the dynamic condensin links, and identify structures where these links form a Z structure. However, a key takeaway of Chapter 2 was that spatio-temporal structure can emerge in these models in a way that is not easily detectable from a single snapshot in time. An open question would be whether temporal information can be incorporated into the search for Z-loops in data in a way that improves detection.

Compared to the form of the model used in (He et al. 2020), we also perform the same conversion by re-expressing the forces using exponential waiting times so that the stochastic switching can be viewed as an instance of a continuous-time Markov chain, in the same way as was shown in Chapter 3. Consequently, the work presented in Chapter 4 could also be applied to this system, or a version simplified for analysis. Under modeling choices from (He et al. 2020), the loop extrusion could be modeled in terms of considering a single loop extrusion protein connecting pairs of beads, meaning that the number of possible stochastic switching states for N beads would be given by N choose 2 (which is quadratic in N) instead of the factorial dependence in the model considered in Chapter 4. This means that the problem of explosion in the dimension of the matrix M that appears in the Hamilton-Jacobi equation could be avoided.

In future work, we could apply the Hamilton-Jacobi derivation from Chapter 4 along with the algorithms presented to investigate Z-loops from the perspective of favored states in an effective energy landscape. One of the most attractive aspects of this avenue is that one of the significant challenges with the WKB average approach that we developed in Chapter 4 is the size of the state space; however, the condensin loop extrusion model exhibits a significantly lower number of possible switching states compared to the pairwise crosslinking model. Thus, we should be able to apply these methods to a sufficient number of beads to produce Z-loops without needing to develop new methods for the sake of computational feasibility. Of course, as we showed in Chapter 4, there remains the possibility for additional numerical challenges to arise as the number of dimensions considered increases. Nevertheless, this remains a promising avenue for analysis, and the extent of attention given to Z-loops in recent literature further necessitate adding a mathematically principled averaging approach to the understanding of the mechanism by which Z-loops arise.

BIBLIOGRAPHY

Albert, Benjamin, Christine Colleran, Isabelle Leger-Silvestre, Axel B Berger, Christophe Dez, Christophe Normand, Jorge Perez-Fernandez, Brian McStay & Olivier Gadal. 2013. "Structure-function analysis of Hmo1 unveils an ancestral organization of HMG-Box factors involved in ribosomal DNA transcription from yeast to human." *Nucleic Acids Research* 41(22):10135–10149.

Albert, Benjamin, Julien Mathon, Ashutosh Shukla, Hicham Saad, Christophe Normand, Isabelle Léger-Silvestre, David Villa, Alain Kamgoue, Julien Mozziconacci, Hua Wong et al. 2013. "Systematic characterization of the conformation and dynamics of budding yeast chromosome XII." *J Cell Biol* 202(2):201–210.

Bassett, Danielle S, Mason A Porter, Nicholas F Wymbs, Scott T Grafton, Jean M Carlson & Peter J Mucha. 2013. "Robust detection of dynamic community structure in networks." *Chaos: An Interdisciplinary Journal of Nonlinear Science* 23(1):013142.

Belmont, Andrew S. 2006. "Mitotic chromosome structure and condensation." *Current Opinion in Cell Biology* 18(6):632–638.

Belmont, Andrew S & Aaron F Straight. 1998. "In vivo visualization of chromosomes using lac operator-repressor binding." *Trends in Cell Biology* 8(3):121–124.

Belmont, Andrew S, Gang Li, Gail Sudlow & Carmen Robinett. 1999. Visualization of large-scale chromatin structure and dynamics using the lac operator/lac repressor reporter system. In *Methods in Cell Biology*. Vol. 58 Elsevier pp. 203–222.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte & Etienne Lefebvre. 2008. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

Brackley, Chris A, Benno Liebchen, Davide Michieletto, Francois Mouvet, Peter R Cook & Davide Marenduzzo. 2017. "Ephemeral protein binding to DNA shapes stable nuclear bodies and chromatin domains." *Biophysical Journal* 112(6):1085–1093.

