MARKOV MODELS FOR THE CONFORMATIONAL KINETICS IN DNA BREATHING FLUCTUATIONS

by

PABLO ROMANO

A DISSERTATION

Presented to the Department of Chemistry and Biochemistry
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2017

DISSERTATION APPROVAL PAGE

Student: Pablo Romano

Title: Markov Models for the Conformational Kinetics in DNA Breathing Fluctuations

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Chemistry and Biochemistry by:

Michael Kellman	Chair
Marina Guenza	Advisor
Andrew Marcus	Core
Peter von Hippel	Core
Alice Barkan	Institutional

and

Sara D. Hodges	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2017

DISSERTATION ABSTRACT

Pablo Romano

Doctor of Philosophy

Department of Chemistry and Biochemistry

September 2017

Title: Markov Models for the Conformational Kinetics in DNA Breathing Fluctuations

As the genetic content is internally located within DNA duplexed form, it has long been hypothesized that DNA undergoes a series of thermally induced conformational changes that assist in protein recognition events. The biological mechanisms for protein-DNA interactions have long been sought after, as little is still known mechanistically about how these complexes form. To study the local contributions to these breathing modes several atomistic simulations of DNA oligonucleotides were generated and analyzed by statistical models to predict metastable conformational states, the system timescales, and the kinetic pathways between states.

In order to sample time-series DNA constructs, microsecond molecular dynamics (MD) simulation were performed. MD simulations provide atomstic resolution of macromolecules in explicit solvent and with modern computational workflows can extend well into microsecond timescales. While MD is a powerful tool, it creates a tremendous amount of time-dependent data. In recent years, Markov State Models (MSM), which project the dynamics of MD simulations

onto discrete coordinates that follow a Markov chain, have become an invaluable tool to model and describe the kinetics of these large datasets. These models can be coarse-grained for chemical insight, however there does not yet exist a method which consistently and "crisply" describe the metastable barriers.

To address this, I developed a new method, called Gradient Adaptive Decomposition (GRAD), which optimizes the macrostate model by refining borders with respect to the gradient along the free energy surface. The proposed method requires only a small number of initial microstates because it corrects for errors produced by limited number of seeds. Whereas many methods rely on fuzzy, or overlapping, partitions for proper statstical analysis of timescales, GRAD retains accuracy and crisp decomposition.

I present a workflow of GRAD refined MSM to analyze the long timescale MD simulations of DNA oligonucleotides to assess the stacking conformational dynamics of DNA. Evaluating the complex network of transitions accessible found evidence suggesting that chiral directed mechanisms are critical in how DNA bases unstack. I explore how these local effects may be significant to long timescale dynamics and the biological impact in relation to breathing fluctuations.

This dissertation includes unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Pablo Romano

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
St. Edward's University, Austin, TX

DEGREES AWARDED:

Doctor of Philosophy, Physical Chemistry, 2017, University of Oregon
Bachelor of Science, Chemistry, 2012, St. Edward's University

AREAS OF SPECIAL INTEREST:

Data Science
Machine Learning
Statistics
Physical Chemistry
Markov Chains

PROFESSIONAL EXPERIENCE:

Graduate Teaching Fellow, Department of Chemistry & Biochemistry, 2012-2015

Graduate Research Fellow, Department of Chemistry & Biochemistry, 2015-2017

GRANTS, AWARDS AND HONORS:

Promising Scholar Award, Department of Chemistry & Biochemistry, 2012

Biophysic Training Grant, Institute of Molecular Biology, 2015-2017

PUBLICATIONS:

P.G. Romano, & M.G. Guenza. (2017). The GRadient Adaptive Decomposition (GRAD) method: Optimized refinement along macrostate borders in Markov State Models. Submitted to *Journal of Chemical Information and Modeling*. In review

P.G. Romano, C. Garcia, T. Fencl, & M.G. Guenza. (2017). Chiral Directed Mechanisms in the Unstacking DNA Oligonucleotides. In preparation

ACKNOWLEDGEMENTS

I give tremendous thanks to my advisor, Dr. Marina Guenza. For the insights you've provided, the wealth of knowledge you've shared, and challenging me to question every aspect of my research. I've come a long way as a scientist thanks to your mentorship and I hope I continue your level of rigour in all my future endeavors. To the members of the Guenza lab, and in particular Dr. Jeremy Copperman, thank you for the countless conversations (related to my research and otherwise), sitting through my group meetings, and making my time at the UO great. To the members of my thesis committee thank you for sharing your experiences and expertise. Your continued enthusiasm to find connections between my simulations and the experiments in biology have been invaluable at every stage of my graduate career.

This work was mostly supported by NIH training grant T32 GM007759 (to P.G.R.). Computational time was provided by NSF Grant No. ACI-1053575 through Extreme Science and Engineering Discovery Environment (XSEDE) resources.

I dedicate this to my parents who introduced me to math and science at an early age and sat with me constantly to explore my every question. I owe so much to the courage, strength, and the endless opportunties you both have provided for me. And to Olivia, you've been my everything while I worked through my graduate career. I can't thank you enough for your unwavering encouragement, love, and support.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

While much is known about the structure of DNA and the several proteins and the macromolecular machinery that interacts with it, there still exists several fundamental questions about how proteins interact directly with DNA and what role the conformational dynamics of DNA plays in relation to them. From Watson and Crick's identification[63] of the duplexed DNA, it became clear that the genetic information of DNA was stored within the double stranded helix. The chemical components were later identified as nucleic acids adenine (A), cytosine (C), guanine (G), and thymine (T). Soon after it became evident that as the nucleic residues are read by macromolecular systems, these bases must be directly accessible. In several events proteins, if not full macromolecular machinery, directly interact and scan through the nucleic residues and as such must gain access to DNA's internal content. The exact mechanism of how this occurs is unknown but there is evidence to suggest DNA undergoes a series of "breathing" fluctuations, the role of which has remained a fundamental question in the biological community for decades[60].

Two paradigms of thought exists in the explanation of this recognition process, protein and nucleic centric. Protein centric interpretations postulated that proteins induce the conformational change in DNA allowing for direct access, such as a proteins first binding to the exterior of DNA producing a nucleation site on DNA. This would suggest that the access to genetic codes must be induced by an external force.

Alternatively, the nucleic centric view posited instead that conformational changes in DNA are thermally induced and act as a sort of signaling mechanismfor

proteins. This view is based on early melting studies of DNA where lower temperatures still observed levels of fluctuations about closed states. Additional support was earn by proton exchange experiments[45, 46, 33] which showed that the Watson-Crick hydrogens underwent solvent exposure although these conformational changes may have only been large enough to introduce solvent. As DNA bases become solvent exposed within these breathing modes, it has been speculated that the underlying mechanism allows for direct interaction with proteins.

Recent computational advances brought upon by massive parallelization and new technology, have extended the timescales and system sizes accessible by Molecular Dynamic (MD) simulations. Simulations can now sample the conformational landscape at biologically relevant times (micro and millisecond timescales) and now more than ever the atomistic mechanism of local DNA base interactions can be immediately explored. This doctoral work presents several simulations of short length oligonucleotides to explore the stacking conformational landscape inherent in single stranded DNA to model the disorder conformational dynamics of bubbled or frayed residue.

These advances in MD have also required with them more sophisticated methods that are capable of sufficiently and accurately capturing the underlying kinetics from MD trajectories. One such popular method has been Markov State Models (MSM), a statistical network model which captures the dynamics of the macromolecular simulations as a Markov process. In this framework, the free energy surface (FES) is decomposed into disjoint regions which define discrete states, the transitions between which can be treated as memoryless, or Markovian, jumps.

To gain insight into the stochastic breathing fluctuations inherent in DNA as it undergoes large conformational changes, it was necessary to analyze the metastable conformations associated with base stacking in short oligonucleotides. While several coarse-graining schemes exists, they require the fine discretization of the conformational landscape which can be computationally costly, or they employ fuzzy clustering which makes it difficult to identify barriers between states. Towards this end, a new method, Gradient Adaptive Decomposition (GRAD), which refines the decomposition of coarse-grained Markov state models (MSM) was developed. While this dissertation details the method in full and describes the underlying theory, in short given a conformational landscape that has been clustered as a MSM the refinement method uses the gradient along the energetic landscape to correct for poor discretization. The direct result is a refined model which accurately describes the crisp separation between metastable states. GRAD refined MSM are used to analyze the complex conformational dynamics and provide insights into stacking mechanisms of DNA.

Chapter II of this text describes theory of Markov State Models (MSM) in great detail, presenting the methodology used within this doctoral work, as well as presenting alternative methods used in the field. It concludes with Transition Path Theory (TPT), and its application to MSM. It is intended to be used as a primer or review to prepare the reader for the material. A working understanding of these concepts is necessary to the remainder of this dissertation.

Chapter III describes methodology used to atomistically simulate DNA oligonucleotides in explicit water, as well as presents a two-site per nucleotide structural model used to capture the stacking conformations sampled by MD

simulations. These simulations and model are the core work presented in chapters IV & V.

Chapter IV details the novel method of Gradient Adaptive Decomposition, which builds upon the standard approach used in MSM and shows consistent correction for sampling errors. The method uses information from the gradient of the underlying free energy surface to correct the decision boundary between metastable states. By maximizing the boundary location, the timescales are maximized producing a better decomposition, and likewise a reduce discretization error, of the metastable states. This chapter presents the application of this workflow on ideal diffusive models of varying landscape complexity, and a more complicated yet biologically relevant Molecular Dynamics simulation of Adenosine Dinucleotides. This work was co-authored with Dr. Marina Guenza and has been submitted for publication and is currently under review.

Chapter V studies the unstacking mechanisms in DNA dinucleotides, which are a helpful model system in the study of DNA breathing fluctuations. These nucleotide systems are not restricted in motion by long range π-stacking effects and as such are useful in modeling the conformational dynamics associated with base stacking. These results are extended by a length dependent study of tri and tetranucleotides where sequences are extending by flanking thymine. A full review of sequence specificity and strand polarity is analyzed via GRAD refined MSM, and the pathways are evaluated using TPT. Some portions of this work have been co-authored by Dr. Marina Guenza and are in preparations for publications.

CHAPTER II

MARKOV STATE MODELS: A PRIMER ON THEORY, GENERATION, AND ANALYSIS OF KINETIC NETWORKS

This chapter introduces a short primer on Markov State Models (MSM) a popular kinetic network model used in the analysis of simulation data. The underlying theory, methods for generating these models, coarse-graining, and analysis of characteristic timescales and pathways are presented. The focus of this chapter is to provide sufficient information for a succinct review such that readers can understand the remaining material of this doctoral work. While within the context of this work the use and application of these methods are primarily used towards the analysis of nucleic acid systems, MSM are a general method for any time-series data where one might prefer to cluster by correlations in time over geometric features in the dataset. For a more detailed account of the MSM field, readers are encouraged to review the vast available literature [21, 8, 47, 11].

In recent years, along with the advances in computing power, molecular dynamics (MD) simulations have been able to reach longer timescales for larger systems. MD has become exceedingly useful for understanding the atomistic properties of biological systems, however the complexity of these trajectories has made modeling the conformational dynamics challenging. MSM have grown in popularity as they have proven to be a useful method by which to address this complexity and create detailed and robust models.

Within the context of analyzing MD simulations, MSM typically refer to the application of machine learning algorithms to partition the conformational landscape into a number of states. Within this "state decomposition" the full space

is discretized into disjoint configurational sets and all transitions in time between sets are allowed with no memory of the previous times. Several procedures have been established and have been extensively reviewed elsewhere[11], which describe methods by which to partition the conformational landscape. While it is often useful to conceptualize these nodes as minima along the free energy surface, it is important to understand that this is not necessarily true. These states instead are divisions in the landscape that separate conformations along energetic barriers.

Typically, simulation data is clustered in a microstate model, where effectively the full landscape is discretized "geometrically" by centroid based algorithms. While there are several variations which accomplish this, in essence all place "seeds" or "centroids" at regions of dense sampling from the simulation data. Once all seeds are generated, the simulation data are assigned by nearest neighboring centroid, creating a discrete region of the conformational landscape referred to as a microstate.

These microstate models provide accurate statistics, and can properly discretize the coordinates, however for kinetic insights they are typically coarse-grained. This refers to a second level of clustering, which aims to "kinetically" cluster the microstates into "macrostates", which are disjoint metastable sets which minimize the likelihood of interconversions between macrostates. The procedure most commonly used in this step is the Robust Perron Cluster Analysis (PCCA+), however other methods have been proposed to coarse grain the microstate kinetics[67, 7, 39]. Within PCCA+ a spectral decomposition of the stochastic transition matrix uses the structure of the related right eigenvectors to group microstates into larger metastable states by how rapidly they interconvert.

Kinetic Markov Models

Markov state models refer to method of analysis where time-series data is projected onto discrete states which are characterized by a Markov chain. The Markov property, named for the Russian mathematician Andrei Markov, simply states that transitions between states are memoryless, and the conditional probability to transition to a new state is given only by the current state. In other words the evolution of coordinates is modeled at a lag time where previous conformations are no longer correlated.

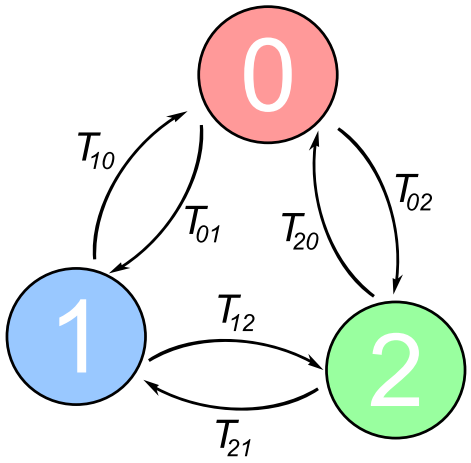

FIGURE 1. A simple network representation of a arbitrary 3 state Markov State Models.
The nodes, or Markov states, are shown as labeled circles, with the conditional likelihood of transitioning between states given by directed arrows.

In essence, MSM create a network of nodes (see figure 1), or conformational states, wherein the edges between all nodes represent the kinetics or transitions between nodes. Simulation data is described in the continuous state space Ω such that all timesteps t transition between discrete states $S(t)$. By employing a Markov approximation, such that the jumps between nodes are uncorrelated

in time or without memory of previous transitions, the kinetics from MD can be succinctly described by a simple yet eloquent mathematical formulism governed by a stochastic transition matrix, defined as

$$T_{ij}(\tau) = P(S(t+\tau) = j | S(t) = i) \ , \tag{2.1}$$

where the conditional probability for a random process transitions from state i to j over a given lag time τ.

The kinetics of the MSM network is described by a Master Equation (ME) formalism[58], which describes the molecular kinetics of a process as a Markov-chain of uncorrelated jumps among conformational states. This formalism describes rate of leaving a state by

$$\dot{\boldsymbol{p}}(t) = \boldsymbol{K}\boldsymbol{p}(t) \ , \tag{2.2}$$

where $\boldsymbol{p}(t)$ is the population state vector at time t, and $\mathbf{K}$ the kinetic rate matrix. We adopt the convention that a dot denotes the rate with respect to time. The formal solution to the ME is therefore simply expressed as

$$\boldsymbol{p}(t) = \mathrm{e}^{\boldsymbol{K}t}\boldsymbol{p}(0) \ , \tag{2.3}$$

from which it is simple to define the relation of the transition matrix to the kinetic rate matrix,

$$\boldsymbol{T}(\tau) = \mathrm{e}^{\boldsymbol{K}\tau} \ . \tag{2.4}$$

The transition matrix, has elements T_{ij} which represent the conditional probability that a simulation will arrive to state j from i over a lag time τ. As

will be discussed in a later section, the transition matrix is simpler to estimate from simulation data than $\boldsymbol{K}$, and as such the MSM community primarily uses this form.

The ME formalism can therefore be rewritten in the following form

$$\boldsymbol{p}(n\tau) = [\boldsymbol{T}(\tau)]^n \boldsymbol{p}(0) \ , \tag{2.5}$$

where a population vector $\boldsymbol{p}(t)$ at some time t can be evolved according to the transition matrix for some lag time, τ.

Estimating the Transition Matrix

In order to obtain an accurate model that describes the underlying kinetics from simulation, there must be a suitable estimation of the transition matrix, $\boldsymbol{T}$. While for an infinitely long simulation with ergodic sampling it is somewhat trivial to calculate the conditional probabilities of transition between states, generally this is not the case. more realistically simulations of macromolecular systems are typically sparse and struggle with undersampling. For this reason, the method of estimating the transition matrix have resorted to statistically rigorous methods. In this section we present some of the more standard estimators, although this is by no means an exhaustive list.

Counting Transitions Between Discrete States

In order to successfully describe the simulation with any meaningful statistics, it is necessary to collect information directly that details how trajectories transition between discrete states. To that end, a important metric is the number of observed

counts for all transitions $i \rightarrow j$. For any simulation, or more generally a time-series dataset, we can describe the simulation, $S(t)$, in terms of the discrete state space $\mathbf{\Omega}$. The count matrix can be estimated by counting all transitions from state i at time t to state j at time $t + \tau$. The sum of all counts from $S(t)$ to $S(t + \tau)$ is therefore defined as the observed count matrix $\boldsymbol{C}^{obs}$. The numerical form can be written as as a summation across time frames k as

$$C_{ij}^{obs} = \sum_{k=1}^{N-l} S_{k,k+l} \; \forall \; S_k = i, S_{k+l} = j \; , \tag{2.6}$$

where N is the total number of time steps, Δt, in the simulation, and l is the number of time steps associated with the model lag time, $\tau = l\Delta t$.

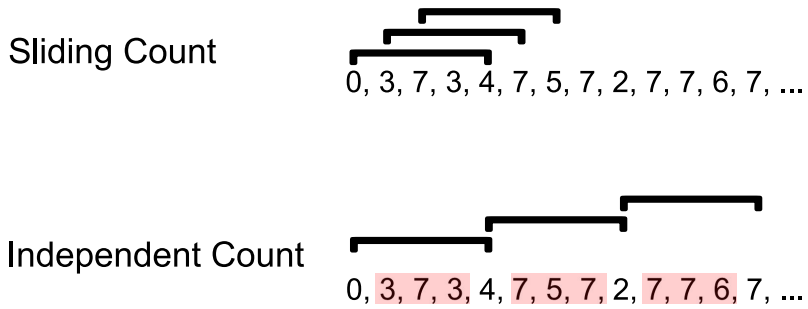

FIGURE 2. The counting methods for sliding window (top) and independent counts
(bottom) are compared for the same trajectory projected onto the coordinates of discrete state space $\mathbf{\Omega}$ from the Markov model. Notice that in the independent counts all states shaded in red are discarded.

Note that it is important to use all time steps from the simulations from frames $k = 1 \rightarrow N - l$ or equivalently $t_0 \rightarrow N\Delta t - \tau$. By using independent counts, that is $t = 0 \rightarrow \tau, \tau \rightarrow 2\tau, ..., (N-1)\tau \rightarrow N\tau$, with the exception of $\tau = 1$ a large

portion of data is being discarded. This is due to only sampling every τ values in time, and therefore all information in between this lag time is lost. Instead, it is recommended that a sliding window approach be used, where the observed counts are sampled from $t = 0 \rightarrow \tau, \Delta t \rightarrow \Delta t + \tau, 2\Delta t \rightarrow 2\Delta t + \tau, ..., (N-1)\Delta t \rightarrow (N-1)\Delta t + \tau$. In figure-2, it is shown visually, how to maximize the available data present from MD simulations.

Non Reversible Estimation of the Transition Matrix

For a simulation as it approaches an infinite time length with reversible sampling, we could trivially define the conditional probability of a transition by directly counting all observed jumps between states within the simulation.

$$T_{ij}(\tau) = \frac{C_{ij}^{obs}(\tau)}{C_{i}^{obs}(\tau)} \tag{2.7}$$

We define the number of observed transitions from state i to j over lag τ as $C_{ij}^{obs}(\tau)$, and we employ the shorthand

$$C_{i}^{obs}(\tau) = \sum_{j} C_{ij}^{obs}(\tau) \ , \tag{2.8}$$

to represent the summation across rows. This estimation of the transition matrix is referred to as the non-reversible estimator. As simulations are rarely long enough to fully sample, this non-reversible estimation is usually inaccurate.

It is important to note that from these definitions, the transition matrix is defined as a row stochastic or row normalized matrix. While this is ultimately arbitrary as theory allows for column or row stochastic matrices as they are related by a transpose operation, this dissertation will adopt the convention of row

normalization. In the literature while row stochastic is more frequently used, many papers still use the column normalized form. Therefore, readers should be aware that operations with the transition matrix may be transposed to the equations used here, with right and left eigenvectors being flipped.

Maximum Likelihood Estimation by Symmetrized Transition Counts

In accordance with detailed balance, for perfectly reversible simulation one would expect the counts forward and backwards be equivalent, such that $C_{ij}(\tau) = C_{ji}(\tau)$. However, it is such the case that in general this is not fulfilled as simulations will tend to have limited sampling and as such will be biased towards the starting conditions. Improved accuracy can be obtained by requiring detailed balance[21, 8] by symmetrizing the observed counts. This maximum likelihood estimator (MLE) takes the average of forward and backward transition as

$$\bar{C}_{ij}(\tau) = \frac{C_{ij}(\tau) + C_{ji}(\tau)}{2} \ . \tag{2.9}$$

and the transition matrix for the ME is then row normalized as

$$T_{ij}(\tau) = \frac{\bar{C}_{ij}(\tau)}{\bar{C}_i(\tau)} \ . \tag{2.10}$$

This simple MLE works well to build the transition matrix in the limit of infinite sampling, or more practically when the length of the simulations is considerably greater than the predicted timescales of the transition matrix. Commonly, this is not the case and more rigorous methods are required. In the following section, we detail a reversible MLE method which is regarded as the standard in the field.

Reversible Maximum Likelihood Estimator

In the work of Prinz et al[47], they developed a method which predicts the MLE for reversible systems, such that detailed balance is enforced. Where in section-2.2 the MLE tried to find the maximum likelihood from the observed count matrix, this method instead maximizes the posterior probability of a transition matrix given a set of observed count matrices as shown in equation-2.11.

$$p(\boldsymbol{C}^{obs}|\boldsymbol{T}) \propto \prod_{i,j=1} T_{ij}^{C_{ij}^{obs}} \tag{2.11}$$

In practice, this is typically solved as the log likelihood, as it is more computationally efficient to deal with logarithms than large exponentials. To ultimately simplify the maximization scheme, the method begins by defining the marginal probability for a transition from i to j as a new set of variables

$$x_{ij} = \pi_i T_{ij} \ , \tag{2.12}$$

where we define

$$x_i = \sum_j x_{ij} = \pi_i \ . \tag{2.13}$$

Notice that in this form, the transition matrix can easily be recovered as x_{ij}/x_i. Therefore, this method requires a prior estimate to the transition matrix, which can be either the non-reversible case which does not necessarily obey detailed balance, or the estimation described in equation-2.10. The problem then proceeds by maximizing the log form of equation-2.11 according to

$$\sum_{i,j} C_{ij} \log \frac{x_{ij}}{x_i} \ , \tag{2.14}$$

where we constrain the problem such that $x_{ij} = x_{ji}$.

The method is iterated until a convergence in the posterior probability, and for x_{ij} which maximizes this form the transition matrix is defined as $T_{ij} = x_{ij}/x_i$.