Brandes, Ulrik, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski & Dorothea Wagner. 2008. "On modularity clustering." *IEEE Transactions on Knowledge and Data Engineering* 20(2):172–188.

Bressloff, Paul C & Jay M Newby. 2014. "Path integrals and large deviations in stochastic hybrid systems." *Physical Review E* 89(4):042701.

Bressloff, Paul C & Olivier Faugeras. 2017. "On the Hamiltonian structure of large deviations in stochastic hybrid systems." *Journal of Statistical Mechanics: Theory and Experiment* 2017(3):033206.

Brzeźniak, Z, Sandra Cerrai & Mark Freidlin. 2015. "Quasipotential and exit time for 2D Stochastic Navier-Stokes equations driven by space time white noise." *Probability Theory and Related Fields* 162(3):739–793.

Cabreros, Irineo, Emmanuel Abbe & Aristotelis Tsirigos. 2016. Detecting community structures in hi-c genomic data. In *Information Science and Systems (CISS), 2016 Annual Conference on*. IEEE pp. 584–589.

Cameron, MK. 2012. "Finding the quasipotential for nongradient SDEs." *Physica D: Nonlinear Phenomena* 241(18):1532–1550.

Cao, Xue-Zheng & M Gregory Forest. 2019. "Rheological tuning of entangled polymer networks by transient cross-links." *The Journal of Physical Chemistry B* 123(5):974–982.

Chen, Jie, Alfred O Hero III & Indika Rajapakse. 2016. "Spectral identification of topological domains." *Bioinformatics* 32(14):2151–2158.

Chen, Zhen, Jinjie Zhu & Xianbin Liu. 2019. "Non-differentiability of quasi-potential and non-smooth dynamics of optimal paths in the stochastic Morris–Lecar model: Type I and II excitability." *Nonlinear Dynamics* 96(4):2293–2305.

Cheng, Tammy MK, Sebastian Heeger, Raphaël AG Chaleil, Nik Matthews, Aengus Stewart, Jon Wright, Carmay Lim, Paul A Bates & Frank Uhlmann. 2015. "A simple biophysical model emulates budding yeast chromosome condensation." *Elife* 4:e05565.

De Domenico, Manlio, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A Porter, Sergio Gómez & Alex Arenas. 2013. "Mathematical formulation of multilayer networks." *Physical Review X* 3(4):041022.

Dekker, Job. 2008. "Mapping in Vivo Chromatin Interactions in Yeast Suggests an Extended Chromatin Fiber with Regional Variation in Compaction." *The Journal of Biological Chemistry* 283(50):34532–34540.

Dekker, Job, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'shea, Peter J Park, Bing Ren et al. 2017. "The 4D nucleome project." *Nature* 549(7671):219.

Dekker, Job, Karsten Rippe, Martijn Dekker & Nancy Kleckner. 2002. "Capturing chromosome conformation." *Science* 295(5558):1306–1311.

Dekker, Job & Leonid Mirny. 2016. "The 3D Genome as Moderator of Chromosomal Communication." *Cell* 164(6):1110–1121.

Dekker, Job, Marc A Marti-Renom & Leonid Mirny. 2013. "Exploring the Three-Dimensional Organization of Genomes: Interpreting Chromatin Interaction Data." *Nature Reviews Genetics* 14(6):390–403.

Di Pierro, Michele, Bin Zhang, Erez Lieberman Aiden, Peter G Wolynes & José N Onuchic. 2016. "Transferable model for chromosome architecture." *Proceedings of the National Academy of Sciences* 113(43):12168–12173.

Duan, Zhijun, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C Anthony Blau & William S Noble. 2010. "A three-dimensional model of the yeast genome." *Nature* 465(7296):363.

Duch, Jordi & Alex Arenas. 2005. "Community detection in complex networks using extremal optimization." *Physical Review E* 72(2):027104.

Ea, Vuthy, Marie-Odile Baudement, Annick Lesne & Thierry Forné. 2015. "Contribution of topological domains and loop formation to 3D chromatin organization." *Genes* 6(3):734–750.

Eser, Umut, Devon Chandler-Brown, Ferhat Ay, Aaron F Straight, Zhijun Duan, William S Noble & Jan M Skotheim. 2017. "Topologically associating genomic domains in yeast." *Proceedings of the National Academy of Sciences* 114(15):3061–3070.

Feric, Marina, Nilesh Vaidya, Tyler S Harmon, Diana M Mitrea, Lian Zhu, Tiffany M Richardson, Richard W Kriwacki, Rohit V Pappu & Clifford P Brangwynne. 2016. "Coexisting liquid phases underlie nucleolar subcompartments." *Cell* 165(7):1686–1697.

Fisher, JK, M Ballenger, ET O'Brien, J Haase, R Superfine & K Bloom. 2009. "DNA relaxation dynamics as a probe for the intracellular environment." *Proceedings of the National Academy of Sciences* 106(23):9250–9255.

Fortunato, Santo. 2010. "Community detection in graphs." *Physics Reports* 486(3-5):75–174.

Fortunato, Santo & Darko Hric. 2016. "Community detection in networks: A user guide." *Physics Reports* 659:1–44.

Fortunato, Santo & Marc Barthelemy. 2007. "Resolution limit in community detection." *Proceedings of the National Academy of Sciences* 104(1):36–41.

Fudenberg, Geoffrey & Leonid A Mirny. 2012. "Higher-order chromatin structure: bridging physics and biology." *Current Opinion in Genetics & Development* 22(2):115–124.

Fudenberg, Geoffrey, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur & Leonid A Mirny. 2016. "Formation of chromosomal domains by loop extrusion." *Cell Reports* 15(9):2038–2049.

Ganji, Mahipal, Indra A Shaltiel, Shveta Bisht, Eugene Kim, Ana Kalichava, Christian H Haering & Cees Dekker. 2018. "Real-time imaging of DNA loop extrusion by condensin." *Science* 360(6384):102–105.

Ghosh, Surya K & Daniel Jost. 2018. "How epigenome drives chromatin folding and dynamics, insights from efficient coarse-grained models of chromosomes." *PLoS Computational Biology* 14(5):e1006159.

Girvan, Michelle & Mark EJ Newman. 2002. "Community structure in social and biological networks." *Proceedings of the National Academy of Sciences* 99(12):7821–7826.

Good, Benjamin H, Yves-Alexandre de Montjoye & Aaron Clauset. 2010. "Performance of modularity maximization in practical contexts." *Physical Review E* 81(4):046106.

Granell, Clara, Sergio Gómez & Alex Arenas. 2011. "Mesoscopic analysis of networks: Applications to exploratory analysis and data clustering." *Chaos: An Interdisciplinary Journal of Nonlinear Science* 21(1):016102.

Guimera, Roger, Marta Sales-Pardo & Luís A Nunes Amaral. 2004. "Modularity from fluctuations in random graphs and complex networks." *Physical Review E* 70(2):025101.

Hajjoul, Houssam, Julien Mathon, Hubert Ranchon, Isabelle Goiffon, Julien Mozziconacci, Benjamin Albert, Pascal Carrivain, Jean-Marc Victor, Olivier Gadal, Kerstin Bystricky & Aurélien Bancaud. 2013. "High-throughput chromatin motion tracking in living yeast reveals the flexibility of the fiber throughout the genome." *Genome Research* 23(11):1829–1838.

Han, Qiuyi, Kevin Xu & Edoardo Airoldi. 2015. Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning*. pp. 1511–1520.