Generating Microscopic Models from Simulation Data

The primary difficulty in defining a network model, such as MSM, from macromolecular simulations is that the states themselves are nontrivial to define given the complexity and hyperdimensionality of the conformational landscape inherent within MD simulations. Even considering small and simple molecules with low degrees of freedom, it can be difficult to define states for kinetic models as conformational minima are not guaranteed to be metastable if the barrier to escape is relatively low. An additional concern in the classification of states, is that thermally induced stochastic fluctuations can add noise in such a way that the peak barrier is non obvious. This results in large uncertainty, or discretizationerror, in where to define the separation between states.

It is therefore beneficial to use an unsupervised classification method. These are machine learning algorithms which use the structure from data to cluster without labels, thereby decomposing the landscape by geometrical features. In particular, for MSM these methods aim to sample well the underlying distribution of the conformational landscape, and due to the large size of MD simulations they should scale well with computational time. As the conformational landscapes can be fairly complicated for molecular processes, it is often necessary to finely discretize the landscape, and as such this level of modeling is referred to as a microstate clustering. This section will only cover *k-means*, *k-medoids*, and Ward

clustering but for reference we show the difference between various clustering methods in figure 3

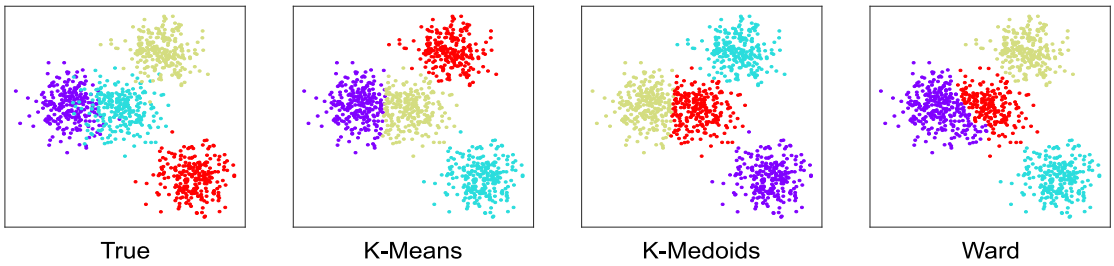

FIGURE 3. The results of unsupervised clustering is performed on randomly generated
data. The clusters compare against the true labels, *k-means++*, *k-medoids*, and hiarchical clustering with Ward's algorithm.

K-Means Clustering

In the standard *k-means* method[32], the trajectory is clustered into microstates by generating state centroids which intuitively can be thought of as decomposing the conformational landscape. The initial centroids are positioned at random within the parameter space spanned by the trajectory data. All data points in the trajectory, which are sampling the energy landscape, are assigned to their nearest centroid, and as such the multidimensional free energy surface is "seeded" by the generation of centroids. Once this step is completed, the position of each centroid is redefined as the center of mass of all points assigned to it. These two steps are repeated until the centroid positions converge. At that point, the resulting energy landscape is partitioned in a set of k microstates. This method requires as input the number of k centroids to be generated.

The advantage of the *k-means* procedure is its speed and simplicity, but limited due to inaccuracies. Without careful seeding, *k-means* is prone to faulty convergence, and the centroids are not necessarily reproducible. In recent years the sampling accuracy has been improved by selecting a better seeding procedure[3, 2], usually referred to as *k-means++*. In this updated clustering method, the acceptance of a centroid is weighted by its square distance from the closest centroids. In this way, the procedure tries to sufficiently sample all the regions in the configurational landscape. The precision of the method increases with increasing the number of centroids.

In practice, configurations that are highly sampled by the simulations have a higher density of centroids, while conformational regions that are weakly sampled have sparsely distributed centroids. In the limit of very large k, high energy barriers are accurately discretized, maximizing the separation in timescales to transition over large barriers. Obviously, this imposes a trade-off in the accuracy of the microstate model and the computational costs. It is often necessary to test for a high optimum number of centroids which increase the accuracy of the *k-means++* clusters. It is common, depending on the complexity of the landscape, to use a few thousand to tens of thousand microstates to sample the free energy surface.

An additional factor is the computational time in which MSM increases due to the diagonalization of the transition matrix. As the eigenvalues and vectors are critical in the kinetic analysis, the computational costs of the model scale with the size of $\boldsymbol{T}$. As $\boldsymbol{T}$ has elements for transitions between all centroids, and becomes sparse when the number of centroids is high, the probability of transitions between some states becomes very low. It is possible, to reach such a large number of microstates, when simulations have sparse sampling, which returns a null populated

microstate violating the Markov conditions. Care must be taken when considering the number of microstates by validating the clustering of simulation data. This can be performed through various methods such as a Silhouette score[50], however the standard workflow to evaluate whether the MSM has a suitable number of microstates is through the analysis of the transition matrix, $\boldsymbol{T}$. In such a case, the predicted timescale

K-Medoids Clustering

As the name suggests, *k-medoids* is similar in concept to *k-means* however instead of finding the mean centroid position, it places the restriction that the centroids must be a data point. The method therefore follows a similar protocol however when updating the centroid position, instead of moving to the center of mass the centroid moves to cluster medoid. This is the data point within each cluster that minimizes the dissimilarity to all other samples. While varying definitions of similarity can be used, the dissimilarity matrix typically refers to the upper or lower triangle of a euclidean distance matrix. The cost is computed at each iteration as the distance of all data within a cluster to its medoid, and it is repeated until convergence.

While slower, there are certain cases in which *k-medoids* is preferred to *k-means* or *k-means++*. In sparsely sampled trajectories averaging the position across data samples per cluster can result in non-physical artifacts. Imagine a case where a cluster is localized around a region of the conformational landscape that is forbidden by a force-field. The mean location may very well place a centroid at the forbidden region. Instead, in *k-medoids* the medoids would have to be a data point and would therefore only be position where the simulation can directly sample.

Hierarchical Coarse-Graining (HC)

While it is desirable to coarse-grain MSM temporaly or according to their kinetics, it has been suggested that such methods are not necessary and geometric clustering can produce satisfactory both microscopic and coarse grained models[28] which equally reproduced the results of kinetic methods in some systems[31]. While various hierarchical clustering methods exist, in this section the agglomerative Ward's[62] algorithm is presented as it has become a standard clustering tool in various fields. Due to the hierarchical nature of the algorithm, it can in essence be used in clustering simulation data for both micro and macrostate decomposition.

As the name suggest the method groups all data points in a branched hierarchy, and grouping data by the minimum variance. Ward's method begins by first computing the matrix of all pairwise distances between data points, and then merges the two closest points. For a system with n data points, this produces $n-1$ clusters where all but one is singleton, or single element sets. The method then updates the distance matrix to incorporate the distance between the new cluster and all others. This is repeated until all data is assigned into a cluster.

Given a case where cluster c_i is compared to a new cluster c_j and some singleton c_s with sizes n_i, n_j, n_s respectively, Ward's algorithm updates the distance matrix by

$$\sqrt{\frac{n_i+n_s}{N}d(c_i,c_s)^2+\frac{n_i+n_j}{N}d(c_i,c_j)^2-\frac{n_i}{N}d(c_s,c_j)^2}\ , \tag{2.15}$$

with the total number of data points given by N and the nonsquared euclidean distance defined by function d.

There are several implementations of Ward's algorithm currently available, however there is not a standard distance metric. It is important to know prior

to use whether an HC implementation uses a squared or nonsquared euclidean distance matrix. Additionally, because the hierarchy is defined within the data points, agglomerative clustering does not typically have a "prediction" method associated with it. As the cluster assignments create branches in the hierarchy by recursively updating minimum distance between every point in the original set of data, new data points cannot simply be assigned by nearest distance without disturbing the nested partitions. Previous work[29] has shown that the recursive update is not necessary and an updated objective function[28] for predicting new assignments can be given by

$$\sqrt{\left(n\sum_{i=1}^{n} d(x_i, P)^2 - \sum_{i\neq j} d(x_i, x_j)^2\right)\left(\frac{2}{n(n+1)}\right)}\,. \tag{2.16}$$

The distance between clusters is simply defined as the weighted distance between an unknown prediction P and data within a cluster $\{x_i, x_{i+1}, ..., x_n\} = C$. Incorporating this scheme allows for new data points to have their position predicted within the hierarchy.

Coarse Graining Network Models

It is often necessary to coarse-grain the microstate model for the purposes of visualization and chemical insight from a set of conformations. The original goal of MSM were to build models which could capture metastable states, and as such coarse-graining plays a significant role in the field. While the microstate model can give accurate predictions of the timescales of the Markov process, it is unable to provide information pertaining to metastable conformations and the transitions across large barriers. There is still no clear consensus on how best to coarse-

grain MSM and remains an active area of research in the field[20, 56, 55].This section covers two methods: Perron Cluster Cluster Analysis (PCCA), and its updated form Robust Perron Cluster Analysis (PCCA+) which cluster by kinetic terms using the spectral decomposition of the ME transition matrix. As discussed previously, Hierarchical Clustering which uses a geometric approach (see section 2.3) can be used to moderate success for coarse-graining as well. It is worth noting that alternative methods have been proposed[36, 7, 66, 39, 31], however are not covered within the scope of this dissertation.

Perron Cluster-Cluster Analysis (PCCA)

The earliest methods for coarse-graining MSM attempted to maximize the separation in timescales by grouping all microstates in such an arrangement where the model metastability is maximized and remained disjoint from one another. Originally, such methods focused on spectral decomposition where long timescale processes could be estimated by the structure in the eigenvectors. While today the standard method for this is PCCA+, the earlier version PCCA is introduced to give the necessary foundation of this clustering scheme.

To initialize the macrostate decomposition recall that microstate clustering defines centroids which sufficiently sample the conformational distribution and project trajectory data onto discrete states. From these discrete state coordinates the transition matrix can be calculated for a given lag time τ according to the methods discussed in section 2.2. The goal of a coarse-graining procedure is to find the grouping of microstates that groups rapidly interconverting microstates, while maximizing the separation in timescales of jumps between macrostates. As the diagonal elements of the transition matrix define the probability of persisting

within the same state over the Markov time τ, the trace of $\boldsymbol{T}$ can be treated as a measure of metastability,

$$M = \sum_i T_{ii}(\tau) \ . \tag{2.17}$$

With this metric defined, how "good" a model has been coarse-grained can be evaluated.

In order to understand how microstates can be grouped into metastable sets, let us consider the properties of the stochastic matrix $\boldsymbol{T}$. In the Markovian limit, all the states are connected and obey detailed balance, and as such the first eigenvalue $\lambda_1 = 1$ with implied time $t_1 = \infty$ which can be trivially shown from equation-2.4. For all other $\lambda_{i>1}$, each processes has decreasing timescales as i increase, where all eigenvalues are bound $|\lambda_i| \leq 1$. Likewise, the associated right eigenvectors $\boldsymbol{\psi}_i$ represent how likely simulations interconvert between centroids at a given timescale. Take for example $\boldsymbol{\psi}_2$, the second right eigenvector. Each element within $\boldsymbol{\psi}_2$, would represent the centroid coordinates on the second eigenvector basis. The distance between all elements show which centroids have the most probable, or fastest, interconversions. By employing a spectral decomposition, it is therefore possible to group all microstates by their kinetic relevance, as opposed to strictly a geometric method. This principle is the foundation to Perron Cluster Cluster Analysis (PCCA).

PCCA clusters microstates based on the separation in timescales, therefore the method begins by first determining how many metastable states can accurately describe the macroscopic properties of the kinetic model. From equation-2.4 we can trivially define the relationship between eigenvalues and the implied timescales of the system as

$$t_i = \frac{-\tau}{\ln |\lambda_i|} \ . \tag{2.18}$$

As all the eigenvalues are in descending order, and as such descending timescales, large separation between λ_i/λ_{i+1} represent a large separation in time. By evaluating these ratios, it is possible to determine how many processes are distinct in time and as such how many metastable states one would expect.

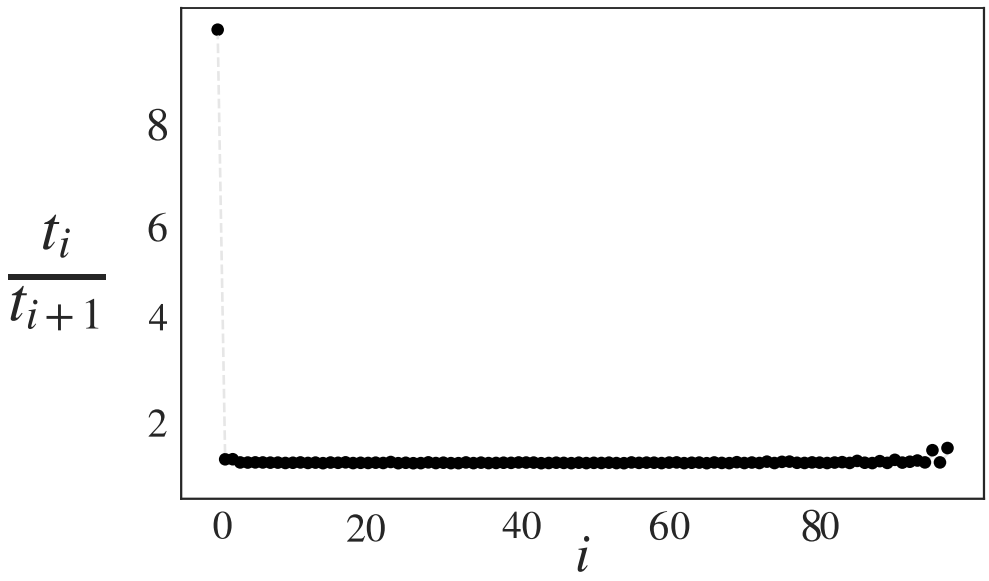

FIGURE 4. The separation in timescales from a diffusion simulation along a two well
potential is shown. Notice that there is one large separation in timescale, implying a two state model.

In figure 4 to illustrate the separation in timescales, a MSM is generated for a Brownian particle freely diffusing on a two well potential. The details of this simulation and generated MSM can be found in chapter 4.3. Notice that in figure 4 there is only one large valued separation, indicated only two metastable states.

This spectral decomposition comes as an extension of the Perron-Frobenius theorem where the stochastic matrix, $\boldsymbol{T}$, with Perron eigenvalue ($\lambda_1 = 1$) has a corresponding right eigenvector of all positive elements. By looking at larger k^{th}

eigenvectors, the elements would transition from positive to negative or vice versa, with increasing number of transitions corresponding to the value of k. Therefore by identifying the eigenvector components where these transitions occur, one can identify which centroids are separated and cluster accordingly.

For example, take the case of Brownian diffusion along a two well potential. To cluster this into a two macrostate model using PCCA, the method would divide all regions of $\boldsymbol{\psi}_2$ by separating the negative and positive valued elements into distinct metastable states. Notice that we do not use $\boldsymbol{\psi}_1$ as all elements equal 1 in accordance with the Perron-Frobenius theorem. Intuitively this makes sense, as the first eigenvector corresponds to an infinite time process, and the probability of any centroids transitions to another at infinite time must be 1 for a connected Markov chain. In this sense, the first eigenvector defines the trivial procedure to create a single macrostates from all microstates.

To make further divisions in the system, such as a three state model, PCCA would first take the clustering predicted from $\boldsymbol{\psi}_2$, and select the macrostate with the largest variance within the eigenvector. Within the selected set of microstates, PCCA then further divides the state by analyzing the transitions according to $\boldsymbol{\psi}_3$. This allows direct selection of the number of macrostates desired, where the system can be clustered up to one less than the total number of microstates.

While this method is theoretically sound, it has the significant drawback in that when the separation in timescales is low, or that if the simulation is too noisy, PCCA is unable to accurately localize the separation between states. In the following section, PCCA+ is introduced as a robust implementation of this theory which better deals with noisy sampling.

Robust Perron Cluster Analysis (PCCA+)

Robust Perron Cluster Analysis[22, 49] (PCCA+) was originally devised as an improvement to PCCA. The goal of this method is still to employ spectral decomposition to kinetically coarse-grain microstate MSM, however to employ more rigorous statistical analysis by defining a fuzzy membership probability to assign all centroids to possible macrostates. To compute these probabilities PCCA+ makes use of the simplex geometry of the eigenvectors of a stochastic matrix. PCCA+has become the standard method to coarse-grain microstate models. Figure 5 illustrates metastable assignment in a two-well potential.

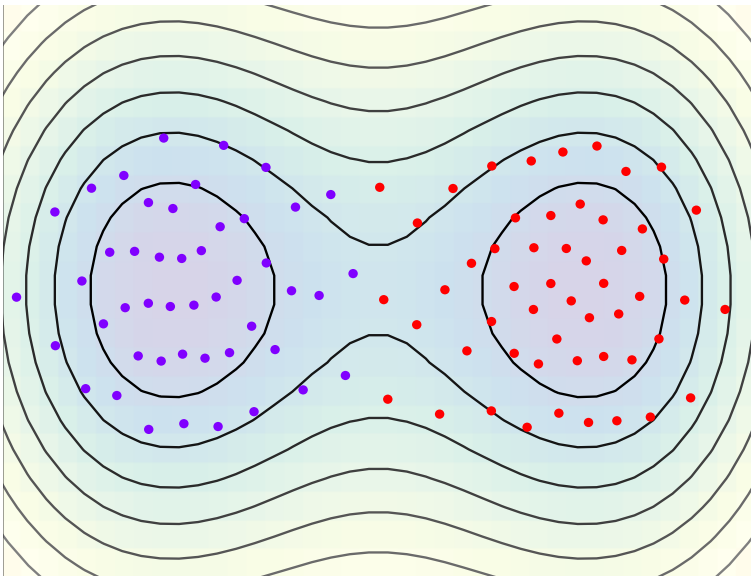

FIGURE 5. A Markov State Model coarse-grained by PCCA+ modeling the diffusion of
a Brownian particle along the shown two well potential. Centroids were computed via *k-means++* and are shown as dots with color representing metastable assignment.

PCCA+ begins by treating the sampled stochastic transition matrix as a perturbation of an ideal uncoupled Markov chain. In the ideal case when a

Markov chain reaches a metastable state the timescales to escape are long enough it does not leave to any other transient states. Generally, this is not the case and so the PCCA+ treats the transition matrix as a first order perturbation of an ideal Markov chain. The goal of PCCA+ then becomes to find a non-singular transformation matrix $A \in \mathbb{R}^{n \times n}$ which transforms the right eigenvectors ($\boldsymbol{\psi}$) into their associated membership probability

$$\chi = \boldsymbol{\psi} A \ . \tag{2.19}$$

As an infinite number of solutions exist[22] to this system of equations, an objective function is maximized to select a solution for χ. As the goal of PCCA+ is to kinetically cluster the microstates, this objective is typically the metastabillity solved from the fuzzy, coarse-grained transition matrix ($\boldsymbol{T}_M$)

$$\boldsymbol{T}_M = D^{-1} \chi^T D T m_\mu \chi \ , \tag{2.20}$$

where $D = \chi^T \pi$, the microstate stationary distribution π solved from the microstate transition matrix $\boldsymbol{T}_\mu$.

Some drawbacks exist to kinetic clustering schemes such as PCCA+. One such drawback to spectral methods in general, is the solution to the eigenvalue problem scales poorly with increasing number of states, and requires the use of sparse linear algebra methods typically when the transition matrix has size greater than 4000 by 4000. At this limit, it is not guaranteed that most methods will converge to a solution for the full decomposition of $\boldsymbol{T}$.

Additionally, one of the benefits of PCCA+, the fuzzy membership, can also be problematic if simulation data is too sparsely sampled or too noisy. At

these regions, the separation between eigenvectors can be difficult to assigned along the simplex geometry of the eigenvector space, and results in no clear macrostate assignment to several microstates. This can be conceptualized to having macrostates so fuzzy, that their regions along the conformational landscape significantly overlap. In the following chapter IV, a new refinement procedure is introduced which corrects for this fuzzy overlap and provides a crisp model.

Validating Markov State Models

A necessary role in the generation of any model, and especially so for MSM, is validating the models themselves. A critical approximation of the MSM is that the kinetics of the system, can be modeled as discrete jumps uncorrelated in time. Therefore, in addition to evaluating the model for accurately reproducing simulation data, it is necessary to validate that the MSM upholds the Markov conditions.

Checking for Convergence in Predicted Timescales

In order to validate the number of microstates, or in other words ensure that the model has adequately discretized the conformational landscape, there should be convergence in the predicted timescales. By spectral decomposition, the slowest, non-infinite time, t_2 can be directly calculated from the second eigenvalue, λ_2 according to equation 2.18. If sufficient number of microstates have been generated then increasing the number of centroids would no longer change the implied timescales of the model.

This same logic can be used to validate the lag time, τ. In order to build an accurate MSM the lag time must satisfy the Markov condition. In accordance with

a Markov process, such that all states are connected and obey detailed balance, the eigenvalues of $\boldsymbol{T}$ are expected to obey $\lambda_i \leq 1$. As shown from equation-2.4 the implied timescales of the MSM can be predicted in the form as shown in equation-2.18. For a real, physical system we must obtain $\lambda_1 = 1$, $t_1 = \infty$ in accordance with infinite time process.

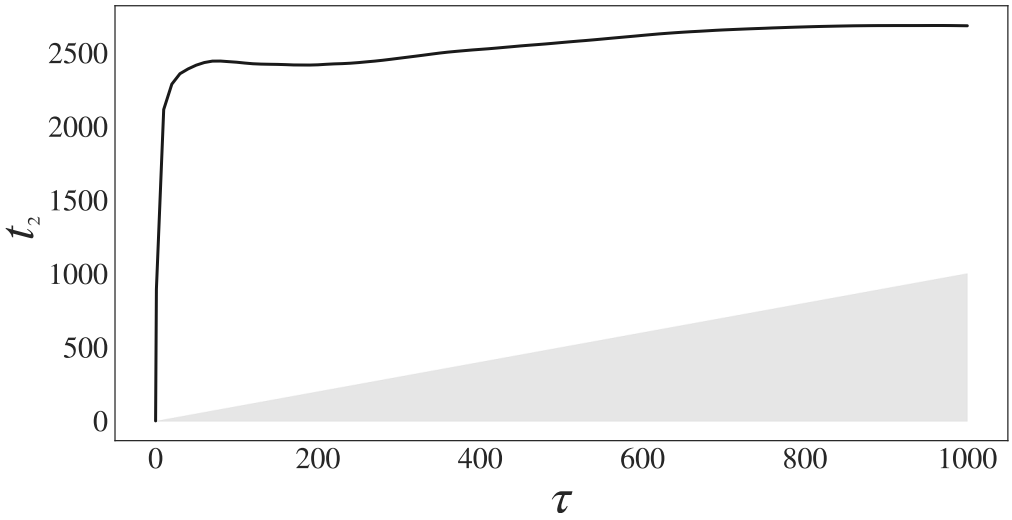

FIGURE 6. The convergence in predicted, or implied, timescales of the Markov State
Models generated from a Brownian simulations freely diffusing along a surface potential. Shaded regions corresponds to timescales predicted are less than the lag times used to generate them.

By computing the transition matrix, and therefore t_2, as a function of the lag time τ it is possible to identify the time lag at which the dynamics becomes Markovian, because when that condition is fulfilled, the time becomes independent of the time lag and levels off. This can be observed in figure 6 for the Brownian MSM describe previous in 2.4. Notice that if the predicted timescales are below the lag times used to estimate them, this is indicative of a model with too much noise.

Chapman Kolmogorov Condition

While selecting a suitable lag time is necessary for a Markov chain, it is not validation of Markovianity. When the kinetic model is Markovian, the Chapman-Kolmogorov (CK) equation

$$\boldsymbol{T}(n\tau) = [\boldsymbol{T}(\tau)]^n \ , \tag{2.21}$$

is fulfilled[58], where n is the number of steps with lag time τ. If the trajectory follows a random walk in configurational space, taking n steps with lag time τ is equivalent to taking one step with lag time $n\tau$. This is an extension of a memoryless process, where the likelihood of a transition should only be dependent on the initial state, and not on any previous.

Fulfilling the CK equation ensures that the relaxation time of a process is independent of the number of uncorrelated steps that are used to model the process. If λ_i is the i^{th} eigenvalue of the matrix $\boldsymbol{T}(\tau)$, by the CK property one should have that

$$\lambda_i = \lambda_i^n \ . \tag{2.22}$$

It is reasonable to assume that the system becomes Markovian at large enough τ, as all kinetic events become uncorrelated if they are sampled at times that differ by an interval larger than their correlation time. However, caution should be taken as MD simulations are not infinitely long, and as such numerical sampling issues can arise if τ is too long relative to the length of simulation data.

Transition Path Theory

In the course of a kinetic pathway, multiple possible states are sampled by the system during the random walk between the initial and the final state within the MSM. The final picture in MSM is often a complex network of metastable states connected by kinetic transitions[9]. This picture of a network of possible states that are populated during the kinetic process defines an ensemble of folding pathways, where one multidimensional kinetic pathway becomes kinetically most probable, and determines the timescale of the kinetic process. It is often the case that a number of paths converge to a given state, called a "hub", which becomes a high priority state as a most travelled state in the main kinetic pathway. Thus, the relevance of a state is not determined by its stability, but by the shape of the energy basin, which allows for multiple pathways to converge to the state and populate it.

The kinetics of the transition between states thus depends on the accurate definition of the coarse-grained MSM, which depends on the precise determination of the position of the borders between them, as well as on the accurate solution of the shape of the basin and surrounding energetic barriers. In counting how many times, in the course of a simulation the system transitions between states, it is important to define without ambiguity if a point in the trajectory belongs to one state or its adjacent one. It is sometimes the case that the membership of a configurational point to a state is ambiguous, for example in the regions where the energy landscape is somehow featureless, or in the regions where the statistical errors of the simulations, or lack of sufficient sampling, gives a rough landscape with no well-defined, large, energy barriers. Thus, it is useful to employ methods

that can be used to study and define with accuracy the border between metastable states.

To gain insights on the pathways associated with generated MSM we applied Transition Path Theory[64] (TPT) to calculate the dominant pathways in the coarse-grained MSM. TPT finds the reactive flux from states A to B, where a pathway is denoted as reactive if it leaves a set of states A and transitions to intermediate state i before finishing in set B. This implies that any path that returns to set A be discarded as non-reactive. This method is useful as MSM themselves can be fairly complicated due to the number of microstates necessary for adequate sampling. Additionally, as these simulations have a biological context, it is often desirable to focus on conformational states of physical significance. Here we shortly describe theory employed, however we point the reader to more detailed accounts[59, 65, 12].

We begin with a transition matrix of our Markov chain, $T_{ij}(\tau) \; \forall \; i, j \; = 1, 2, ..., N$, which denotes the conditional probability of arriving to state j from state i over a lag time τ. We adhere to the convention of a row stochastic matrix, which satisfies $\sum_j T_{ij}(\tau) \; = \; 1 \; \forall \; i \; = \; 1, 2, ..., N$. For a physical system, $T(\tau)$ must have a spectral decomposition such the eigenvalues are real and satisfy $1 \geq |\lambda_i|$, and that the eigenvectors satisfy the following for $\lambda_1 = 1$,

$$\boldsymbol{\phi}_1^T T(\tau) = \boldsymbol{\phi}_1^T, \boldsymbol{T}(\tau)\boldsymbol{\psi}_1 = \boldsymbol{\psi}_1, \boldsymbol{\phi}_1^T \boldsymbol{\psi}_1 = 1 \; . \tag{2.23}$$

An additional condition of a Markov chain is the expectation that $\boldsymbol{T}(\tau)$ with the equilibrium population for state i, π_i, satisfies detailed balance,

$$\pi_i T_{ij}(\tau) = \pi_j T_{ji}(\tau) \; \forall \; i, j = 1, 2, ..., N \; . \tag{2.24}$$

In order to define transition pathways, we define the probability that an intermediate state i reaches B after having left from A as the commitor function, q_i. As the reactive flux is defined as states that arrive to set B after having left set A, we know q_i is 0 for all $i \in A$ and 1 for all $i \in B$. Additionally, as we are modeling a time-reversible system, we can define the probability of all intermediate transitions as one minus the probability of the previous transition. Therefore, q_i can be simply written as...

$$q_i = \sum_{j \in B} T_{ij} + \sum_{j \notin A \cup B} T_{ij} q_j, \; i \notin A \cup B \; . \tag{2.25}$$

However, as we know that committor probability for sets in A and B are 0 and 1 respectively, we can reduce this equation as...

$$q_i = \sum_{j=1} T_{ij} q_j, i \notin A \cup B \; . \tag{2.26}$$

In addition to the trivial solution of intermediate state $i \in A, B$, we can take advantage of the row normalization of $\boldsymbol{T}$ to solve for the committor probability of state i arriving to B. These committors can be directly calculated from the transition matrix by solving the following system of linear equations in accordance with previous work[35],

$$\sum_{j \notin A \cup B} \left(T_{ij} - \underset{\sim}{I} \right) c_i = 0 c_{i \in A} = 0 c_{i \in B} = 1 \; . \tag{2.27}$$

The goal of TPT is to determine the pathways associated with an MSM, such that we can analyze the timescale as well as the likelihood of a specific pathway, or probability flux. We can then calculate the flux between intermediate states i to j as they progress from sets A to B as

$$f_{i,j} = (1 - q_i)\, \pi_i T_{ij} q_j \ . \tag{2.28}$$

We notice that this definition is derived by noticing that the probability that a trajectory visiting state i be within a reactive path is given by $q_i(1 - q_i)$. As such the equilibrium probability of observing a reactive path is similarly calculated as the product of the two probabilities, $\pi_i q_i(1 - q_i)$. From the probability current, we can also calculate the mean reactive flow from sets A to B as the net of equation-2.28.

$$F_{i,j} = f_{i,j} - f_{j,i} = \pi_i T_{ij} \left(q_j - q_i\right) \ . \tag{2.29}$$

The practical aspect of this framework is that in addition to determining pathways, we can quantify the probability current through them and then sort by dominance. This provides critical information about what mechanisms underlie conformational changes between key molecular structures.

Conclusion

MSM are a kinetic model generated as a network of Markov jump processes governed my a stochastic transition matrix, $\boldsymbol{T}$, which contains the probabilities to transition between discrete states. The method coarse-grains the description of macromolecular MD simulations from atomistic to a series of discrete states, which represent regions of the conformational landscape. In essence, the standard workflow for MSM both decomposes the landscape as well as models the kinetic transitions. The decomposition of the free energy landscape is performed at two levels: geometric and kinetic.

The first is the generation of a microstate model, where the full landscape is broken into fine regions of geometric similarity. Here all conformations within one microstate "near" one another in the euclidean distance of some kinetically relevant parameter space. This level description can return accurate timescales of the system and be used to model experimental systems. It however has a trade off in accuracy versus performance. In order to have a successful model, it is necessary to use a large number of microstates to sufficiently capture the conformational landscape. At the same time as the number of microstates grows so does the computational cost of generating them. This is further complicated by the spectral decomposition of a large state transition matrix is very slow, however necessary for a kinetic analysis of the MSM. Further, as the size of the states reaches above 4000, it is necessary to perform sparse linear algebra methods as a complete solution of the eigenvalue problem is no longer necessary guaranteed.

The second level is the generation of macrostate by kinetic clustering. This typically performs PCCA+ to group all microstates together by how rapidly they interconvert over a given lag time. This method originally stemmed from the PCCA method but is more robust to numerical calculations with noisy simulation data. The method solves the membership matrix which estimates the fuzzy probability of the microstate belonging to all macrostates. This corrects for errors in noisy sampling, which is limited in PCCA, and returns the most likely metastable assignments for the microstate model. The macrostate clustering in PCCA+ groups microstates such that intrastate conversions are fast, while maximizing the timescales to transition between macrostates.

These models are validated to ensure that the Markov conditions are retained. This is performed by testing the system lag time to ensure the predicted or implied

timescales converge, or in essence the kinetics of the system do not change with increased value τ. With a suitable value τ, the centroids are validated by testing for convergence in the longest non-infinite timescale predicted by the model. Finally, the Chapman-Kolmogorov condition is validated using the methods described in [47] which compares the similarity in the MSM model versus that of the simulation by evaluating the condition $\boldsymbol{T}(n\tau) = [\boldsymbol{T}(\tau)]^n$.

The theory of Transition Path Theory, which is a useful analysis tools to discover the dominant pathways within the MSMare fully detailed. This provides a mathematical framework to directly solve the flux through conformational paths from set of states A to set of states B, without having to fully enumerate all possible combinations. TPT provides additional kinetic information which can aid in a chemical understanding of the MSM.

This chapter has presented a primer on Markov State Models including the fundamental theory, practical methodology to generate, validation, and analysis via Transition Path Theory. With this review, the reader should be able to understand the working theory well enough for the following chapters.

CHAPTER III

STRUCTURAL MODEL OF DNA BASE STACKING

The purpose of this chapter is two-fold. The first, to detail the methodology used to generate simulations of DNA oligonucleotides and secondly to introduce the structural model that describes the base stacking of DNA. While the short length oligonucleotides studied in this work have few atoms, due to the structural geometry of DNA, bases stacking has several parameters with non-restricted degrees of freedom. In contrast to peptide structures where protein motion can have several residues restricted due to the peptide bond. Without these geometric restrictions, the DNA structure can undergo far more complicated conformational changes and in order to build an accurate kinetic model, a great deal of care must be taken to define suitable structural parameters of DNA over a simulation.

Such parameters must well describe the physical structure of DNA, and also capture the slow motion of the physical process to ensure an accurate MSM. While several parameters could be used in combination, this can lead to the curse of dimensionality. As the number of dimensions increases so too does the associated volume making it far harder to sample the underlying distribution. Dimensionality reduction has become a useful tool such as the case of Principal Component Analysis or time-lagged Independent Component Analysis[43, 37], however as these methods contain dynamic information, and as such structural degeneracy within a point in PCA or tICA space, this work instead chose to adopt a static structural description. This allows for a visually intuitive space which can be simply related to known DNA parameters.

Molecular Dynamics Methods and Setup

Molecular Dynamics (MD) simulations provide invaluable information as they generate a time-dependent ensemble of structural configurations with atomistic resolution. For all DNA simulations presented within this dissertation, the same computational workflow was used to ensure that all simulations were performed under the same conditions. Structures were generated using the same tools, and atomic coordinates propagated by the same force-field and water model. The starting conformations for all DNA sequences were estimated as B-form using the Nucleic Acid Builder within AmberTools[25] without hydrogens. Hydrogens were later introduced by the force-field of choice. For any simulations with replicate simulations, a starting conformation was selected from a structural minima estimated from a preliminary MSM within the first simulation. Doing so allows two things, the first being better sampling statistics by increased number of simulations, and the second ensuring that simulations are long enough that starting conditions do not effect conformational dynamics.

All-atom equilibrium MD simulations were performed for DNA structures using the GROMACS[40, 1] software package on an allocation to the Comet supercomputer within the Extreme Science and Engineering Diiscovery Environment (XSEDE). Simulations were performed using the Amber99+parmbsc0[42] force-field in explicit TIP3P water, and sufficient sodium ions were added to concentration which neutralize charges along the phosphate backbone[4]. Ionic effects are largely important to stabilize the negative charges along the phosphodiester backbone and as such cations are necessary for accurate solvent effects.

Starting configurations were prepared[19, 18, 5] by energy minimization using a steepest descent algorithm for 5000 steps. After solvation the system was heated to $T = 300$ K and equilibrated as an NVT ensemble for 500 ps with box size large enough for 1 nm distance between all atoms to their nearest wall. NVT equilibration was followed by a secondary 500 ps equilibration in the NPT ensemble using the Parrinello-Rahman barostat[16]. MD production runs were performed in the isobaric-isothermal ensemble with velocity rescaling thermostat and Parrinello-Rahman barostat, evolving atomic coordinates every 2 fs with the Verlet integrator under LINCS constraints[27]. Simulation data was then saved to file every 1 ps.

Structural Model to Base Stacking

In oligonucleotide systems, base stacking can be defined at the pairwise level among all nearest neighbor residues. While the oligonucleotides are typically small molecule, the configurational topology has several degrees of freedom due to non-restricted dihedral rotations[6, 17] along the backbone and in plane rotations.

In the generation of MSM the conformational kinetics of DNA stacking were analyzed along two kinetically relevant coordinates, or collective variables: the stacking distance (r) and twist dihedral (φ) between neighboring nucleic residues. These two parameters are standard in the definition of B-form DNA, and as such give a clear visual intuition as structures deviate away from canonical form. While other parameters, such as base tilting or ribose sugar conformations, also play a role in the definition of B-form, studies of the simulations show that changes in these parameters are fluctuations around equilibrium or occur at relatively rapid timescales.

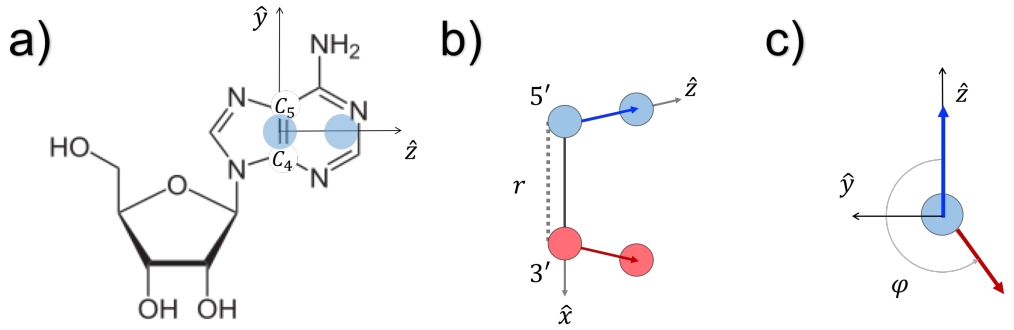

FIGURE 7. Depiction of the conformational model wherein a) the fictitious sites are placed within the plane of the base. The independent order parameters are b) the radial separation between C_4-C_5 midpoint within each Adenine monomer, and c) an aerial view, shown $5' \rightarrow 3'$ into the page, of the dihedral between the in-plane vectors.

In order to capture these two physical properties, a two-site per nucleotide model was employed. In this description two virtual atoms were placed within all nucleotides to capture the structural configuration per frame. The two sites were defined in accordance with the Devoe-Tinnocco convention[15], a reference frame which describes the relative arrangement of two nucleic residues with respect to one another. The first "site" is positioned as the midpoint between the C_4 and C_5 atoms in purines, or the C_4 and N_1 atoms in pyrimidines. The second site is given by a 1Å displacement oriented by an in-plane by a counter-clockwise rotation from the bond which defines the site 1 position (see Figure 7). This rotation keeps the second site restricted to the plane of the nucleotide while also pointing in the direction of the Watson-Crick hydrogen bond donors. Within this two-site per nucleotide model the conformational dynamics can be captured by the distance between nucleotides as the distance between site 1 across nucleotides and the torsion defined by the dihedral of the in-plane vectors. While other structural parameters exist, and are well described by in this model, distance and relative

base twist coordinates are the dominating structural components in defining how nucleic bases separated from one another at long timescales.

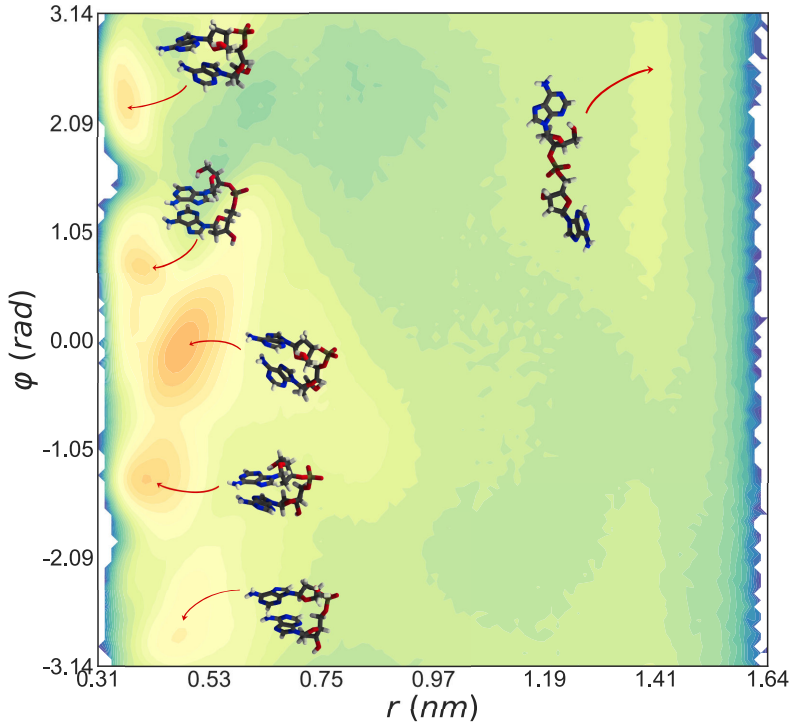

FIGURE 8. The conformational landscape of the Adenosine dinucleotide, depicting the stacking conformations sampled by MD simulations.
Regions of the landscape are labeled with conformations sampled around minima.

Over the course of microsecond simulations these parameters are sampled well showing several structural minima as shown in the case of the Adenosine dinucleotide landscape in figure 8. The landscape is labeled with structures sampled directly from MD simulations near the indicated minima, and illustrates the complex nature of base-stacking. Several stacked structures emerge indicating

right $\varphi > 0$, left $\varphi < 0$, and parallel $\varphi = 0$ orientation between stacking nucleotides. The change in sign of φ over the simulated data provides information into the how stacked conformers transition according to the chirality of the molecule.

Conclusion

In the generation of Markov state models, it is critically important to ensure that a Markov chain be estimated from relevant kinetic parameters. The assumption of a memoryless system necessitates that slow order motion of the physical process is sufficiently captured.

In the case of DNA stacking conformations, several common properties have described the conformational state of DNA as the relative arrangement of the macromolecule has several degrees of freedom. To this end it is important to select parameters that accurately describe the physical process of interest. To accomplish this criterion two parameters, the radial separation between nucleic bases and the relative dihedral orientation between the two planes that sufficiently capture these slow timescale processes. The two-site per nucleotide model presented, provides a structural description suitable for MSM estimation and of large biological significance.

CHAPTER IV

THE GRADIENT ADAPTIVE DECOMPOSITION (GRAD) METHOD: OPTIMIZED REFINEMENT ALONG MACROSTATE BORDERS IN MARKOV STATE MODELS

The following chapter presents work that was co-authored with Dr. Marina Guenza. While the bulk of the work, both the theoretical development and the associated software written, were developed by myself, conversations with my advisor Dr. Marina Guenza and labmate Dr. Jeremy Copperman provided invaluable assistance to the project.

Markov State Models (MSM) have become widely used as a convenient method to model the kinetics of complex processes as a network of discrete states[35, 38, 54, 30]. The system is treated as a random walk of uncorrelated jumps between discrete states along an underlying multidimensional energetic pathway. The dynamical process is modeled using a Master Equation (ME)[48], for a sequence of uncorrelated kinetic transitions between a network of energetic states. While MSM can use several thousand states and remain at the microscopic level it is often desirable to coarse-grain these kinetic models to gain chemical insight into molecular systems.

Several coarse-graining schemes have been proposed[36, 7, 66, 39, 31], many of which require large number of microstates for accurate decomposition[8], or relying heavily on fuzzy membership probabilities such as Robust Peron Cluster Analysis[22, 49] (PCCA+). In these cases, the number of microstates is not known a priori and as such requires the user to generate several MSM and validate an appropriate number of microstates sufficient for coarse-graining. Additionally,

in fuzzy methods, when simulation noise is too large it can be difficult to clearly identify the boundaries between metastable states. It is therefore desirable to produce a method that can coarse-grain a model while retaining a crisp partition between metastable states.

By clustering all states with fast transitions into a macrostate, one separates them from the slow transitions that occur between macrostates. When this separation of timescales between fast and slow kinetic processes is well-defined in the process, which happens for example in the transition between two states separated by a high barrier, the metastable states are easily identified. In that case, the diagonalization of the ME transition matrix gives eigenvalues that present a distinct "gap" in their sequence, and are then separated into fast and slow processes.

Most commonly, kinetic processes occur over a range of timescales, and fast and slow processes are not clearly distinct[24]. This can be easily seen, for example, when considering the fractal nature of the free energy landscape in protein conformational dynamics and folding[61]. Transitions within a state are in most cases not clearly separated in timescale from transition among states. In those cases, the application of MSM can be challenging.

To maximize the separation of timescales for transitions between metastable states, it is important to exclude the presence of large energy barriers within a metastable state. When the system transitions over the barrier inside the metastable state its dynamics is dramatically slowed down. In that case, it can happen that there is not a clear separation of timescale between processes within the metastable state and jumps between states, and Markov statistics does not apply.

In general, complex kinetic pathways do not follow a random walk in phase space, and memory function contributions are also important. However, assuming Markov statistics of the dynamical process allows for a straightforward mathematical solution of the system kinetics. The burden in the MSM approach becomes the identification of the macrostates that are energetically uncorrelated. These macrostates are necessarily metastable, in the sense that they occur along the pathway of the process, but are not necessarily the initial or final state of the transition in which the system "spends" some amount of time before it exits the energetic basin to move to another macrostate[53].

In this chapter Gradient Adaptive Decomposition method (GRAD) is presented as a novel method for accurate decomposition of the conformational landscape. The goal of GRAD is to provide an accurate definition of the energy barriers and related borders, even when the number of centroids used to define a MSM is too small to sufficiently sample the conformational topology. This is performed by refining the borders between metastable states according to the slope along the free energy landscape. The method estimates the free energy surface by histogram binning followed by noise-filtering procedure to smooth the energy distribution. Smoothing reduces the roughness present due to noise while retaining the important the dominant features of the energetic distribution. GRAD then determines the energy barriers and related border by directly incorporating information about the smoothed slope of the free energy landscape, and iteratively refines the state decomposition to maximize transition timescales.