He, Yunyan, Josh Lawrimore, Diana Cook, Elizabeth Erin Van Gorder, Solenn Claire De Larimat, David Adalsteinsson, M Gregory Forest & Kerry Bloom. 2020. "Statistical mechanics of chromosomes: in vivo and in silico approaches reveal high-level organization and structure arise exclusively through mechanical feedback between loop extruders and chromatin substrate properties." *Nucleic Acids Research* 48(20):11284–11303.

Henkelman, Graeme, Blas P Uberuaga & Hannes Jónsson. 2000. "A climbing image nudged elastic band method for finding saddle points and minimum energy paths." *The Journal of Chemical Physics* 113(22):9901–9904.

Henkelman, Graeme & Hannes Jónsson. 1999. "A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives." *The Journal of Chemical Physics* 111(15):7010–7022.

Heymann, Matthias & Eric Vanden-Eijnden. 2008. "The geometric minimum action method: A least action principle on the space of curves." *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 61(8):1052–1117.

Hult, Caitlin, David Adalsteinsson, Paula A Vasquez, Josh Lawrimore, Maggie Bennett, Alyssa York, Diana Cook, Elaine Yeh, Mark Gregory Forest & Kerry Bloom. 2017. "Enrichment of dynamic chromosomal crosslinks drive phase separation of the nucleolus." *Nucleic Acids Research* 45(19):11159–11173.

Imakaev, Maxim, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker & Leonid A Mirny. 2012. "Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization." *Nature Methods* 9(10):999–1003.

Jensen, Melanie A, Ying-Ying Wang, Samuel K Lai, M Gregory Forest & Scott A McKinley. 2019. "Antibody-mediated immobilization of virions in mucus." *Bulletin of Mathematical Biology* 81(10):4069–4099.

Johzuka, Katsuki, Masahiro Terasawa, Hideyuki Ogawa, Tomoko Ogawa & Takashi Horiuchi. 2006. "Condensin loaded onto the replication fork barrier site in the rRNA gene repeats during S phase in a FOB1-dependent fashion to prevent contraction of a long repetitive array in Saccharomyces cerevisiae." *Molecular and Cellular Biology* 26(6):2226–2236.

Johzuka, Katsuki & Takashi Horiuchi. 2009. "The cis element and factors required for condensin recruitment to chromosomes." *Molecular Cell* 34(1):26–35.

Jutla, Inderjit S, Lucas GS Jeub & Peter J Mucha. 2011-2019. "A generalized Louvain method for community detection implemented in MATLAB.".
URL: *http://netwiki.amath.unc.edu/GenLouvain*

Kanungo, Tapas, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman & Angela Y Wu. 2002. "An efficient k-means clustering algorithm: Analysis and implementation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7):881–892.

Kipf, Thomas N & Max Welling. 2016. "Semi-supervised classification with graph convolutional networks." *arXiv preprint arXiv:1609.02907* .

Kivelä, Mikko, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno & Mason A Porter. 2014. "Multilayer networks." *Journal of Complex Networks* 2(3):203–271.

Lawrimore, Josh. N.d. "Code Release: Bloom Lab Yeast – Nucleolar Analysis; available at `https://github.com/BloomLabYeast/nucleolarAnalysis`.".

Lawrimore, Josh, Joseph K Aicher, Patrick Hahn, Alyona Fulp, Ben Kompa, Leandra Vicci, Michael Falvo, Russell M Taylor & Kerry Bloom. 2016. "ChromoShake: a chromosome dynamics simulator reveals that chromatin loops stiffen centromeric chromatin." *Molecular Biology of the Cell* 27(1):153–166.

Le, Tung BK, Maxim V Imakaev, Leonid A Mirny & Michael T Laub. 2013. "High-resolution mapping of the spatial organization of a bacterial chromosome." *Science* 342(6159):731–734.

Lieberman-Aiden, Erez, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner et al. 2009. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* 326(5950):289–293.

Liu, Sijia, Haiming Chen, Scott Ronquist, Laura Seaman, Nicholas Ceglia, Walter Meixner, Lindsey A Muir, Pin-Yu Chen, Gerald Higgins, Pierre Baldi et al. 2017. "Genome architecture leads a bifurcation in cell identity." *bioRxiv* p. 151555.

Liu, Sijia, Pin-Yu Chen, Alfred Hero & Indika Rajapakse. 2018. "Dynamic Network Analysis of the 4D Nucleome." *bioRxiv* p. 268318.

Maeshima, K, S Hihara & H Takata. 2010. New Insight into the Mitotic Chromosome Structure Irregular Folding of Nucleosome Fibers Without 30-nm Chromatin Structure. In *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 75 Cold Spring Harbor Laboratory Press pp. 439–444.

Magnus, Jan R. 1985. "On differentiating eigenvalues and eigenvectors." *Econometric Theory* pp. 179–191.

Marko, JF & ED Siggia. 1995. "Statistical mechanics of supercoiled DNA." *Physical Review E* 52(3):2912.

Marko, John F & Eric D Siggia. 1994*a*. "Bending and twisting elasticity of DNA." *Macromolecules* 27(4):981–988.

Marko, John F & Eric D Siggia. 1994*b*. "Fluctuations and supercoiling of DNA." *Science* 265(5171):506–508.

Marko, John F & Eric D Siggia. 1997. "Polymer models of meiotic and mitotic chromosomes." *Molecular Biology of the Cell* 8(11):2217–2231.

Moore, Christopher M, Christopher R Stieha, Ben C Nolting, Maria K Cameron & Karen C Abbott. 2015. "QPot: an R package for stochastic differential equation quasi-potential analysis." *arXiv preprint arXiv:1510.07992* .

Mucha, Peter J, Thomas Richardson, Kevin Macon, Mason A Porter & Jukka-Pekka Onnela. 2010. "Community structure in time-dependent, multiscale, and multiplex networks." *Science* 328(5980):876–878.

Naumova, Natalia, Maxim Imakaev, Geoffrey Fudenberg, Ye Zhan, Bryan R Lajoie, Leonid A Mirny & Job Dekker. 2013. "Organization of the mitotic chromosome." *Science* 342(6161):948–953.

Newby, Jay, Jennifer L Schiller, Timothy Wessler, Jasmine Edelstein, M Gregory Forest & Samuel K Lai. 2017. "A blueprint for robust crosslinking of mobile species in biogels with weakly adhesive molecular anchors." *Nature Communications* 8(1):833.

Newby, Jay M. 2014. "Spontaneous excitability in the Morris–Lecar model with ion channel noise." *SIAM Journal on Applied Dynamical Systems* 13(4):1756–1791.

Newman, Mark EJ. 2006. "Modularity and community structure in networks." *Proceedings of the National Academy of Sciences* 103(23):8577–8582.

Newman, Mark EJ & Michelle Girvan. 2004. "Finding and evaluating community structure in networks." *Physical Review E* 69(2):026113.

Olsen, RA, GJ Kroes, G Henkelman, A Arnaldsson & H Jónsson. 2004. "Comparison of methods for finding saddle points without knowledge of the final states." *The Journal of Chemical Physics* 121(20):9776–9792.

Oluwadare, Oluwatosin & Jianlin Cheng. 2017. "ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data." *BMC Bioinformatics* 18(1):1–14.

Otsu, Nobuyuki. 1979. "A threshold selection method from gray-level histograms." *IEEE Transactions on Systems, Man, and Cybernetics* 9(1):62–66.

Pearson, Chad G, Paul S Maddox, ED Salmon & Kerry Bloom. 2001. "Budding yeast chromosome structure and dynamics during mitosis." *The Journal of Cell Biology* 152(6):1255–1266.

Pope, Benjamin D, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield et al. 2014. "Topologically associating domains are stable units of replication-timing regulation." *Nature* 515(7527):402.

Porter, Mason A, Jukka-Pekka Onnela & Peter J Mucha. 2009. "Communities in networks." *Notices of the AMS* 56(9):1082–1097.