The GRAD method improves discretization between states by maximizing the timescales associated with metastable states. This method adaptively refines the position of state barriers by randomly sampling microstates along the border

wall and then lumps each microstate in the direction of free energy surface gradient. Newly predicted barriers are accepted on the condition that the system metastability increase, where metastability is given as

$$M(\tau) = \sum_i T_{ii}(\tau) \ . \tag{4.1}$$

The trace of the transition matrix can be used as a metric of metastability as it defines the summed probability that a simulation persists within a macrostate over a characteristic lag time. We illustrate the method and the accuracy of its predictions with a number of test calculations where the energy barriers and their adjacent basins are well defined. In our examples, the barrier can be symmetric or non-symmetric. We also present the predictions of the method for Adenosine dinucleotide monophosphate, a small test molecule of biological interest. In all cases the GRAD refinement method appears to be useful in improving accuracy of crisp, coarse-grained kinetic models.

Traditional "Refinement" Workflow of Markov State Models

MSM are based on the discretization of the configurational landscape into states that need to be kinetically independent in accordance with the Markov condition. By partitioning continuous energy landscapes into discrete states, and by the projection of the continuous trajectory onto discrete state coordinates along the macrostates, the method introduces discretization errors[11]. Automatic decomposition procedures are desierable as they can improve the quality of the prediction of the MSM by tuning the number of initial microstates in which the energy surface is initially partitioned. In cases where the number of microstates

is too small, undersampling can inaccurately identify the position of the barrier between metastable states.

If the number of microstates is too high, i.e. oversampling, this can be equally detrimental as it can lead to "overfitting", wherein the procedure fits the errors present in the simulated energy landscape instead of the real border[10]. *Fuzzy* clustering methods are often used because they mitigate overfitting by allowing for overlaps between centroids in a region at the border between states. While this procedure improves accuracy in timescale prediction, crisp partitioning methods are desirable as they give a clear representation of where boundaries between states are located, bringing an informed molecular insight on the state configurations.

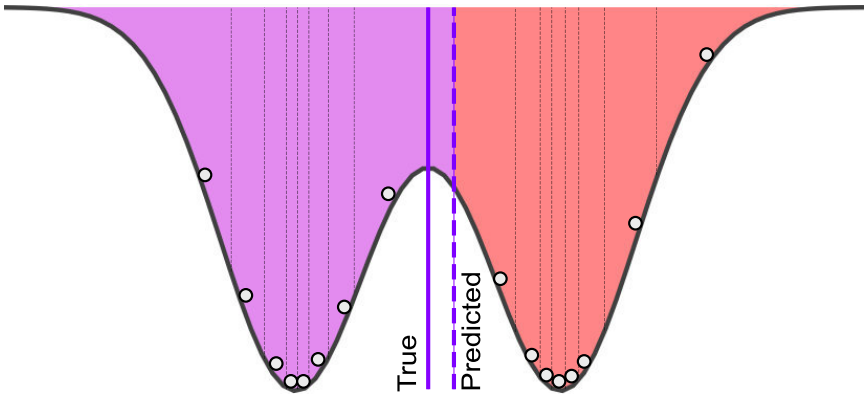

FIGURE 9. Discretization errors can be produced if the number of microstates selected is too small to adequately divide the conformational landscape. As shown here, when too few centroids are generated the predicted (dashed) separation between timescales deviates from the true (solid) barrier location.

Because the number of necessary microstates cannot be known *a priori*, multiple runs of microstate clustering are necessary to adequately define what

number of microstates are sufficient to generate a model with low discretization error[34]. In figure 9 illustrates why too few microstates can result in poor discretization in both micro and macrostate decomposition. Tyically the MSM starts with a limited number of initial centroids which is progressively increased until the predicted kinetics converges to a final timescale. If the predicted timescale (t_2) converges then the procedure is terminated. Otherwise, the procedure is repeated with an increased number of inital microstates, until the process converges. This refinement procedure is, however, computationally quite expensive. Currently, 1000 to 10000 centroids are typically used to sufficiently sample trajectories.

A couple issues add additional complexity to these traditional refinement methods. The first is the scaling of centroid generation methods. K-means++, typically implemented for the best trade-off between scale and accuracy, is still an NP-hard solution. Increasing the number of starting centroids severely affects the computational time needed to generate a MSM[3]. Secondly, having more than $4,000$ centroids requires the use of sparse linear algebra methods to reduce computational complexity. The computational time needed to perform conventional, dense, spectral decomposition for this size matrix is not only slowed, but can also fail to find an adequate solution. Sparse linear algebra methods address this but find an approximate solution to a subset of the full eigenvector space.

In the following section, GRAD is detailed, where the method addresses two criteria: 1) to reduce the number of centroids used, 2) while retaining a crisp partitioning of the energy barrier. By refining macrostates to minimize discretization error, in particular reducing error from under-fitting the kinetic

model, fewer microstates are necessary as initial input. Additionally, as the refinement identifies the barriers between metastable states without resorting to convex hull methods, we can ensure an accurate, crisp decomposition. The proposed method refines the macrostate borders iteratively and maximizes the system metastability while limiting the number of initial centroids necessary in the microstate generation.

Gradient Adaptive Refinement along Metastable Borders

Within the GRAD method, the number of initial microstates is small and accuracy in coarse-graining is achieved by the iterative refinement of the macrostates along their borders. Such a method is desirable as it limits the number of microstates and further reduces the dependence on how many microstates are generated. The method also localizes the refinement procedure only in a sub area of the total free energy space, namely the metastable borders, and as such does not require the full conformational landscape be clustered at each iteration as previous methods have proposed[21].

While traditional MSM validation increases the number of microstate centroids until the predicted timescales converges[8], our proposed method selects a small number of centroid and then refines the borders between macrostates until the metastability is maximized. As opposed to the standard method which uses large number of microstates to reduce discretization error, the gradient along the energetic landscape is used to refine the position of metastable borders. As the free energy surface is estimated from MD simulations, this only requires one calculation and does not need to be updated with every iteration.

In order to retain information from both the free energy landscape as well as state space, a lattice map is made that bins all conformational positions and assigns each bin a discrete state from a coarse-grained MSM. This allows for numerical analysis that combines the information of both the energetic landscape and the kinetic relevance predicted by Markov state decomposition. While numerical lattice methods are limited to only a few dimensions due to the computational costs and extensive memory allocation, the MSM field typically encourages the use of dimensionality reduction such as principal component analysis (PCA), time-lagged independent component analysis[43, 37] (tICA). Reducing the dimensionality identifies largest variance or slowest kinetic processes for PCA and tICA respectively, which are typically regarded as good coordinates for kinetic analysis in MSM[34, 55]. Alternatively, for smaller molecular systems the kinetic process can be described easily by one or two dimensional ordered parameters. The curse of dimensionality remains an active problem in many machine-learning applications.

A benefit to the lattice grid implementation is that periodic boundaries can be corrected when suitable for the coordinates chosen. While kinetic clustering methods deal with periodic boundaries by assigning macrostates based on interconversion, geometric approaches artificially divide barriers by how boundary conditions are defined. For example, an angular or dihedral variable where a full 2π rotation is allowed, state decomposition could separate values 0 from 2π where kinetically no such barrier exists.

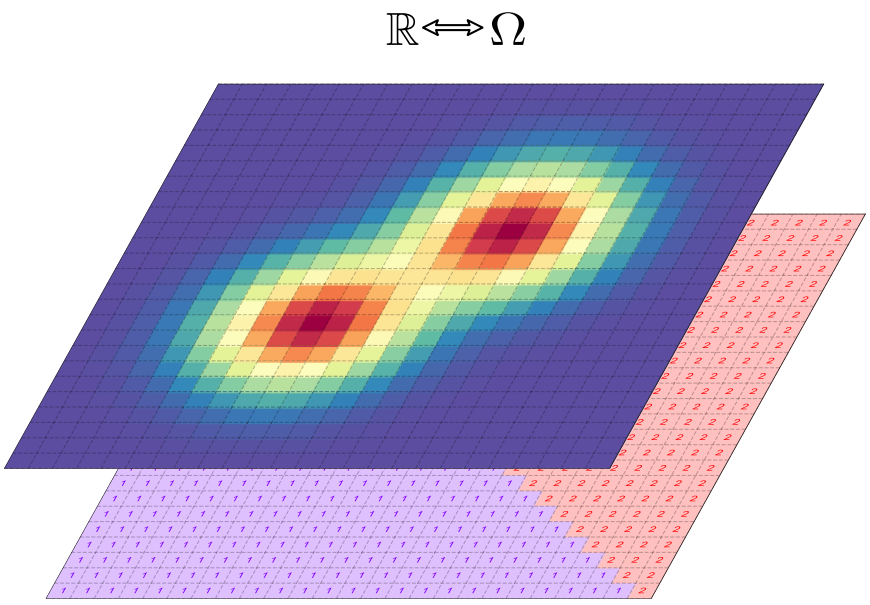

FIGURE 10. All conformations are binned onto a lattice map allowing for mapping between the energetic landscape in $\mathbb{R}$ to the state space Ω. Any information calculated on one space can therefore be transferred.

Splitting the Metastable Barriers

GRAD samples the free energy along macrostate borders by generating new microstates exclusively along the internal wall between macrostates. We refer to these as *micro-borders* so as to avoid confusion with the initial microstate seeding. Micro-borders are generated following a multistep procedure. First, the internal walls of a specific macrostate are padded with a region as shown in Figure 11. The shortest axis in the area of the macrostate is selected and the padding length is defined as a percentage of this axis length. Typically small ratios less than 10^{-1} over the course of the refinement to help ensure convergence.

Larger ratios can allow a local configurational minima in the state space decomposition to be escaped, whereas small width ratios produce a fine-scaled refinement. Once the padded region is built, centroids are randomly initialized and

micro-borders are built as a Voronoi cell restricted to the padded area. Standard uniform distribution methods are typically plagued by issues of densely packed centroids which create micro-borders of varying shapes and sizes, therefore a Poisson Disk method[26] is used to generate centroids. While other centroid generation methods exists, the computational time of this Poisson method scales linearly with the total number of centroids. Disk generation enforces that the micro-border centroids have uniform density by evaluating whether any centroids are within a radial disk from a newly proposed centroid. While the centroids are still randomly generated, they have a minimal radial separation between all other centroid positions. Without such a restriction, there would be no guarantee that the contributions from each micro-border would refine at the same lengthscale. Changes in the configurational decomposition would therefore be unevenly weighted, creating non-smooth, and eve non-physical, division between states.

Poisson disk generation uniformly places centroids with blue-noise characteristics[14] smoothly increasing the mean number of centroids as the disk radius decreases. For micro-border generation centroids are restricted to the padded region along the metastable border. Additionally, the radial separation between centroids is defined as two times larger than the padding length. This ensures that centroids are placed alongside one another parallel to the direction of the border. Uniform distribution methods provide micro-borders with even, convex shapes and as the density is uniform, it also automatically determines the number of micro-borders necessary to fill the padded region (see Figure 11).

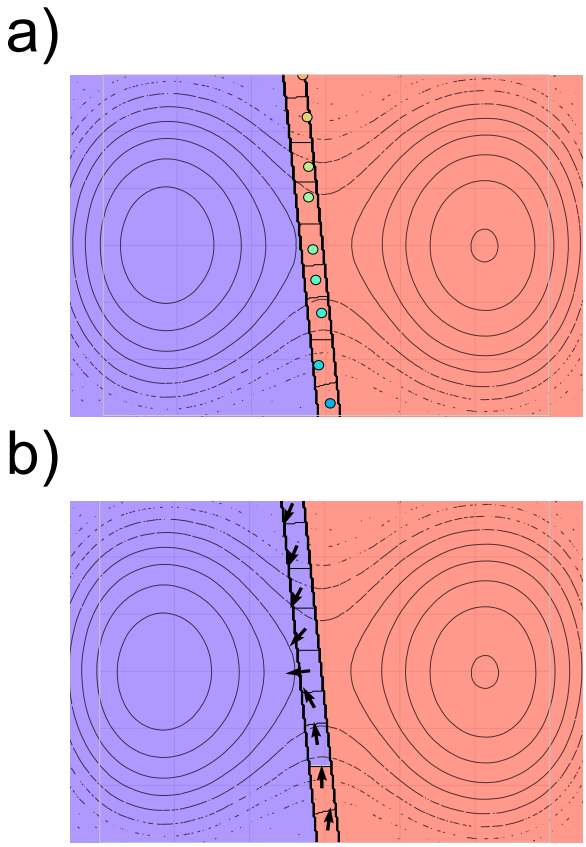

FIGURE 11. Shown in a) micro-borders were generated along the wall of an arbitrarly
defined macrostate at a fixed padding length. Each micro-border is
shown with a unique color and the centroid predicted by the Poisson Disk method.
As shown in b) each micro-borders is clustered to a
macrostate (denoted by color) assigned by the direction along the
mean gradient within each micro-border.

Assigning Micro-Borders to Macrostate

The discrete states generated are stored within the lattice map allowing for easy one-to-one mapping of surface properties such as the FES. As such, micro-borders can be clustered to corresponding macrostates by using the gradient along the FES to identify barrier maxima. The mean gradient along the FES within each newly generated micro-border is therefore used to determine in what direction the microstate should be regrouped across the border (see Figure 11). Note however, that the FES gradient is only computed once, and only the mean per micro-border is recalculated per iteration.

Each micro-border is assigned to a neighboring macrostate by lumping the micro-border in the direction of the gradient along the free energy surface. This pushes the border between macrostates closer to the barrier and reduces the discretization error from the initial MSM. Within each micro-border, the mean gradient along the energy surface is computed by a numerical gradient[44]. As these numerical methods can have error at low number of gridpoints, and MD simulations often produce sparse regions of poor sampling, a 2D Savitzky-Golay filter[51] is applied to the energy surface prior to calculating the gradient of the full landscape.

This filters statistical noise and ensures that the gradients are accurately calculated even in the presence of sparse or noisy sampling, (see Figure 12), and that the mean gradient within the micro-borders are not plagued by noise. The Savitzky-Golay filter has two critical parameters, the first is the size of the window used to fit the data with a polynomial function, and the second is the degree of the polynomial that fits the data. As the surface has a fixed number of points along the lattice, the window size is kept fixed as 10% of the total number of gridpoints. The order of the polynomial is determined by direct inspection, where the surface

is smoothed without changing the characteristic shape of the FES and improving continuity in the gradients(see Figure 12).

As stated earlier, the border between macrostates is moved in the direction of the mean gradient. The direction is defined from the median centroid of a micro-border along the state space lattice using the Bresenham Algorithm[13]. This algorithm "rasterizes" a line along a lattice (i.e. represents a line by the shortest corresponding path on the lattice) by moving step wise to discrete points that minimize the error away from the line. The rasterized line is drawn until it reaches a point outside the micro-border, and the micro-border is clustered into either the macrostate from which it originated or into a neighboring macrostate (see for example Figure 11).

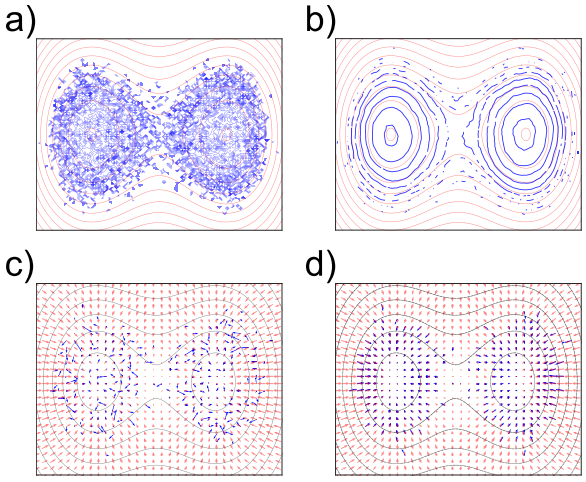

FIGURE 12. Free energy calculated from a single diffusion simulation along a symmetric two-well potential. Images left to right show the smoothing process by 2D Savitzky-Golay filter on the energy calculated from the simulation trajectory (top panels), and as well on its gradient (bottom panels). Red lines and vectors are calculated by an analytical function (noise free), while blue lines and vectors are from simulated data.

Once all micro-borders within a macrostate have been clustered, the metastability is computed. If the metastability increases with respect to the

previous iteration and within an established threshold, the new arrangement is accepted; otherwise, it is rejected, and a new macrostate is selected. A single iteration in the refinement scheme ends when all macrostates, selected in random order, have been refined.

The procedure terminates when the change in metastability for a complete iteration becomes smaller than a pre-established metastability threshold M_c, e.g. for step i we test that

$$0 \leq \frac{M_{i+1}(\tau) - M_i(\tau)}{M_i(\tau)} \leq M_c \ . \tag{4.2}$$

This ensures that all macrostates are refined, and that convergence is met not for individual, but for all states.

Validating GRAD with Ideal Model Systems

In this section two test cases are provided to evaluate the validity of the GRAD method. The first are a series of ideal Brownian diffusion simulations along a known potential. This allows for direct validation of GRAD against known results and compares between varying levels of resolution. The second is the Adenosine dinucleotide monophosphate (AA), which is a far more complex and biologically relevant molecule. While toy models are important to testing the method for success, the method must also be general and robust enough to work within the biophysical community.

Brownian Diffusion Simulations

In order to evaluate the GRAD method, a series of ideal simulations with known "states" were performed. These simulations define the potential energy landscape as summed Gaussian distributions, and allows for direct comparison

between the analytical terms and the estimated values from MSM. Brownian particle were simulated freely diffusing along a potential energy landscape. In order to evaluate the decomposition with GRAD both symmetric and asymmetric double and triple-well potentials were used. This allows for comparison with ideal cases while still regulating the complexity of the landscape.

In all simulations, the potentials were defined as the sum of N elliptical Gaussians

$$V = -10\ k_B T \sum_{i}^{N} exp[-a_i(x - x_{0,i})^2 - 2b_i(x - x_{0,i})(y - y_{0,i}) + c_i(y - y_{0,i})^2]\ , \quad (4.3)$$

where N is also the number of minima in the potential, k_B is the Boltzmann constant, and T is the temperature. The coefficients of the potential are defined as

$$\begin{aligned} a_i &= cos^2(\theta_i)/(2\sigma_{x,i}^2) + sin^2(\theta_i)/(2\sigma_{y,i}^2)\ , \\ b_i &= -sin(2\theta_i)/(4\sigma_{x,i}^2) + sin(2\theta_i)/(4\sigma_{y,i}^2)\ , \\ c_i &= sin^2(\theta_i)/(2\sigma_{x,i}^2) + cos^2(\theta_i)/(2\sigma_{y,i}^2)\ . \end{aligned} \quad (4.4)$$

In Eq. (4.4) σ_x and σ_y represent the width at half max of the curve along the x and y axis respectively; and the parameter θ is the angle by which the axes are rotated with respect to the x axis.

Each diffusion model defines a test case, where the number of macrostates is well-defined as the number of minima in the free energy surface. By easily tuning the complexity of the potential landscape, we were able to directly evaluate the accuracy of the MSM generated with and without our refinement procedure.

The single particle diffusion was modeled by a Langevin equation

$$\vec{r}(t+\Delta t) = \vec{r}(t) + \frac{\Delta t}{\gamma}(-\frac{\partial V}{\partial \vec{r}}) + \vec{r}_{random} \ , \tag{4.5}$$

where $\vec{r} = \{x, y\}$, with γ the friction coefficient, and $\vec{r}_{random}$ a random displacement obeying the white-noise fluctuation-dissipation condition. For simplicity, we reduced the energy scale such that the simulated particle had thermal energy, $k_B T = 1$. All the simulations were performed for one million time steps, and were repeated to allow the initialization from all possible minima. The potentials that were studied, as shown in Figure 13, were i) symmetric two well, ii) asymmetric two well, iii) symmetric three well, and iv) asymmetric three well potentials.

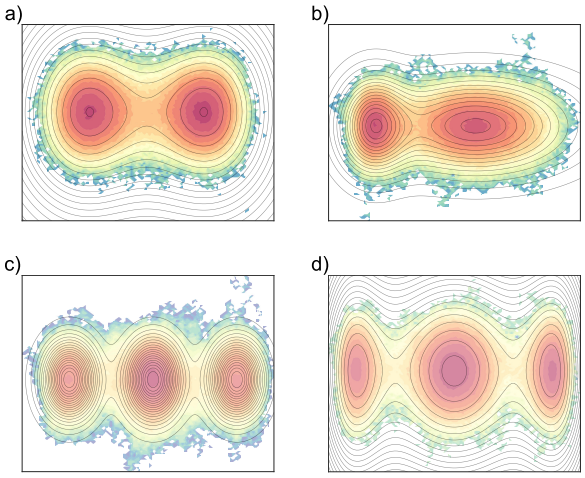

FIGURE 13. The free energy surface of the four model potentials a) symmetric two well, b) asymmetric two well, c) symmetric three well, and d) asymmetric three well. The free energy surface calculated from the analytical equation is shown as smooth contour lines, while the free energy sampled by the diffusive simulation trajectory is shown as filled contour surfaces. As the energy scale increases from red to blue, the figure indicates that the diffusive simulations preferentially sample the states with lowest energy.

Comparing GRAD Markov State Models

For all diffusion simulations, an initial 10 microstate MSM were generated, and the trajectories were clustered by PCCA+ into macrostates, determined by the number of well minima. This 10-centroid MSM analysis was refined following two different procedures. In the first, we a MSM performed with an increased number of centroids, and in the second the MSM(10) model was refined by Gradient Adaptive Decomposition. The purpose is to assess how the shape and symmetry of the potential affects the accuracy of the GRAD procedure compared to MSM with increased sampling.

As a shorthand notation we identify each MSM analysis of the simulation trajectory as MSM(N), where N is the number of microstates. To distinguish between standard MSM and those refined by GRAD, we adopt the notation GRAD(N) to indicated a N-centroid MSM refined by GRAD.

Using the conventional MSM method, we first performed calculations with microstates generated using k-means++ with $10, 500, 1000$ centroids for the four potentials presented in Figure 13. Then, to test the GRAD method, we started from the MSM(10), and refined the borders following the procedure described in the previous section. Refinement for all diffusion potentials were carried out until reaching the convergence of the metastability parameter, Eq.4.2, while the calculations were performed using the trajectories from the diffusive simulations for the four potentials presented in Figure 13.