Prieto, José-Luis & Brian McStay. 2007. "Recruitment of factors linking transcription and processing of pre-rRNA to NOR chromatin is UBF-dependent and occurs independent of transcription in human cells." *Genes & Development* 21(16):2041–2054.

Rajapakse, Indika, David Scalzo, Stephen J Tapscott, Steven T Kosak & Mark Groudine. 2010. "Networking the nucleus." *Molecular Systems Biology* 6(1):395.

Rajapakse, Indika & Mark Groudine. 2011. "On emerging nuclear order." *The Journal of Cell Biology* 192(5):711–721.

Rajapakse, Indika, Mark Groudine & Mehran Mesbahi. 2011. "Dynamics and control of state-dependent networks for probing genomic organization." *Proceedings of the National Academy of Sciences* .

Rao, Suhas SP, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander et al. 2014. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." *Cell* 159(7):1665–1680.

Reichardt, Jörg & Stefan Bornholdt. 2006. "Statistical mechanics of community detection." *Physical review E* 74(1):016110.

Ren, Weiqing & Eric Vanden-Eijnden. 2013. "A climbing string method for saddle point search." *The Journal of chemical physics* 138(13):134105.

Rosa, Angelo & Ralf Everaers. 2008. "Structure and Dynamics of Interphase Chromosomes." *PLoS Computational Biology* 4(8):e1000153.

Sakai, Yuji, Atsushi Mochizuki, Kazuhisa Kinoshita, Tatsuya Hirano & Masashi Tachikawa. 2018. "Modeling the functions of condensin in chromosome shaping and segregation." *PLoS Computational Biology* 14(6):e1006152.

Schalbetter, SA, A Goloborodko, G Fudenberg, JM Belton, C Miles, M Yu, Job Dekker, Leonid Mirny & J Baxter. 2017. "Structural maintenance of chromosome complexes differentially compact mitotic chromosomes according to genomic context." *Nature Cell Biology* 19(9):1071–1080.

Schmitt, Anthony D, Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L Tan, Yun Li, Shin Lin, Yiing Lin, Cathy L Barr et al. 2016. "A compendium of chromatin contact maps reveals spatially active regions in the human genome." *Cell Reports* 17(8):2042–2059.

Schroeder, Holly A, Jay Newby, Alison Schaefer, Babu Subramani, Alan Tubbs, M Gregory Forest, Ed Miao & Samuel K Lai. 2020. “LPS-binding IgG arrests actively motile Salmonella typhimurium in gastrointestinal mucus.” *Mucosal Immunology* pp. 1–10.

Serra, François, Marco Di Stefano, Yannick G Spill, Yasmina Cuartero, Michael Goodstadt, Davide Baù & Marc A Marti-Renom. 2015. “Restraint-based three-dimensional modeling of genomes and genomic domains.” *FEBS Letters* 589(20):2987–2995.

Tanizawa, Hideki, Osamu Iwasaki, Atsunari Tanaka, Joseph R Capizzi, Priyankara Wickramasinghe, Mihee Lee, Zhiyan Fu & Ken-ichi Noma. 2010. “Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation.” *Nucleic Acids Research* 38(22):8164–8177.

Taylor, Dane, Rajmonda S Caceres & Peter J Mucha. 2017. “Super-resolution community detection for layer-aggregated multilayer networks.” *Physical Review X* 7(3):031056.

Taylor, Dane, Saray Shai, Natalie Stanley & Peter J Mucha. 2016. “Enhanced detectability of community structure in multilayer networks through layer aggregation.” *Physical Review Letters* 116(22):228301.

Terakawa, Tsuyoshi, Shveta Bisht, Jorine M Eeftens, Cees Dekker, Christian H Haering & Eric C Greene. 2017. “The condensin complex is a mechanochemical motor that translocates along DNA.” *Science* 358(6363):672–676.