In order to evaluate how closely the proposed refinement method correct undersampling of the MSM, we compared in Figure 14 GRAD(10) (panel d) to MSM(10) (panel a), MSM(500) (panel b), and MSM(1000) (panel c) models. Each panel in the figure displays how the MSM partitions the energy surface

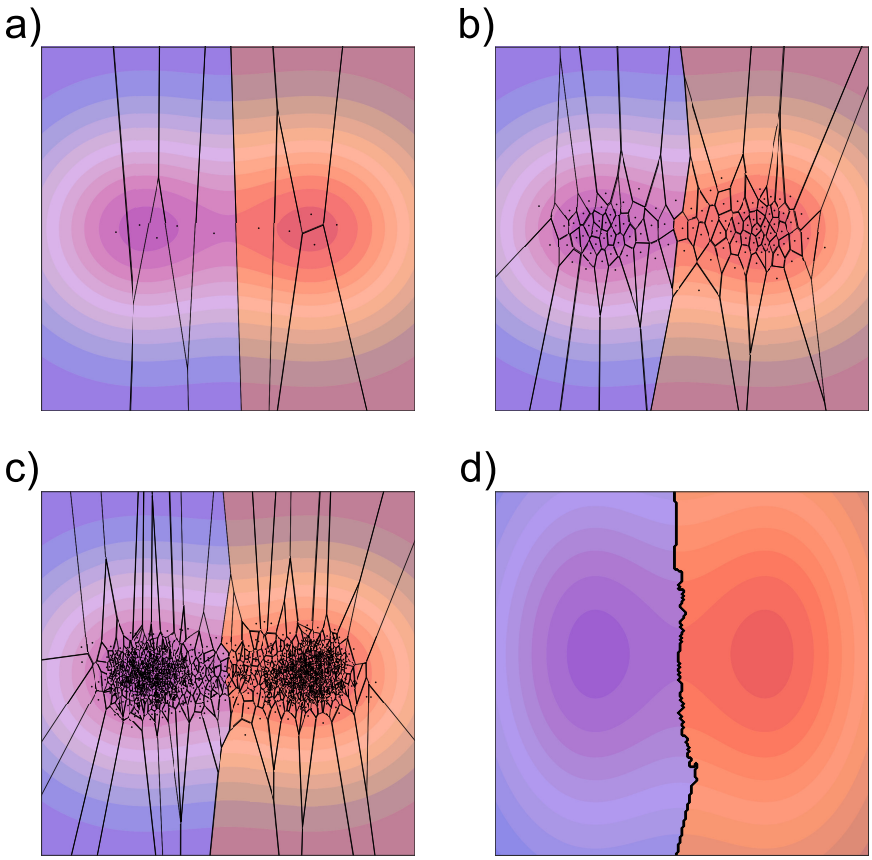

FIGURE 14. Markov state models for a) 10, b) 100, and c) 1000 centroids where black
lines represent the crisp borders of microstates, and color fill denotes membership in macrostates. Panel d) illustrates the macrostate MSM initialized by 10 centroids and refined with
GRAD.

into macrostates, given a fixed number of initial centroids, or microstates. In this representation the black-lines represent the division between states, and macrostate assignments are divided by color coding using a rainbow colormap.

Panels a), b), and c) in Figure 14 show the conventional MSM microstates grouped by PCCA+ into macrostates, after convergence to the Markovian statistics for the symmetric two-well potential. Notice that in the case of panels $a - c$ the discretization by microstates is shown as this is what defines the metastable border. In the GRAD refined method (panel d), only a single line is shown, as refinement does not require updating dependent on centroid position. At the lag time τ selected for this figure, the macrostate are optimized and do not show further modification of their areas.

The initial number of centroids in which the free energy surface is partitioned is MSM(10) in the top left panel, MSM(100) in the top right panel, and MSM(1000) in the bottom left panel. Because of the analytical structure of the potential, the exact border is well-defined and given by a straight vertical line exactly positioned at equal distance between the minima in the two wells. The figure shows that by increasing the number of microstates the resolution of the energy border between macrostates improves. The last, bottom-right, panel shows the two macrostates obtained from MSM initialized with 10 microstates (MSM model of panel a)) and refined with the GRAD along macrostate borders procedure. Even in the undersampled limit of our GRAD(10) model, the refinement of the border leads to a precise definition of the border between macrostates. Similar results are obtained for all the three other potential shapes (data not shown).

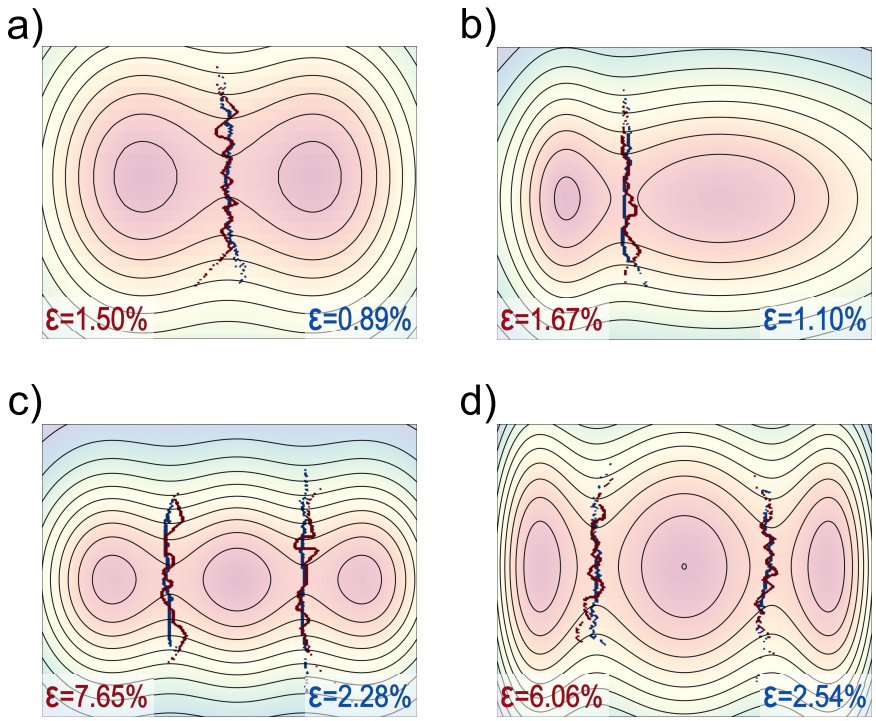

FIGURE 15. Illustration of the refined decomposition of the conformational space into macrostate for the a) symmetric two well, b) asymmetric two well, c) symmetric three well, and d) asymmetric three well diffusion models. Lines represent the crisp partition between metastable states predicted from 1000 centroids (red) and refined from 10 centroids (blue). The error, (ϵ), reported is the mean squared error predicted via harsh boolean metric against analytical barrier, for MSM(1000) (red, bottom left), and GRAD(10) (blue, bottom right).

The comparison between the GRAD(10) and the MSM(1000) is shown, for all potentials, in Figure 15. Specifically, the figure shows, for each potential, how the free energy surface is decomposed in a number of macrostates for the two refinement procedure (MSM(1000) and GRAD(10)). The border between macrostates is depicted in black for GRAD(10) refinement method, and in red for the MSM(1000). The demarcation line is crisper for the GRAD refinement method with low centroid number for both symmetric and asymmetric potentials, with two or three wells.

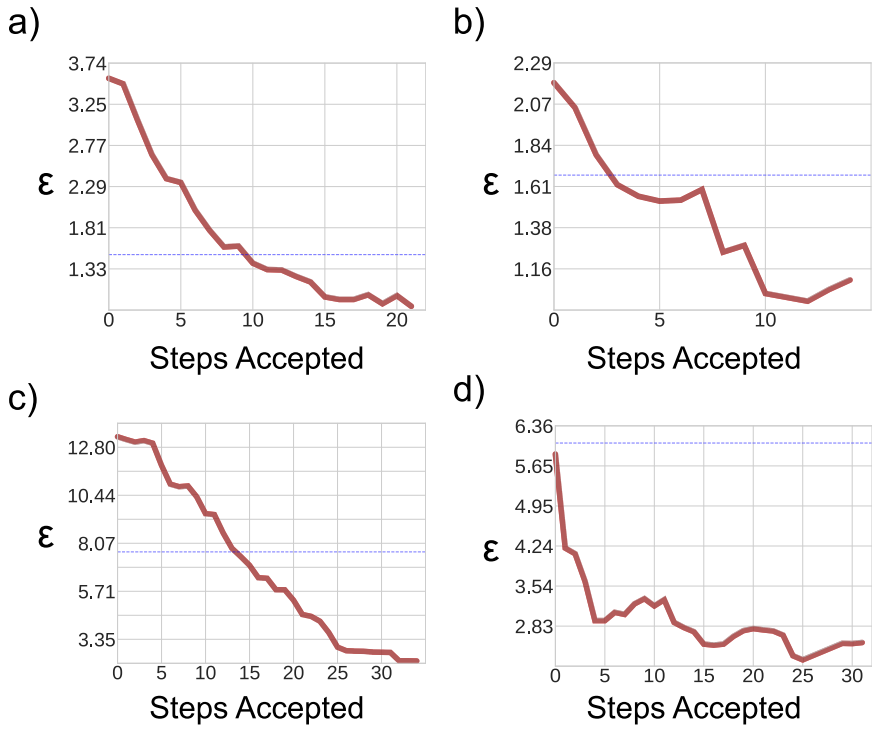

FIGURE 16. Mean squared error predicted via harsh boolean metric between analytical
barrier and GRAD(10) refinement method for all the accepted iterations (red). The error is additionally shown for MSM(1000) model (dashed blue line). While at a small number of iterations MSM(10) is less precise than the 1000 centroids MSM, with increasing number of iteration GRAD(10) method converges to a smaller error.

To further test the precision of the GRAD refinement method versus the MSM with sufficient microstate sampling we evaluated how accurately the barriers between macrostates were reproduced according to the analytical potential for which the true gradients along the barrier is easily calculated. First we evaluated the mean squared error between MSM(1000) and the analytical barrier of the potential, using a harsh boolean metric, where all matching points along the lattice receive a score of 0, and all differences a penalty of 1. The error is reported, as the mean squared error over all sampled positions, at the bottom of each panel in Figure 15.

Figure 16 reports the error per accepted iteration of the GRAD(10) method, and, as a horizontal blue line, the error of the MSM(1000) calculation. For three of the potentials, initially the MSM(10) is evidently incorrect when compared with the better sampled MSM(1000). However, as the GRAD refinement procedure proceeds, the error is reduced, and the GRAD method rapidly finds the correct energy barrier decomposition. The final predicted error in GRAD(10) is less than the error in the well sampled MSM(1000) calculation. These results are significant as they demonstrate the ability of the GRAD methods to reduce discretization error even below those predicted by undersampled centroid models.

For the asymmetric two-well potential, Figure 13 illustrates an issue with overfitting in MSM. For the asymmetric two-well potential, MSM(10) performs better MSM(1000) even in the absence of refinement. It appears that the proposed refined method can correct for both undersampling and for the discretization error of oversampling. The reason for this is that the GRAD method is directly informed by the free energy landscape at the barrier, while in the MSM approach the energy landscape enters only through the sampling performed by the centroids. In this

way the quality of the method used to sample the free-energy-landscape determines both the computational efficiency of the method and the precision of the results. One could object that the error in the free-energy-landscape, i.e. the roughness of the surface and its associated noise, could affect the local slope at the border. However, this error is accounted for by the procedure that smooths the energy surface.

In addition refining the crisp partitioning of the conformational landscape, a key aspect of the GRAD method is its ability to successfully recover timescales predicted by the kinetic model.

In figure 17, the predicted long timescales (t_2) of all refined cases, GRAD(10), are compared to the predicted times from low-sampled MSM(10) and well-sampled MSM(1000) models. For each MSM(N), the centroids were clustered in macrostates using PCCA+ at a lag time, τ. The lag time was calculated by finding the time at which the slowest relaxation time, t_2 defined in equation 2.18, converged and the Chapman-Kolmogorov condition was fulfilled. For all four model potentials we observe that the MSM converged to Markovian statistics within the time of the simulation run.

In MSM generation, the larger the number of centroids initialized during the microstate clustering, the smaller the discretization error. In the case of the simple diffusion models, we find that MSM(500) gives an accurate enough decomposition of the macrostates in most cases, as comparing models MSM(500) to MSM(1000). In its predictions of the timescale for the slowest kinetic process, t_2, the GRAD method with 10 centroids is comparable in accuracy to the MSM with a high number of centroids. Statistical errors were calculated by the reversible transition

matrix sampling algorithm. [57] In all cases, convergence via GRAD refinement produce significant t_2 values.

The diffusion models illustrate the benefit of adopting the *MSM+GRAD* refinement approach because with iterative refinement the method converges to a decomposition of the molecular landscape, successfully improving the accuracy of both under and over-sampled models. Additionally, as the refinement method is not centroid dependent, it more accurately decomposes the macrostates along the barriers because it can accurately represent non-convex shapes, while Voronoi tessellation only does so in the limit of a large number of centroids.

The results presented in this section can be explained considering that the conventional MSM method assumes that the barriers can be defined as the midpoint between centroids, which allows for the use of Voronoi cells in the procedure of energy-surface decomposition. However, if the shape of the barrier is asymmetric, the Voronoi cells procedure can introduce errors in the calculation of the kinetic. The use of fuzzy clustering methods have aimed to address this, however can reduce insights into conformations associated with metastable states. More specifically, if the number of starting centroids is small, i.e. the system is undersampled, the use of Voronoi cells introduces discretization errors because of possible underfitting the MSM. Thus, in the case of undersampled systems, our method can be useful as it provides an accurate prediction of border decomposition independent of the number centroids of the shape of the barrier.

Extending GRAD to Molecular Systems

While the ideal systems provide invaluable information, it's necessary to show that method is also robust enough to refine biological systems. Doing so provides

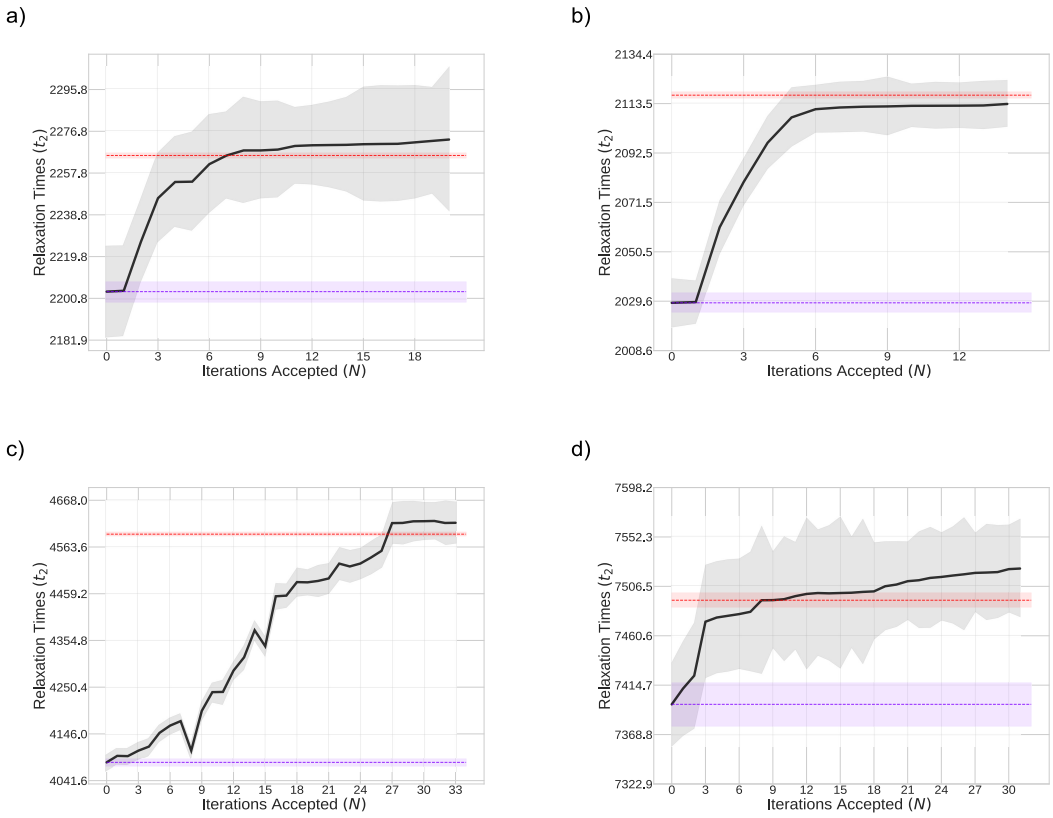

FIGURE 17. The four panels display calculations for the four diffusion potentials: a) symmetric two well, b) asymmetric two well, c) symmetric three well, and d) asymmetric three well. The t_2 relaxation times of the *MSM+GRAD* refinement approach are reported as black lines, and show how t_2 evolves per accepted step in the refinement procedure. The predicted t_2 for MSM(10) and MSM(1000) are shown as purple and red lines, respectively. Errors are displayed as shaded regions of the same corresponding color, where statistical uncertainty is calculated by the reversible transition matrix sampling algorithm[57].

added cause to the relevance of the method, and tests the GRAD methods ability to capture more complex landscapes. To this end the validation was extended to Deoxyribose Adenosine Dinucleotide Monophosphate (AA). For this system, the free energy landscape has complex features including several minima. More challenging to capture, the roughness of the landscape varies significantly, making it difficult to calculate surface properties robustly. The combination of the complexity and roughness of the landscape, provides further tests to evaluate the capabilities of GRAD to define, with accuracy, the border between macrostates. In total, 10 independent, 1 μs simulatons were performed. Each independent run used the same methodology but had different initial configurations, starting from energy minima identified from a preliminary MSM analysis of the first simulation trajectory. In total, the simulation data gave a cumulative sampling of 10 μs. The details of simulation methods are described in Chapter 3.1.

Deoxyribose Adenosine Dinucleotide Monophosphate

From the free energy surface of AA shown in Figure 18 along the coordinates previously defined, the landscape appears complex, with several minima characterized by a variety of base-stacking, which can not be trivially separated into a two-state model, based on the transition between "stacked" and "unstacked" configurations. For this complex landscape a MSM analysis of the complex transitions between states is appropriate. We generated a MSM for the Adenosine Dinucleotide using 100 starting microstates and repeated the calculations with higher resolution using 10000 centroids. We selected a lag time of 500 ps, where the system converges to Markov statistics. The free energy landscape is coarse-grained with PCCA+ to 5 macrostates for both the MSM(100) and MSM(10000)

	MSM(10)	GRAD(10)	MSM(10000)
t_2 (ps)	2484.09	3851.34	3686.19

TABLE 1. Timescale for the slowest kinetic process, t_2, in the dynamics of the AA.

(see Figure 19). The resulting times, t_2, for both MSM calculations are reported in Table 1.

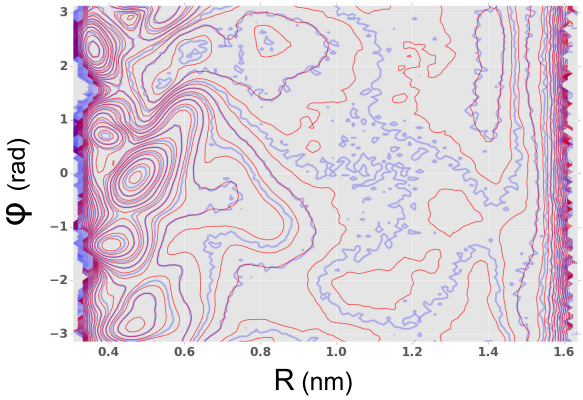

FIGURE 18. Free energy landscape calculated from simulations of AA smoothed via
2D Savtizky-Golay filter. Blue contours are calculated directly from simulation data, where as red contours have noise reduced via filtering.

The predicted MSM, while informative, required the use of 10000 centroids which is computationally costly and memory exhaustive, both in the k-means++ seeding and in the diagonalization of large matrices, which involves sparse linear algebra. Thus our refinement method could present an opportunity to reduce the number of microstates needed and return an accurate kinetic description from an undersampled MSM. A kinetic model was generated starting from MSM(100), and analyzed at the lag-time $\tau = 500$ ps. The GRAD(100) refinement was performed iteratively until convergence. The predicted slow time, t_2, for GRAD(100) is reported in Table 1, and it is found to be consistent with MSM(10000).

When exploring complex energy landscapes, convergence can be slowed down due to complications in gradient minimization as the barrier line can become locally trapped, and not further explore neighboring maxima. As the full configuration of all macrostate decompositions is too large to sample ergodically, overcoming this problem requires an intelligent exploration to find the "true" division between metastable states. This is addressed within the GRAD method by defining a padding length, which is refined using coarse sizes to extend beyond local peaks, and then finishing the refinement at finer padding lengths. The procedure is repeated for consistency: initially the padding length is set to be constant until convergence is reached for the metastability. Then the procedure is repeated with a padding length decreased by an order of magnitude. This allows for exploration of local minima, and then after several iterations, fine-scale refinement to ensure the largest increase in metastability.

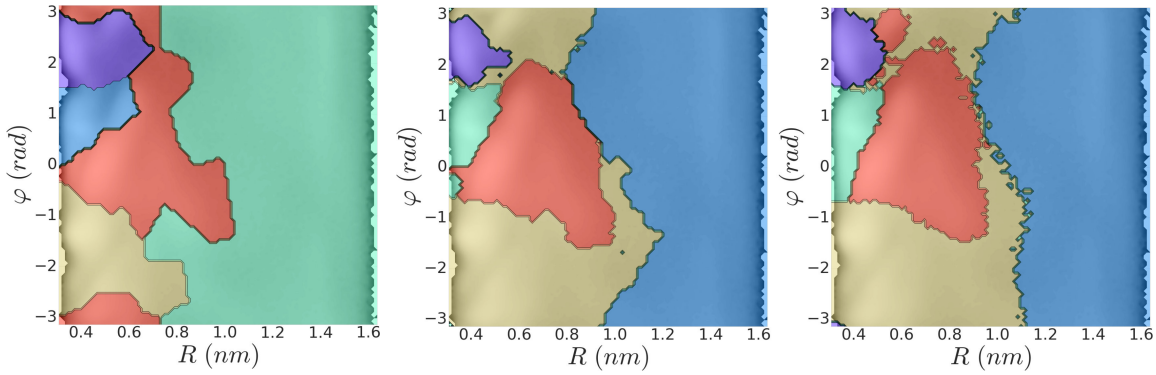

FIGURE 19. Decomposition of the free energy surface for AA, as predicted by a) MSM(10), b) GRAD(10), and c) MSM(10000).

For the AA system, a key evaluation of the refinement was to recover the landscape decomposition of a system generated from significantly more centroids, in this case GRAD(100) method was compared with MSM(10000). As this system

is non-trivially defined (as opposed to the diffusion potentials), the landscape decomposition predicted via MSM is far more challenging and as such more prone to error. While the number of AA simulations provide enough statistics to reduce sparse sampling, several regions of the free energy surface are still not sufficiently well sampled by simulations to minimize noise due to numerical error. Therefore, the energy surface is first analyzed after applying Savitzky Golay filtering, which smooths the surface and reduces error during refinement. Figure 19 shows a comparison between the two refinement methods. While the method does converge to a slightly different decomposition, it is evident that the refinement is able to largely correct the decomposition predicted by MSM(100) and produce a macrostate model that is similar to that MSM(10000) model. The mean squared error between the two refinement methods, shown in Figure 20, demonstrates the improvement capabilities of the proposed method.

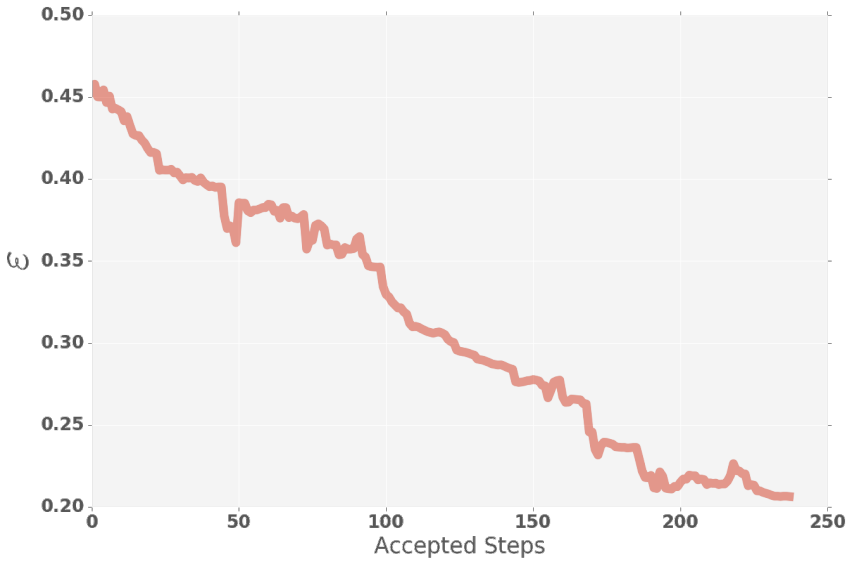

FIGURE 20. Mean squared error (ε) predicted via harsh boolean metric between the
MSM(10000) decomposition and the GRAD(100) decomposition, as a function of the accepted iterations.