Tjong, Harianto, Ke Gong, Lin Chen & Frank Alber. 2012. “Physical tethering and volume exclusion determine higher-order genome organization in budding yeast.” *Genome Research* 22(7):1295–1305.

Traag, Vincent A, Gautier Krings & Paul Van Dooren. 2013. “Significant scales in community structure.” *Scientific Reports* 3:2930.

Vanden-Eijnden, Eric & Matthias Heymann. 2008. “The geometric minimum action method for computing minimum energy paths.”.

Vasquez, Paula A, Caitlin Hult, David Adalsteinsson, Josh Lawrimore, Mark G Forest & Kerry Bloom. 2016. “Entropy gives rise to topologically associating domains.” *Nucleic Acids Research* 44(12):5540–5549.

Vasquez, Paula A & Kerry Bloom. 2014. “Polymer models of interphase chromosomes.” *Nucleus* 5(5):376–390.

Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio & Yoshua Bengio. 2017. “Graph attention networks.” *arXiv preprint arXiv:1710.10903* .

Verdaasdonk, Jolien Suzanne, Paula Andrea Vasquez, Raymond Mario Barry, Timothy Barry, Scott Goodwin, M Gregory Forest & Kerry Bloom. 2013. “Centromere tethering confines chromosome domains.” *Molecular Cell* 52(6):819–831.

Vian, Laura, Aleksandra Pękowska, Suhas SP Rao, Kyong-Rim Kieffer-Kwon, Seolkyoung Jung, Laura Baranello, Su-Chen Huang, Laila El Khattabi, Marei Dose, Nathanael Pruett et al. 2018. “The Energetics and Physiological Impact of Cohesin Extrusion.” *Cell* 173(5):1165–1178.

Walker, Benjamin, Dane Taylor, Josh Lawrimore, Caitlin Hult, David Adalsteinsson, Kerry Bloom & M Gregory Forest. 2019. “Transient crosslinking kinetics optimize gene cluster interactions.” *PLoS Computational Biology* 15(8):e1007124.

Wang, Renjie, Julien Mozziconacci, Aurélien Bancaud & Olivier Gadal. 2015. "Principles of chromatin organization in yeast: relevance of polymer models to describe nuclear organization and dynamics." *Current Opinion in Cell Biology* 34:54–60.

Weinan, E, Weiqing Ren & Eric Vanden-Eijnden. 2002. "String method for the study of rare events." *Physical Review B* 66(5):052301.

Weir, William H, Ryan Gibson & Peter J Mucha. 2017. "CHAMP package: Convex Hull of Admissible Modularity Partitions in Python and MATLAB.".
URL: *http://github.com/wweir829/CHAMP*

Weir, William H, Scott Emmons, Ryan Gibson, Dane Taylor & Peter J Mucha. 2017. "Post-processing partitions to identify domains of modularity optimization." *Algorithms* 10(3):93.

Welte, Michael A. 2004. "Bidirectional transport along microtubules." *Current Biology* 14(13):R525–R537.

Wong, Hua, Hervé Marie-Nelly, Sébastien Herbert, Pascal Carrivain, Hervé Blanc, Romain Koszul, Emmanuelle Fabre & Christophe Zimmer. 2012. "A predictive computational model of the dynamic 3D interphase yeast nucleus." *Current Biology* 22(20):1881–1890.

Wong, Hua, Jean-Michel Arbona & Christophe Zimmer. 2013. "How to build a yeast nucleus." *Nucleus* 4(5):361–366. PMID: 23974728.
URL: *https://doi.org/10.4161/nucl.26226*

Yang, Shuo, Samuel F Potter & Maria K Cameron. 2019. "Computing the quasipotential for nongradient SDEs in 3D." *Journal of Computational Physics* 379:325–350.

Zhou, Joseph Xu, MDS Aliyu, Erik Aurell & Sui Huang. 2012. "Quasi-potential landscape in complex multi-stable systems." *Journal of the Royal Society Interface* 9(77):3539–3553.