Conclusion

Markov State Models are widely employed to evaluate the kinetics of transition in systems that have a complex energy landscape, by analyzing simulation trajectories of the time evolution of the system. To calculate the kinetics of the process, one has to construct the Transition Matrix in the ME formalism, and to do so one has to precisely count, during the simulation trajectory, the number of fast transitions that keep the system still inside each macrostate, and separate them from the slow transitions that occur between macrostates; the latter forming the Markovian pathway. For macroscopic models, wherein the MSM is coarse-grained using kinetic clustering techniques such as PCCA+, a crisp decomposition can be fairly difficult to achieve. As it is not known prior the necessary number of microstates to accurately decompose the landscape for a coarse-grained kinetic model, macrostate MSM are often computationally expensive and require several models be validated. In the traditional MSM the validation of the macrostate borders is performed by progressively increasing the number of microstates used to build the transition matrix, until the slowest kinetic time converges. This process is precise but computationally costly, both in the seeding procedure and in the numerical diagonalization of the sparse transition matrix of the ME formalism that is required in the analysis of microstate models.

In this chapter, we propose an alternative method, GRAD, to refine metastable borders, producing a coarse-grained kinetic model with crisp partitions. The new method, which is easily integrated in the traditional MSM workflow, starts from a MSM performed by using a minimal number of centroids, larger than or equal to the number of minima in the free energy landscape. A region along the border is then clustered into microstate borders. These micro-borders

are subsequently grouped into to macrostates by using the mean gradient along the energy landscape within the center of the micro-border to direct assignment. The decomposition and recomposition of the borders is methodically performed on all the macrostates and iteratively performed while the method checks that the overall metastability of the macrostates increases and then converges. The metastability is calculated as trace of the transition matrix, as this is the summed probability that a simulation remains within a state over time.

We find that the GRAD method is capable of producing accurate coarse-grained kinetic models regardless of the number of microstates generated, given that there is at least one centroid sampling each minima along the conformational landscape. This provides a framework for macroscopic kinetic models without having to heavily parameterize the coarse-graining method. As the method incorporates a smoothing procedure in its calculation of the numerical gradients, it is robust to complexity of the energetic surface, both in improving sparse and noisy sampling.

While the number of necessary microstates is substantially reduced, the current implementation of GRAD would still benefit greatly from computational optimization. While the method does not require substantial centroid generation and validation, nor does it require the diagonalization of large matrices, its time scaling could be improved. While for simple toy models, it is significantly faster than the traditional refinement method, in more realistic biological examples such as the presented AA, several iterations of GRAD are necessary before convergence. The method lends itself nicely to parallelization schemes such as replica exchange where the state map could be independently refined by parallel processes and accepting the configuration with the largest increase to system metastability.

The GRAD refinement method is a novel protocol that ensures the accurate decomposition of conformational space on discrete metastable states. It works well to accurately and crisply decompose the conformational landscape from undersampled MSM. The method is robust to landscape complexity with respect to sparsity as well as noise inherent in the simulation data. This is predominantly due to the implementation of the Savitzky-Golay filtering, which is used for data smoothing to reduce error from noisy as well as sparse sampling. Implementing information from surface data can prove particularly useful in reducing the effect of limited sampling of energy barriers in simulation trajectories.

CHAPTER V

CHIRAL DIRECTED MECHANISMS IN THE UNSTACKING OF DNA OLIGONUCLEOTIDES

The following chapter presents work that was co-authored with Dr. Marina Guenza. Some contributions to the presented simulations were provided by fellow labmates: Cinthia Garcia who ran many simulations of DNA dinucleotides, and Tomas Fencl who ran simulations of the tri, and tetranucleotides. Additionally, Tomas Fencl generated initial Markov State Models for the tetranucleotide simulations. The majority of the written work, analysis, and all interpretations were performed by myself.

Over the course of evolution, proteins have been modified within each generation to optimize the sequence and structure to biological functions. An interesting example of this functional evolution being the adaptation of proteins to target DNA whether by forming oligomer complexes, reading the internal content such as in replication or transcription, or in checking and correcting for replications errors. Within the scope of protein and DNA complexes, several questions still remain as to how these complexes are formed. While the influence of proteins on DNA has tremendous experimental support, an additional nucleic centric hypothesis suggests that the thermally induced conformational changes in DNA, or breathing fluctuations, play a critical role in these events. However, little is known about the exact mechanisms of DNA breathing fluctuations inasmuch how these conformational modes manifest. To address this question, the contents of this chapter investigate structural dynamics of DNA breathing fluctuations by exploring the local scale stacking of oligonucleotides. A better understanding of

the microscopic contributions serves an important role in understanding how DNA facilitates protein recognition processes.

An exhaustive study was performed on all DNA sequence pairs of dinucleotides to evaluate the effects of sequence and strand polarity on the conformational dynamics of DNA stacking. While it is difficult to extrapolate out to long-length double stranded sequences, these systems provide detailed description of the energetic minima in the absence of higher ordered perturbations such as interstrand hydrogen-bonding and long distance π-stacking between bases. Additional analysis on the conformational landscape was performed by length dependent studies to evaluate the effect of increase strand length on pair-wise interactions. These studies were performed on single stranded monophosphate tri and tetranucleotides with flanking thymine residues. This gradual perturbation by introducing weak stacking terms provides insight on how the energetic landscape is affected by increased lengths. By evaluating the preferred conformational states and the transitions between them, some insight may be gained towards conformational mechanisms of breathing modes characteristic of bubble formation or end effects.

Several microsecond simulations of DNA oligonucleotides were performed to act as a model system of the unstacking mechanism in DNA breathing. Longer length systems prove to be too difficult at feasible timescales approachable by MD simulations to sample enough conformational transitions to build an accurate model of the system kinetics. While a simulation may sample some conformational changes, there would be no guarantee that all residue positions, with sequence specificity, would undergo full unstacking. For such reasons, long-length DNA breathing is an intractable process to model. The immediate benefit to short length

oligonucleotide systems being that the small nature of the molecules allows for ergodic sampling providing detailed information of conformational states for any combination of sequences.

In order to evaluate these simulations and gain any relevant insights to the underlying conformations a two-site per nucleotide model was employed to describe the conformational landscape as detailed in chapter III. The kinetics along these structural parameters were modeled with MSM methods as decribed in chapter II clustering microscopic features with *k-means++* and coarse-graining with a combination HC and PCCA+. For chemical insight and ensuring accuracy in the analysis, coarse MSM were refined using GRAD. Transition Path Theory (TPT) was then utilized to determine the kinetic pathways and identified key structural intermediates necessary for unstacking.

This chapter presents the following study on the analysis of DNA stacking conformational dynamics evaluating the stable states and the dominant transition pathways for di, tri, and tetranucleotides. The results presented illustrate a possible chiral directed mechanism which may prove useful to investigations on breathing modes in DNA recognition events.

Kinetic Markov Model Analysis of Base Stacking

Molecular Dynamics

All-atom MD simulations were performed under Amber99+parmbsc0 force-field in explicit TIP3P water at 300K (see Chapter 3.1 for full description of MD methods). In order to capture the effects of sequence and polarity microsecond simulations were performed for all DNA sequence pairs, the details of which are described in table 2. With the exception of homopair dinucleotides, reversed

Sequence	N. Sim	Sim Time (ns)	Sequence	N. Sim	Sim Time (ns)
AA	10	1200	GA	1	1000
AC	1	1000	GC	1	1000
AG	1	700	GG	1	1000
AT	10	1000	GT	1	1000
CA	1	1000	TA	10	1000
CC	1	1000	TC	1	1000
CG	1	1000	TG	1	1000
CT	1	1000	TT	10	1000
AAT	1	1000	$TGGT$	1	500
TAA	1	1000	$TTGT$	1	500
TAT	1	1000	$TGTT$	1	500

TABLE 2. Details of DNA oligonucleotide simulations including sequence, number of
replicate simulations, and the length of simulation time. Note that all
replicate simulations were simulated for the same length of
time.

sequences were run as well adopting the standard 5' to 3' notation. For the study of length effects, series of tri and tetranucleotide sequences were run as well. In these simulations dinucleotide sequences were extended by flanking thymidine.

In all simulations, initial structures were generated from ideal B-form DNA using the Nucleic Acid Builder (NAB) tool from AmberTools[25]. Any replicate sequences built new structures by sampling conformations from minima in states identified by a preliminary MSM of the first simulation. All structures were individually solvated and equilibrated from new seeds, and production runs were saved to file every $1ps$. All simulations were processed removing rotational motion and corrected for periodic boundary effects using the built-in analysis tools in GROMACS[40, 1].

Markov State Models

In the generation of the Markov models, the time-dependent values for the radial separation (r) and planar twist (φ) were collected for all simulations in accordance with the two-site per nucleotide model detailed in chapter 3.2. This structural model, which captures the slow conformational process in DNA stacking, defines a reference frame with the first site centrally located within the nucleic acid and a second site positioned towards the Watson-Crick hydrogen bond donors. The orientation in this two-site per nucleotide model captures the distance between nucleotides as the stacking distance r and the torsion defined by the dihedral of the in-plane vectors as the twist φ.

These coordinates are biologically significant as they directly relate to B-form properties of DNA, and clearly describe the handedness in stacking which is critical in understanding mechanisms of chiral macromolecules. Further, they describe well the conformational state for base stacking and additionally capture the slow timescale dynamics of the system and for these reasons are a good candidate for kinetic analysis.

MSM were generated by clustering trajectory data using k-means++[**?** 3] implemented through the Python API Scikit-Learn[41]. This method finds a set of centroids which minimize the sum of the euclidean distance to all nearest neighbor trajectory data. The updated *k-means++* stands apart from its earlier predecessors by improved initial seeding[2]. For all dinucleotide sequences, the conformational landscape was discretized with 1000 microstates, and 1500 for longer length molecules. Timescales predicted by these MSM found converged timescales when compared to models with a larger number of microstates indicative of a good state decomposition. All simulation data were projected onto the

discrete microstates and from which the ME transition matrix was estimated according to the reversible maximum likelihood estimator described in Prinz et al[47] and detailed in chapter 2.2. A lag time τ equal to 2.0 or 5.0 ns was found to approximate the dynamics of the discretized system for dinucleotides or longer strands respectively. Models showed converged timescales for integer multiples of $n\tau \; \forall \; n \; \in \; \mathbb{Z}$. All generated MSM presented here within used standard validation methods[8, 47, 11] to evaluate convergence in predicted timescales.

Metastable states were determined by coarse-graining the microstate MSM to a macrostate description using a two step procedure. The first using Hierarchical Clustering (HC) with the Ward algorithm for agglomerative clustering implemented in Scikit-Learn which lumped microstates from 1000 to 100 (or 1500 to 150). This was then followed by Robust Perron Cluster Analysis[22, 49] (PCCA+) performed using the software PyEMMA[52] to further coarse-grain to the number of macrostates as defined in table 3.

The PCCA+ method coarse-grains the microstate MSM by solving a set of coupled linear equations which transform the right eigenvectors of $\boldsymbol{T}(\tau)$ to estimate a fuzzy membership probability. This membership assigns a probability that a microstate is grouped within a macrostate. The right eigenvectors are effectively used to determine where separation in timescale localize between microstates, by what probability they overlap, and cluster by how rapidly simulation data interconvert at the Markov time τ. Within this method fast interconversions are lumped thereby maximizing the timescale to transition between macrostates.

When the overlap between clustering is sufficiently large, or the barrier between metastable states is too fuzzy, macrostate models can be plagued by too much noise inherent in the microstate description resulting in non-physical

assignment of macrostates. This results in cases where the coarse model is too opaque for any meaningful chemical insights of the metastable states. As the barriers between stacked states were small, and unstacked states had noisy sampling, additional criteria were necessary for good coarse-graining.

The use of HC removes some uncertainty in metastable assignment by first clustering microstates at an intermediate level by how close the centroids generated by k-means++ are positioned. HC groups densely populated microstates and PCCA+ kinetically clusters these further by grouping rapid interconversions. This combination of methods effectively places a geometric constraint on how PCCA+ coarse-grains the MSM, and reduces the fuzzy overlap between microstates. As the separation between timescales in microstates is not necessarily large, and in particular for noisy regions of sampling, fuzzy memberships can be too large for meaningful coarse-graining. Whereas PCCA+ can be too "fuzzy" and HC tend to poorly distinguish closely neighboring minima, the combination of the two methods produced metastable states which separated stacked conformations from one another while incorporating kinetic information. As a significant portion of this study was to identify *how* dinucleotides unstack, seeing the transition between stacked states was largely desirable.

In order to determine the final number of macrostates, spectral decomposition of the transition matrix was performed to identify the separations in timescales predicted by the microstate eigenvalues of $\boldsymbol{T}(\tau)$ according to equation-2.18. By identifying large separations in timescales, the number of metastable states can be identified. In table 3 the full list of parameters used for the generation of MSM on the simulation data are detailed. Before any analysis was performed with these MSM they were all refined by GRAD to ensure accurate and crisp decomposition.

Sequence	N. Macro.	τ (ns)	t_2 (ns)	Sequence	N. Macro.	τ (ns)	t_2 (ns)
AA	8	2.0	9.3	CA	7	2.0	10.7
AC	6	2.0	8.3	CC	7	2.0	4.1
AG	6	2.0	7.2	CG	6	2.0	6.6
AT	7	2.0	7.4	CT	6	2.0	12.1
GA	6	2.0	9.0	TA	7	2.0	8.0
GC	7	2.0	3.3	TC	7	2.0	5.3
GG	7	2.0	11.2	TG	5	2.0	6.4
GT	6	2.0	10.1	TT	7	2.0	8.1
AAT	7	5.0	21.7	T**GG**T	7	5.0	28.9
T**AA**	7	5.0	42.8	T**TG**T	6	5.0	21.2
TAT	6	5.0	19.0	T**GT**T	6	5.0	30.7
T**AT**	6	5.0	14.8				

TABLE 3. Parameters for the MSM of all simulated DNA oligonucleotides. For tri and
tetranucleotides bolded residues indicate which base pairs were modeled.
Notice for the sequence TAT that two MSM were generated, uniquely
modeling 5' and 3' ends.

This is further detailed in section 5.3. These models provide valuable information pertaining to the kinetics of stacking dynamics. The crisp decomposition of all generated MSM can be found in appendix B. In the following section the conformational minima and the kinetic pathways between the metastable states are analyzed.

Conformational Dynamics of DNA Base Stacking

As an initial analysis of the base-stacking properties of the DNA, the structural properties are first reported for the dinucleotide systems. These can be evaluated as a model of base stacking in the absence of higher ordered perturbation, such as the case that is later discussed by length studies of tri and tetranucleotides.

As the conformational landscape share the strongest similarities based on base type, simulation data for all sequences were grouped by structural type purine (R) and pyrimidine (Y). When accounting for strand polarity this resulted in four possible combinations: RR, RY, YR, and YY. The distinction in strand polarity is emphasized by writing sequences in the standard 5' to 3' notation. In this way simulation data can be distinguished not only by base type but additionally by the connectivity along the phosphate backbone, emphasizing that the structural orientation between the two nucleic bases is inherently dependent on order.

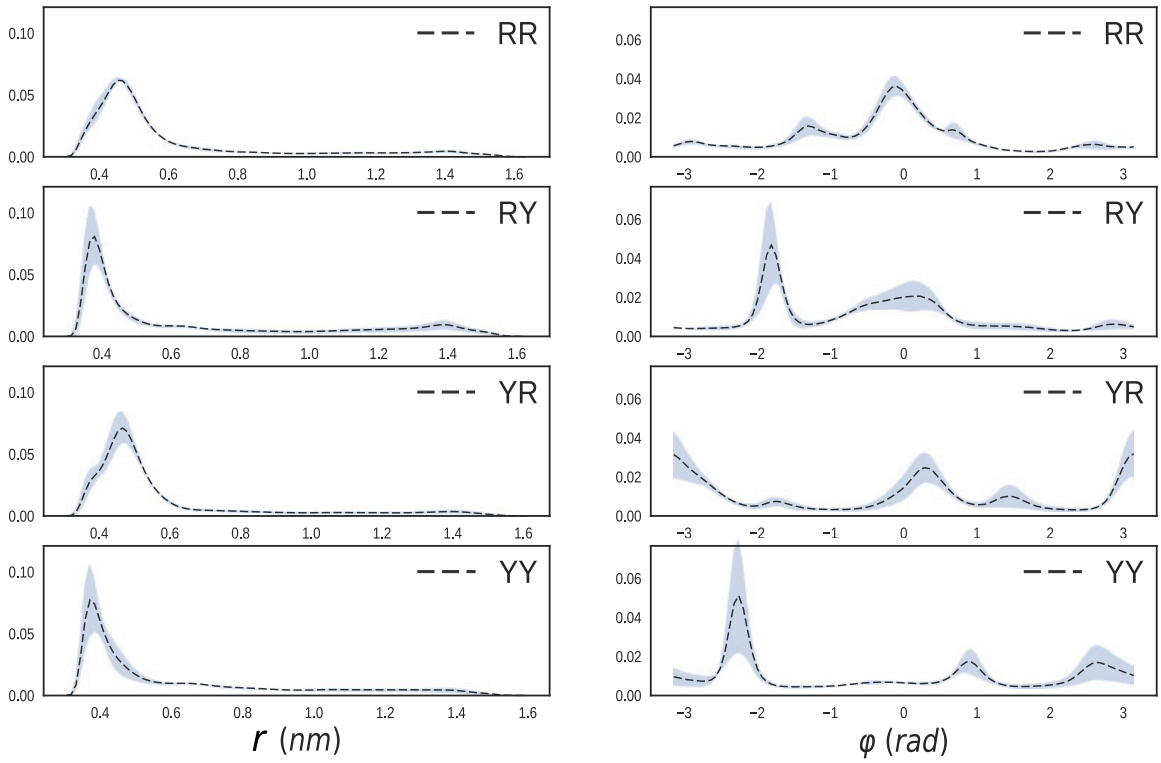

FIGURE 21. Probability distribution for Purine and Pyrimidine structural motifs along
collective variables r and φ. For example the structural motif RR contains the set of simulations AA, AG, GA, GG. Shaded region represents one standard deviation estimated by the probabilities of independent sequences in accordance with equation-5.1.

In order to compare the effect of base *type* (R or Y) in relation to the role than base *residue* (A, C, G, T) by comparing the marginal probability distributions for the structural types as opposed to their individual distribution by residue. For example, if AA, AG, GA, and GG can be successfully grouped as one system, then the marginal distribution of all four sequences should vary minimally with respect to the probability of each sequence independently. As such, in figure 21 we plot the estimated distribution of the four structural types by the independent collective variables, $X = \{r, \varphi\}$, and compute the standard deviation according to equation 5.1 quantifying by how much the distribution deviate with respect to collective variable across all individual residues.

$$\sigma^{(\alpha,\beta)}(X) = \sqrt{\langle P(X^{(\alpha,\beta)})^2\rangle - \langle P(X^{(\alpha,\beta)})\rangle^2} \quad \forall \quad \alpha, \beta \in \{R, Y\} \tag{5.1}$$

From figure 21 there is clearly a minimal deviation for all systems regardless of collective variable and base type. While the distributions are not purely equivalent, and some level of discrepancy undoubtedly results from differences across averaging residues, the peaks of all distributions are located at nearly the same position as can be seen by minimal deviation in the x direction. Additionally, the general shape of the distributions is qualitatively the same although the relative amplitudes vary when analyzing the twist dihedral. This can further be seen by evaluating the joint probability distributions $P(r; \varphi)$ shown in appendix B. Physically this implies that the stability of stacking is dependent on contacts between nucleotides. By mixing purines and pyrimidines, the bases have more variability in arranging themselves and therefore stacking interactions are destabilized.

The probability of the collective variables, as shown in figure 21 for the dinucleotide system, illustrates how structural differences can play a role in conformational dynamics. While the relative probability amplitudes, $P(r)$ and $P(\varphi)$, can vary between specific residues, the overall qualitative distribution is the same for coordinates so long as they are within the same base type. A surprising result from this analysis is that while the distributions do vary across structural groups, there is a striking similarity along the radial distributions. When comparing RR to YR we notice the stacked state is broadened with a shoulder, whereas comparing RY to YY the stacked state has a narrow monotonic distribution. The stability of the stacked states is dominated by the relative orientation along the twisting of two bases rather than on proximity. While bases can aggregate just as closely regardless of structural motif, how they lay on top of one another is determined by the number of stabilizing contacts in the 3' end. As purines are heterocyclic structures, the increased number of residues destabilizes the native stacking contacts, whereas pyrimidines favor a left handed orientation.

Additionally, when comparing the dihedral distribution between RY and YR, it is clear that polarity introduces a significant effect on the stability of base stacking as can be observed by the relative distributions which are significantly different. For all sequence pair, no symmetric operation results in a similar orientation of stack states. Further evaluation with individual sequences (appendix B) presented similar results, and is consistent with the analysis of grouped residues.

Kinetic Pathways in Base Pair Unstacking

To evaluate the kinetic model and the transition pathways between metastable conformations, the coarse-grained MSM for all sequences were refined with GRAD to convergence with tolerance $M_c = 10^{-5}$ as defined in equation 4.2. Refinement was performed under periodic boundary conditions along the φ dihedral at padding lengths 10^{-1}. The method refines the coarse-grained model, finely decreasing discretization error by maximizing the slope along the landscape, and as a consequence increasing the separation in timescales. The transition matrices were recomputed and analyzed by TPT to determine the dominant pathways.

While the underlying MSM were unique to each sequence in that all state decompositions were different, similar to the structural trends as reported in 5.2, the kinetic pathways between groups RR, RY, YR, and YY could shared strong similarities. As such the sequences AA, AT, TA, and TT are reported as a representative example of each. The GRAD-refined MSM with predicted pathways from TPT are shown in figure 22. Additionally, the full data for all dinucleotides, both the joint distributions and the MSM can be found in the appendix B.

GRAD facilitated the analysis of TPT by including insights on the conformations associated with the transition paths through crisp decompositions. While TPT analysis can be performed on transition matrix defined with fuzzy overlap, the interpretation of the conformations associated in the kinetic pathways becomes difficult to interpret. As PCCA+ was used to generate macrostates the boundaries between metastable states are difficult to assign, and refinement by GRAD improved chemical insight. This workflow provided critical information about what mechanisms underlie conformational changes between key molecular structures.

For the purposes of this chapter, the macrostate MSM of dinucleotides were analyzed with the desire to understand the transitions between right-handed stack states to fully unstacked, or in other words from small values of r at $\varphi > 0$ to large values of r for any values of φ. Analyzing the dominant pathways provided information concerning any underlying structural preference in base unstacking from any origin to target states.

In all MSM generated the majority of metastable states localized around stacked conformations of both left ($\varphi < 0$), right ($\varphi > 0$), and parallel ($\varphi = 0$) orientations. This is interesting within a biological scope as DNA bases typically have a right handed twist largely due to the ionic repulsions between the phosphates along the backbone. However, as these sequences are monophosphate backbones, no such ionic repulsion exists and the prominent configurational stability is due to favorable stacking contacts or minimized steric repulsion. At short sequence lengths and large flexibility in the phosphodiester backbone, bases were free to orient in both syn or anti states. Many sampled structures had mixed conformations in which one residues could occupy the syn and anti conformations simultaneously. The changes between dihedral and sugar orientation played a interesting role in how bases unstacked when not restricted by higher order interactions such as stacking or hydrogen-bonding.

Additionally, the unstacked region for all dinucleotides have a large radial separation of at least $1nm$, with no significant preference on the torsional dihedral φ. The large conformational entropy of these unstacked states is contrasted with the narrow well defined stacked states. Moving from large φ to small is very reminiscent of the standard protein folding funnel[23] where energetic states are

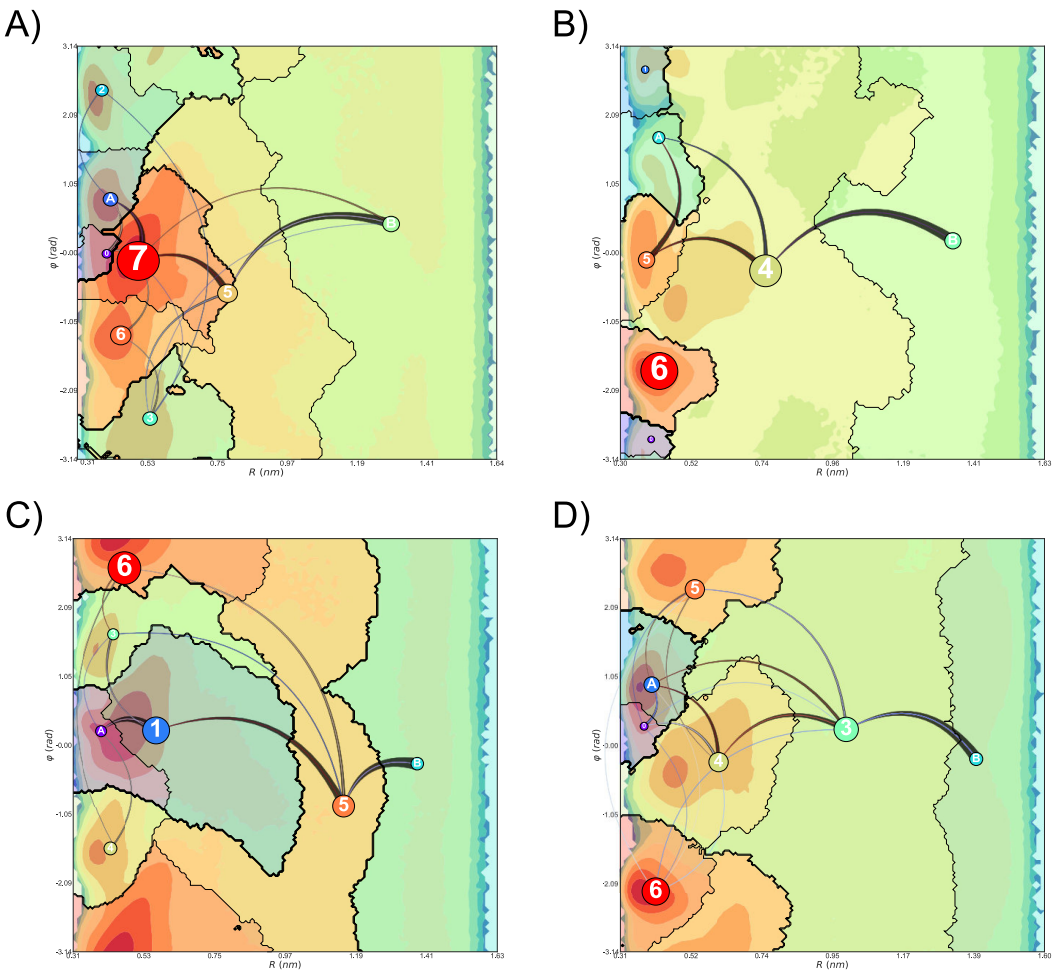

FIGURE 22. The crisp decomposition predicted by MSM+GRAD of the free energy
landscape for a) AA, b) AT, c) TA, and d) TT. The states are labeled according to their macrostate assignment indexed from 0 and color coded violet to red. Transition pathways start from macrostate A and end in B. The relative size of state labels represents the equilibrium populations $\boldsymbol{\pi}$, and the relative with of the arrows represents the net flux likelihood.

favorable, despite a decrease in entropy, by favorable enthalpic contributions in the stacked configurations.

To gain insights into the DNA unstacking mechanism, the transition pathways between B-form like stacked states to the fully denatured region were analyzed, labeled as states A and B respectively in figure 22. The B-form characteristics are defined according to separation and twist between the bases, however within these states base-flipping events occur for both 5' and 3' residues, and for this reason structural analysis was performed within each conformational state. While the state decompositions were unique per sequence, a consistent trend emerged in that bases underwent a left handed (counter-clockwise) rearrangement before a large radial separation characteristic of unstacking. While all sequences could, and did, undergo a right handed twist, the likelihood of such a mechanism was substantially reduced.

In YR and YY sequences, such as TA and TT, where a metastable state exists at $\varphi = \pi$, it was not necessarily clear whether transitions from $A \rightarrow i$ underwent a left or right handed transition. To elucidate this mechanism, any metastable state that was divided by this boundary, was split into a left and right handed state. The TPT analysis was performed then to determine whether states at the periodic boundary entered first in left handed or right handed orientations. In these cases, the right handed transitions had slightly larger probability to commit along the pathway than left handed. However, in pathways were right handed transitions occurred first they generally continued rotating to a left handed state before terminating at the unstacked state B. For sequences where right handed pathways did emerge, barriers were too large to allow for a direct transition to unstacked states and had to rotate further into left handed orientation which more easily facilitated the separation between bases.

A significant aspect to the stacking transition emerged as a consequence of the relative orientation of the base to the deoxyribose sugar. Base or sugar flipping effectively acted as a sufficient reorganization allowing or prohibiting transitions to unstacked states. While the syn and anti configuration play a role in the mechanism, the timescales were significantly faster than that of the change in the distance and twist of base stacking, and as such were not necessarily slow enough to be kinetic parameters for the MSM.

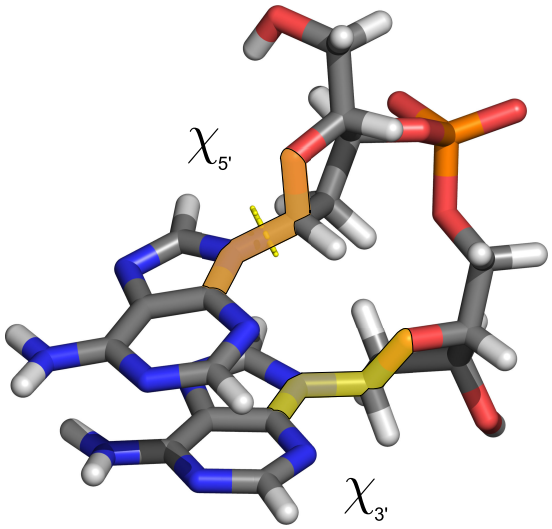

FIGURE 23. The χ dihedral which measures the relative orientation of the plane of the
nucleic base relative to the orientation of the deoxyribose-sugar. The
atomic sites $O_{4'}, C_{1'}, N_9, C_4$ within the AA moleculare are labeled orange
and yellow for 5' and 3' dihedrals, respectively.

To measure the effect of base flipping, the standard χ dihedral was measured time-dependently for all simulations. This dihedral measures the relative orientation of the plane of the nucleic base relative to the orientation of the deoxyribose-sugar. This was computed for both 5' and 3' residues using atomic

sites $O_{4'}, C_{1'}, N_9, C_4$ for purines, and $C_{1'}, O_{4'}, N_1, C_2$ for pyrimidines, which is shown in figure 23 for the AA dinucleotide.

The structural analysis of the dihedral orientation evaluated the relative orientation of the sugar with respect to the base and was compared in relation to the standard syn (60^o) or anti (260^o) dihedrals. Due to the limited conformational restrictions in the dinucleotide structures, molecules were able to stably occupy non-standard conformations. The space of 5' and 3' χ dihedrals are shown in figure 24 for the AA dinucleotides. In dinucleotide simulations all combinations of syn and anti were allowed however with the trend that RR mostly occupies syn-anti, RY syn-syn, YR anti-anti, and YY anti-syn. These conformational states emerge from non-restricted motion, and stabilized as they reduce the steric interactions among side groups.

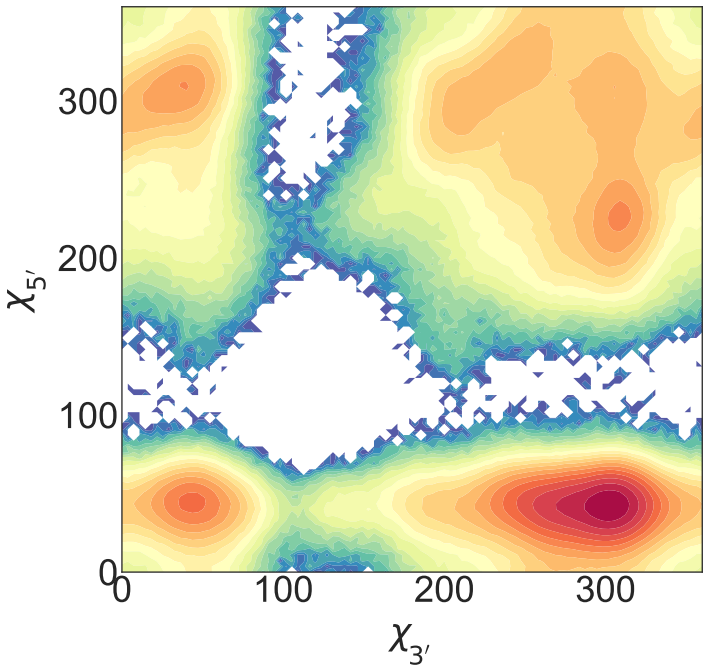

FIGURE 24. The free energy surface for the 5' and 3' χ dihedrals.

To further study the effect of base flipping the conditional transition probabilities of the χ dihedrals were analyzed with respect to transitions between states in the MSM. For all states that transitioned between states $i \rightarrow j$ over a lag time τ the conditional probability for 5' and 3' χ dihedrals were given by $P(\chi(t+\tau); j|\chi(t); i)$. Evaluating the likelihood of χ of each respective residue before and after specific transitions illustrated the role of base flipping on conformational changes. As an example of this the probabilities for AA is shown in figure 25 for three cases in both the 5' and 3' residues: the right handed $1 \rightarrow 2$ transition, and the left handed $1 \rightarrow 6, 7$ transitions. Evaluating the χ dihedral in relation to the base stacking helps elucidate the preference towards left handed transitions.

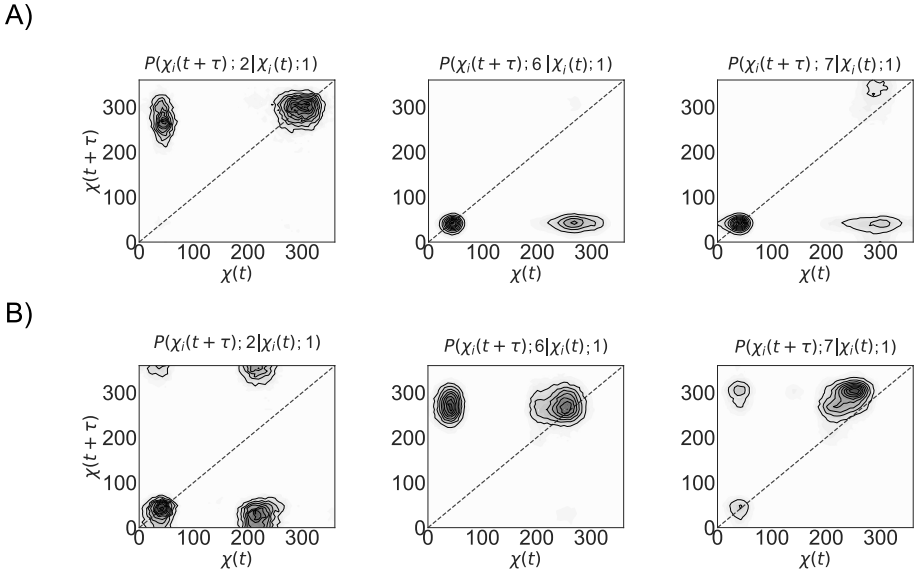

FIGURE 25. The conditional probability of χ transitions from states $i \rightarrow j$ over lag time τ given as $P(\chi(t+\tau); j|\chi(t); i)$ for dihedrals for A) 5' and B) 3' residues.

The conformational change associated with right handed transitions required a reorganization of the $\chi_{5'}$ and $\chi_{3'}$ dihedrals. As the 5' residues in AA were most stable in syn conformations, a comparison as shown in figure 25 illustrates the

necessity for a base flipping event. The 5' base would need to undergo a syn $\rightarrow$ anti transition which was less likely. In contrast, the left-handed transitions were typically found in the syn conformation and did not necessitate any further flipping mechanism allowing for an easier transition to left-handed states. It was evident from these results that the right handed transitions required an intermediate base-flip however base-flipping within these states were more restricted. Alternatively, in many of the left-handed states, several syn and anti combinations were accessible, allowing for faster transitions in and out via left-handed states.

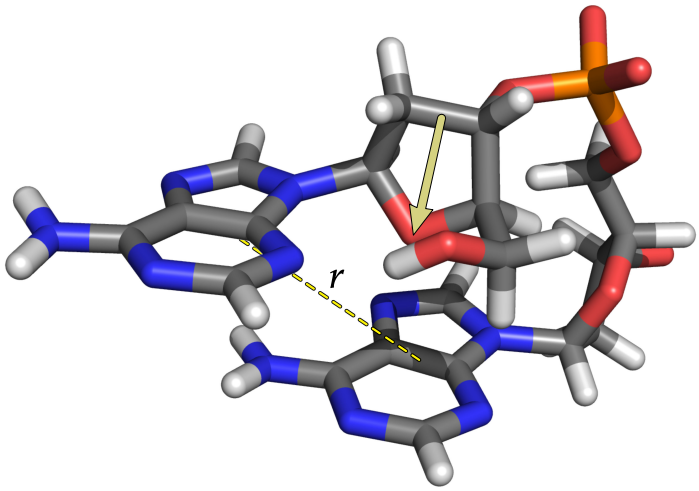

FIGURE 26. The rearrangement of the 5' deoxyribose sugar, labeled by the yellow arrow, allows for a stabilized offset of the 5' residue off and over the 3' base.

To further explore this, the evaluation of the structural changes also showed that left-handed transitions were aided by a reorientation of the ribose-sugar facilitating the $1 \rightarrow 0, 6, 7$ transitions by pushing nucleic bases off and over one another without having to significantly increase the radial separation as shown in figure 26. From this structure, the $\chi_{5'}$ rotation is significantly less restricted as the number of steric contacts is reduced. This likely explains why barriers to

left-handed states are shallow and why the landscape topology has a wide basin for parallel stacked states where $\varphi = 0$. In contrast, right handed structures the rotations about $\chi_{5'}$ and $\chi_{3'}$ dihedrals were restricted. This may explain the large barrier that is present between the transitions from states $1 \rightarrow 2$ producing kinetically slow transitions in and long-lived escapes.

In general, the cases of RR and RY as the the 5' base is a purine, and therefore with more possible steric interactions, we found that the rearrangement acted as a kinetic trap with long timescales before escape. In the cases of YR and YY, the pyrimidine bases seem more able to undergo this $\chi_{5'}$ rotation relatively easy, which might explain why right handed transitions occur more readily for these structures.

Length Dependence on DNA Stacking Landscape

When evaluating the dinucleotide systems, as done in the previous section (5.3), the kinetic pathways of DNA stacking were explored for all pairwise sequences of DNA dinucleotides. This provided insights into the local contributions of nucleotide stacking on the global structure of DNA. DNA in biologicalsystems however, in addition to interstrand contributions in duplexed form, has more than just short ranged contributions from nearest neighbor base stacking. From layered π-stacking, ionic repulsions along the phosphodiester backbone, and increased solvent exposure; length plays a significant effect on the intrastrand stability. Several questions still remained in how the increased stacking contributions effects the conformational landscape of DNA base stacking, and how those contributions play a role in the chiral directed mechanism in dinucleotide systems.

In the two residue stacking mechanism, conformational dynamics were dominated by left over right handed transitions. When further explore the chiral directed pathway found a large contribution to this phenomena originated from structural dependence of the deoxy ribose sugars. Reorientation of the DNA base in relation to the deoxyribose sugar were critical in stabilizing intermediate states and facilitating several stacking conformations at fast timescales. Whereas transitions to right handed states were limited due to steric interactions preventing transitions between syn vs anti conformations, left handed transition underwent a sugar reorientation which reduced unfavorable base contacts allowing for quick transitions. A great deal of these mechanisms however were undoubtedly facilitated by the high degree of freedom in dinucleotides as no additional residues contributed to stacking. This raised the question, what happens in the introduction of higher order terms, such as when strand length is increased? In order to further investigate the stacking landscape, additional simulations were performed on longer length tri and tetranucleotide sequences.

An underlying expectation to these results was that as the length of DNA increased B-form parameters should become a more stable structure along the conformational landscape and the variability in base flipping should be reduced. As the length of the DNA strand is extended, the number of stacking interactions along with ionic repulsions between phosphates along the backbone, shift the energy landscape to favor B-form or more generally right-handed conformations. With this restriction, bases would be unable to easily facilitate the sugar reorientation that allowed for transitions between syn and anti conformations. To explore length dependence on base stacking, simulations of tri and tetranucleotides were performed on a subset of the dinucleotide sequences studied previously and

were extended with flanking thymine bases the parameters of which are detail in table 2.

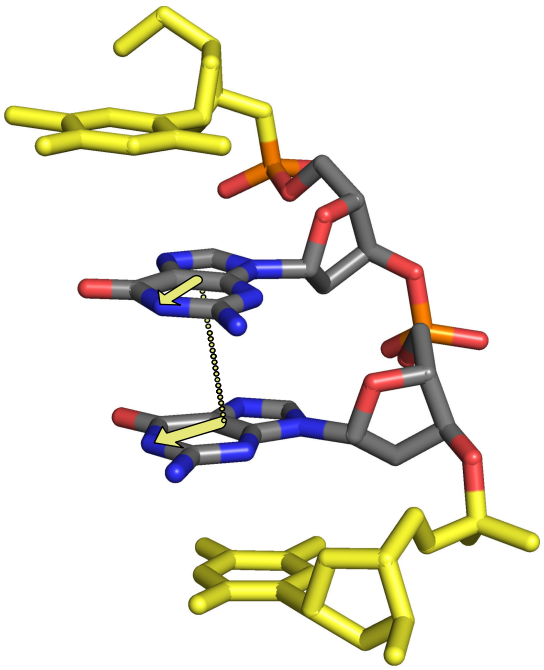

FIGURE 27. The structural two-site per nucleotide model for Guanine-Guanine with
flanking thymine residues. The flanking T are highlighted in yellow, with distance between sites and in plane vectors (as described in chapter 3.2) shown.

The two-site per nucleotide model was evaluated on internal bases, with implicit contributions from flanking T residues included from simulation data as shown in figure 27. By extending sequences in this manner a consistent perturbation to the DNA unstacking mechanism, where the stacking stability of the two residues could be evaluated under influence of increasing length. Additionally, by adding flanking T bases, the structure is only weakly perturbed due to the known weak stacking stability of T residues. Extending sequences by purines such

as A and G could have over stabilized the strands making transitions more difficult to sample.

While trinucleotide sequences had A residues and tetranucleotides G residues, the analysis of the free energy surfaces for dinucleotide systems shows minimal differences as shown in the supplemental content B, and as such we evaluate their trends. The general trend observed was that as the length of the oligonucleotide increased the conformational landscape shifted, deepening the energetic minima in right handed orientation and decreasing the likelihood of left-handed conformations.

Trinucleotides

The conformational landscape of the four trinucleotide models is depicted in figure 28 below. The sequences for AAT, TAA, and TAT are shown modeled by the two-site per nucleotide parameters. To evaluate the effect of thymine perturbation for a dinucleotide sequence XX, the structure would simulate either TXX or XXT. This method allows for studies of polarity specific perturbation on the conformation stacking dynamics. For example, the AAT sequence was modeled as an AA model with a perturbation introduced at the 3' thymine. Therefore, in AAT and TAA the stacking for 5' and 3' AA were evaluated as individual models. Similarly the TAT system was analyzed for TA and AT as 5' and 3' slices. This framework allowed for the effect of the flanking thymine to be directly compared with respect to the dinucleotide systems.

Observing the complex landscape, while many of the populated minima that were observed were reminiscent of the dinucleotides, some critical differences provided interesting biological implications. The first noticeable difference, was between energetic favorability between left, right, and parallel orientations.

Whereas in the case of dinucleotides, the left handed and parallel orientations were more energetically favorable, in the trinucleotide these configurations were far less sampled. While some conformational minima were identified with left-handed characteristics (such as AAT and TAA), in others these states were very weakly sampled by simulations, and had no discernable distribution about a conformation (TAT).

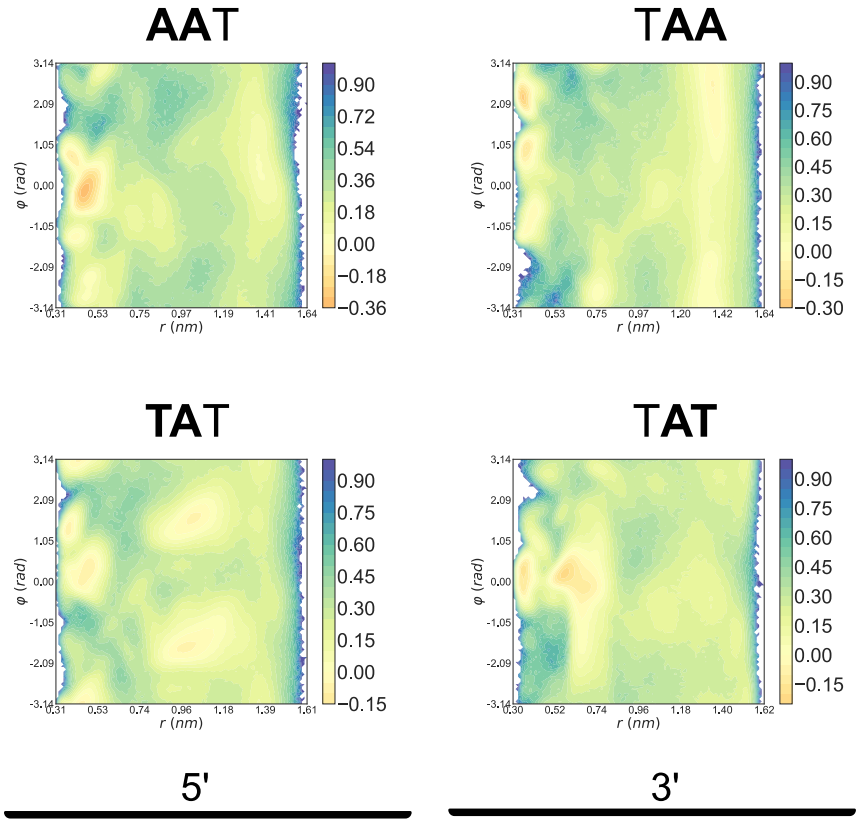

FIGURE 28. The conformational landscapes for trinucleotide systems.

An additional difference in relation to the dinucleotide well size, was the right-handed states for trinucleotides clearly stabilized. While this was observed for all cases, the manner was unique to each sequence. AAT still had parallel stacking as the most likely conformations but B-form characteristics emerged as the next

most sampled conformations. TAA had no left or parallel stacking and had a strong likelihood of sampling right handed forms. Interestingly enough, in all sequences unstacked conformations had well defined minima. A comparable radial separation was found in the trinucleotide unstacked states as the dinucleotides was observed, however these long valleys were more energetically favorable.

A final difference between was a prominent feature in the TAT sequence. Both the 5' and 3' models had "weakly", or semi stacked structures. The use of weak here points out the conformations that had a distance parameters $0.5 < r < 1.2$ nm. In both the TAT models the molecules stabilized in intermediate conformations. Whereas the 3' case the stacking distance was relatively close, the 5' case had a stable distance of over 1 nm. These conformational states depict how length effects can drastically change the complexity of the structure. For this distance range, it's not entirely stacked, as the traditional cut-off for measuring stacking contacts is below this value.

Further, when compared to unstacked states in all other sequences, the distance were too small to be regarded as unstacked. When viewed structurally, these conformations illustrate the ability for nucleic bases to stack along the side of the DNA strand as opposed to directly on top of it. As the DNA strand increases in length, undoubtedly so does the possibility of "weakly" stacked intermediate states, that effectively get trapped between standard stacking and unstacked states.

Tetranucleotides

Tetranucleotide sequences demonstrated a significant change in the conformational landscape (see figure 29). While in the trinucleotides there was a decreased sampling of left-handed states, in the tetranucleotides this sampling

was virtually non-existent over the course of the 500 ns of simulation time. In the sequences analyzed the number of "strongly" ($r < 0.5\ nm$) stacked states were substantially reduced and several "weakly" stacked states similar to trinucleotides dominated. The absence of these states indicated the existence of very large barriers and thereby restricting transitions away from the initial B-form configuration. The stacked states observed in the dinucleotide systems were most likely a result of the large degrees of freedom, and with the flexibility of the sugar groups to reorganize (when internally flanked by residues) the barriers to transition between stacked states were largely reduced.

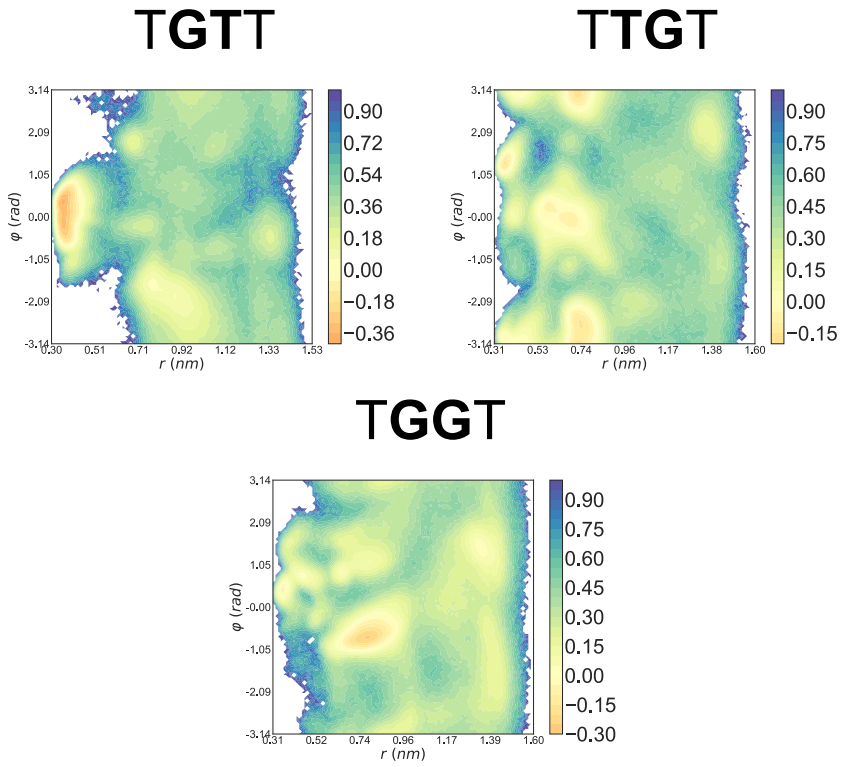

FIGURE 29. The conformational landscapes for tetranucleotide systems.

In general, a prominent feature of the tetranucleotide sequences was the multiple available stacked conformations observed in shorter strand become

destabilized. While several new minima were populated, stacking at short radial separation typically only populated one minima. While in TGTT and TGGT two minima could emerge in this radial range, only one minima was strongly sampled. In addition, the analysis of the conformational surfaces indicated that structural minima stabilized in the weakly stacked states. These states were reminiscent of the TAT sequence in which some kinking or lifting of the bases introduced a non-canonical structure that could persist over the course of the simulation. With the introduction of increasing number of thymine residues, the more significant the change on how DNA bases unstack. As mentioned in previous sections the parallels to the protein folding funnel were pronounced. As the strand length increased the conformational entropy minimizes, and the structures all "funnel" into a single stacked conformation.

Evaluating Length Dependence on the Unstacking Kinetics

As before, all kinetic models (parameters shown in table 3) were refined by GRAD and then analyzed by TPT. From the conformational analysis (see section 5.4), the free energy landscapes shifted due to the introduction of flanking T residues and as such some minima observed at shorter strand lengths were not sampled in MD. This raised the questions as to how might the kinetics between these states change. As the intermediate "weakly" stacked states had no direct parallel with regards to the dinucleotide system it was not immediately obvious whether or not the patterns observed with dinucleotides would propagate out to longer length systems. While the strongly stacked states were dominated by right handed characteristics, weakly stacked states provided new nodes along chiral pathways.

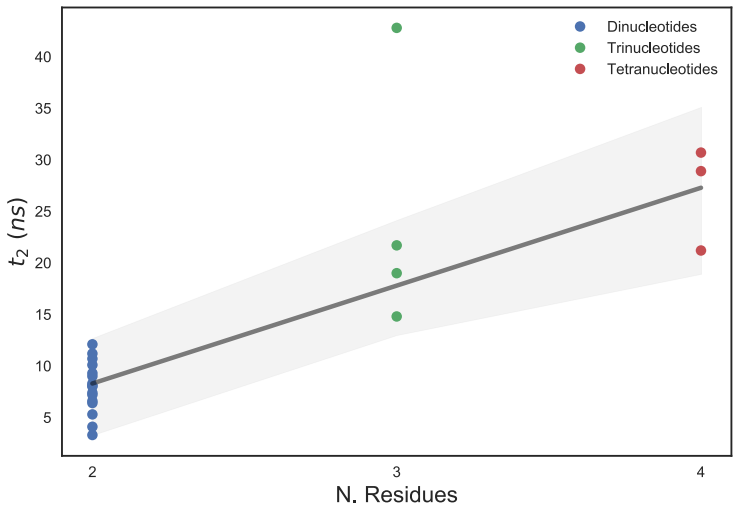

FIGURE 30. The slowest timescale predicted by the MSM were fit by linear regression
shown as a black line, and 95% confidence intervals are shown shaded in gray estimated as a normal distribution about datasets. Fit was obtained by discarding the t_2 predicted by TAA as this was well outside the standard deviation of the trinucleotide set.

An immediate interest was in evaluating the effect on timescales as predicted by the MSM for all molecules. Comparing the timescales as reported in table 3, as might be expected the amount of time for the slowest process (t_2) of the system, increased with increasing length. These timescales were evaluated by spectral decomposition of the microstate transition matrix. With increasing length the jumps between discrete states were showed an increasing trendline.

The implied timescale by length are shown in figure 30 and were fit by linear regression. While the data for all sequences fit this trend well, the TAA molecule was found to be outside one standard deviation from the trinucleotide dataset and therefore was excluded from fitting. With this the only exception, the data otherwise followed a linear trend with increasing residues staying within a 95% confidence interval. The change in timescale by residue increased rather slowly and showed a large overlap between varying lengths, indicating the dynamics may have behaved similarly but slowed down with complexity. Biologically this was in line with the expectation as strand length should act to augment the complexity in the conformational landscape. As the volume of the configurational space increases the number of possible intermediate states thereby slow transitions between stacked and unstacked conformations.

The landscape of the tri and tetranucleotides illustrated several weakly stacked configurations, and when refined with GRAD many of these features were isolated as metastable states. In dinucleotide structures while some features could be observed in unstacked regions, the metastable discretization typically separated stacked and unstacked conformations indicating a single barrier to unstacking. Longer length sequences however illustrated a far more complex network of

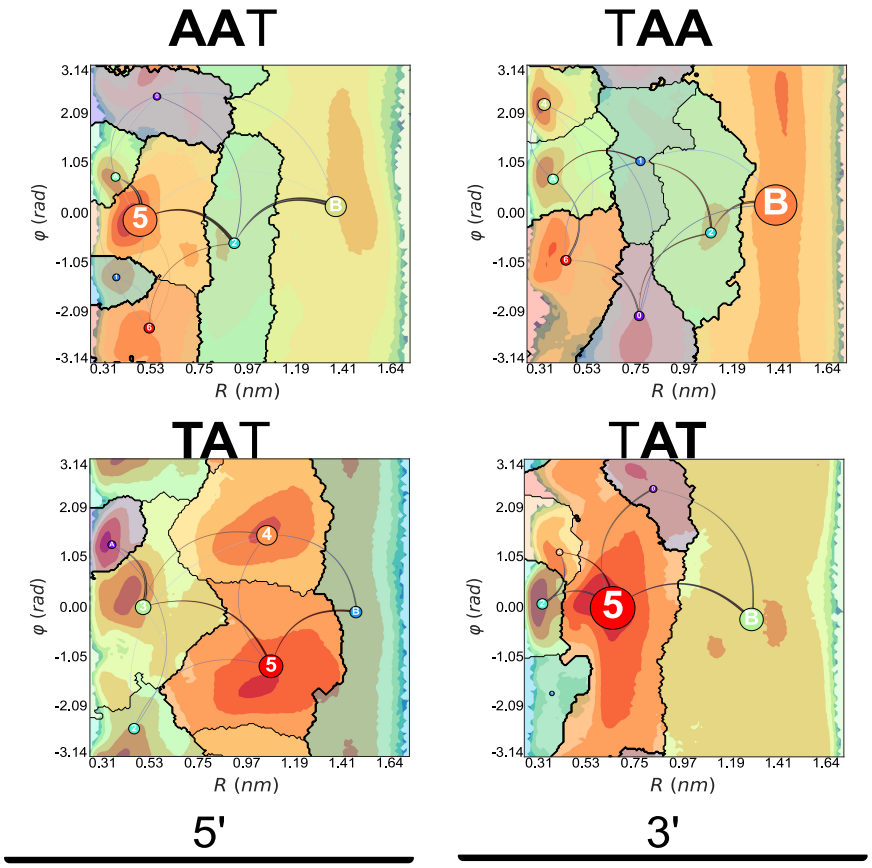

FIGURE 31. The transition pathways predicted from TPT analysis of trinucleotide sequences.

metastable states where transitions could persist in semi-stacked regions and fully lifting bases off one another required step-wise transitions.

In the transition pathways evaluated for trinucleotide sequence, as many of the strongly stacked states were still sampled, it was observed that left-handed transitions through stacked conformations (such as $A \rightarrow 5 \rightarrow 1$ in AAT) still played a significant role in the unstacking mechanism. However, these transitions were no longer the dominant pathway. Instead transitions to weakly stacked states occurred with a higher commitor probability. With the exception of the 3' TAT model, all transitions to weak intermediates followed a left handed transition. Whereas in the dinucleotides right handed transitions had a very low flux probability, in the trinucleotides they were far more likely. One of the possible explanations of this, was that sugar reorientation were not as dominant a feature. As this was the primary source of stacked transitions between right to left handed forms, this could explain the decreased flux of these pathways.

Many of the features observed in the trinucleotides extended into the tetranucleotide sequences. Semi-stacked conformations played a significant role on the transition pathways between right handed to fully unstacked conformations. One noticeable feature of these structures was that as the nucleotides were flanked on both ends, there was very limited base flipping events. The ability to transition between syn and anti conformations only seemed to occur after bases had separated. With increasing strand length, the likelihood of these non-canonical conformations were substantially reduced.

While in the tetranucleotide sequences there were not a significant number of strongly stacked states, molecules still had an increased likelihood to transition through left handed orientations in weakly stacked conformations. The TGGT

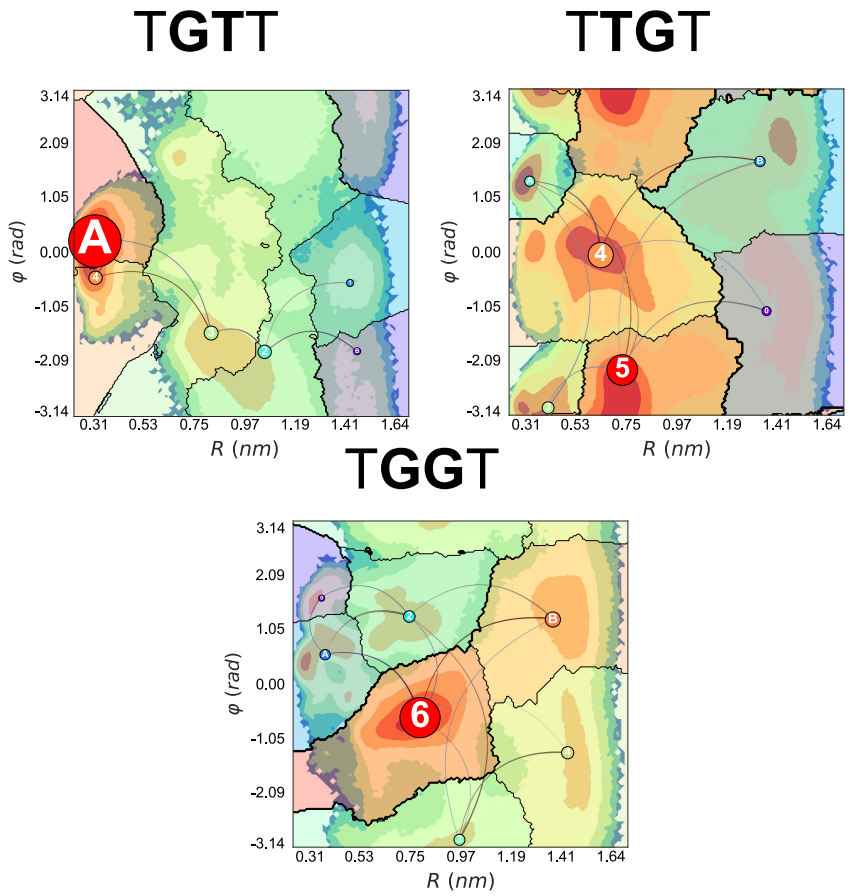

FIGURE 32. The transition pathways predicted from TPT analysis of tetranucleotides.

sequence provides an interesting case of this as states 2 and 6 occupy different chiral orientations. While the transition $A \rightarrow 2$ had a larger flux in, the commitor likelihood through state 6 was larger, and therefore this pathway dominated. Observing the equilibrium populations, state 6 was additionally far more populated at equilibrium.

In all cases, the unstacked regions were far more prominent than in shorter length analogs. Within these states right and left handed flavors emerged (B, 0 in TTGT or B, 4 in TGGT). Continuing with the TGGT case, the equilibrium populations predicted showed a greater chance to occupy right handed unstacked states at infinite times. However, when analyzing the dominant pathway, the molecule first transitioned to a left handed weakly stacked state ($A \rightarrow 6 \rightarrow B$ and $A \rightarrow 6 \rightarrow 3 \rightarrow B$) indicating that chiral directed transitions along the φ parameters were prevalent during unstacking.

Where the tetranucleotide MSM deviated away from shorter length sequences were in the unstacked states. Di and trinucleotides had a single metastable unstacked region, whereas in all tetranucleotides two metastable states emerged. This was indicative that by increasing strand length new barriers were created between rotations of unstacked states.

A dominant result of the tri and tetranucleotide was that several long-lived states emerged at $0.5 < r < 1.2\ nm$ suggesting that deformations in the stacking of DNA became trapped. While simulations could eventually escape due to thermal energy, the timescales were substantial. Additionally, chiral directed mechanisms were still prominent in longer length strands. While the energetic landscape shifted to favor right handed stacked states, and various new unstacked states emerged, the twisting involved in the unstacking mechanism still preferred negative $\Delta\varphi$.

This could provide insights into protein recognition, where when some structural change may occur (bubble formation, base flipping, or some general nucleation event), conformations being trapped may act as a suitable nucleation event or more directly be recognized by proteins.

Conclusion

Kinetic Markov models of long timescale MD simulations provide an elegant means by which to analyze the conformational dynamics of biological systems. Applying them and the GRAD refinement method towards the analysis of DNA base stacking provides new avenues to gain chemical insight on both what the metastable conformations are associated in base stacking, and additionally the timescales associated to local conformational changes in DNA rearrangement. By performing atomistic simulations for all oligonucleotide sequences, the preferred conformations and transition pathways were detailed showing a structural mechanism which indicated a preference for unstacking to occur due to left-handed over right-handed rotations.

In the cases of tri and tetranucleotides, while left-handed pathways still dominated, however as left-handed stacking states were less energetically favorable over the course of the simulation a larger portion of the transition pathways could emerge through right-handed reorientation or could occur by a "kinking" mechanism where bases just lifted off each other without a twisting rearrangement.

While B-form parameters were stabilized in tri and tetranucleotide systems and left-handed stacked characteristics were reduced in the case of non-restricted stacking such as in the dinucleotide case, the nucleic acids have a chiral directed pathway.

By increasing strand length of oligonucleotides, even by only increasing the number of residues by 1 to 2 thymine, illustrated a significant change in the energy barriers that shift the prevalence of DNA stacked conformations to right-handed states. These evaluations illustrate that even small lengthscale contributions may play a large role in the standard B-form conformations and that by better understanding the dynamics at this scale, more insight can be obtained at the macroscopic. The chiral directed mechanisms introduce a new scope in evaluating bubble or fraying dynamics DNA as they transition away from equilibrium and may address new strategies for experimentation to evaluate these results.

CHAPTER VI

DISCUSSION

With the advances in computational tools, both in the simulation capabilities of the MD community and the theoretical advances, the biophysical community has begun to adopt more rigorous studies in the analysis of the dynamics and kinetics of conformational changes. The use of machine learning algorithms to automatically predict the conformational landscapes from data alone, allows for reduced bias in the modeling of conformational states as neither the number of states nor the location of where states are separated are manually selected.

These methods such as Markov state models, and the GRAD method developed within this doctoral work, provide a network description of the biological simulations at varying timescales capturing not only geometric similarities in chemical data (ie how densely populated conformational minima are) but also kinetically by how rapidly conformations interconvert.

While MSM have proved invaluable to the biophysical community, coarse-graining methods have left much to be desired. In particular a crisp depiction of the metastable states provides great chemical insight however have traditionally sacrificed accuracy. In GRAD developed within this doctoral work, a novel method is described where a crisp decomposition is maintained while maximizing the timescales and metastability of the discretized conformational landscape.

Using these techniques, a study is performed on the stacking transition pathways in DNA oligonucleotides. The presented results illustrate a chiral directed mechanism which provide some interesting insight into how thermally induced fluctuations of DNA may play a role in biological roles such as protein recognition

or any function that accesses the internal content of DNA. Future directions for these studies include a further exploring the effect of sequence and polarity on the stacking conformational landscape. Longer timescales as well as additional sequences for longer length DNA would provide critical information.

The structural analysis of DNA oligonucleotides found chiral directed mechanisms in the kinetic pathways between stacked and non-stacked conformers. While dinucleotides had the freedom of several conformational states and likewise the flexibility to transition between them, the complexity of longer length strand found a conformational mechanism which took advantage of the chiral structure of the genetic building blocks. Within these pathways lies many exciting questions. How do these mechanisms play a role in the interaction with DNA protein machinery? How might proteins have evolved to maximize their efficiency in recognizing these thermally induced mechanisms? How much does this effect specific and non-specific interactions? Future studies hope to address these questions and more to evaluate the biological impact of the local stacking dynamics to the macroscopic properties critical in how proteins scan and edit the genetic code of life.

APPENDIX A

MATHEMATICAL SYMBOLS

$\boldsymbol{T}(\tau)$	The transition matrix, a row stochastic conditional probability matrix in $\mathbb{R}^{n\times n}$ with elements T_{ij} define the likelihood of witnessing a transition from states $i \to j$ over a lag time τ.
$\boldsymbol{C}(\tau)$	The count matrix in $\mathbb{Z}^{n\times n}$ with elements C_{ij} define the number of observed transition from discrete states $i \to j$ over a lag time τ.
$\bar{\boldsymbol{C}}$	The symmetrized count matrix in $\mathbb{Z}^{n\times n}$ with elements $\bar{C_{ij}}$ and $\bar{C_{ji}}$ are equivalent. This is typically used to enforce detailed balance.
τ	The model lag time.
$\boldsymbol{p}(t)$	The population vector in $\mathbb{R}^n$ with elements $p_i(t)$ represnting the probability to be in state i at time t.
$\boldsymbol{\pi}$	The stationary population vector in $\mathbb{R}^n$, with elements π_i the probability of of being in state i at equilibrium.
$\boldsymbol{\phi}_i$	The i^{th} left eigenvector of the transition matrix in $\mathbb{R}^n$.
$\boldsymbol{\psi}_i$	The i^{th} right eigenvector of the transition matrix in $\mathbb{R}^n$.
λ_i	The i^{th} eigenvalue of the transition matrix in $\mathbb{R}^n$.
$\boldsymbol{\Omega}$	The continuous state space where all positions are mapped to a discrete decomposition.
$S(t)$	The state in $\boldsymbol{\Omega}$ at time t.
$P(A)$	The marginal probability of observing event A.
$P(A;B)$	The joint probability of observing events A and B.
$P(A\|B)$	The conditional probability of observing event A given even B.

$\mathbb{R}$	The space of real valued numbers.
$\mathbb{Z}$	The space of real valued integers.

APPENDIX B

EXTENDED RESULTS FOR DINUCLEOTIDES MODELS

While all results were analyzed and discussed in chapter V, as the total number of sequences analyzed in this work is extensive, the data is presented here as supplemental material.

Here the free energy surfaces are plotted for all dinucleotides. The energy values for all sequences were scaled to the same range of -1 to 1 to ensure that relative changes in barriers could be compared across all surfaces, and to facilliate the assignment of energy values to the color map. For all graphs the x-axis represents the radial separation between DNA bases, and the y-axis the stacking twist represented by the plane dihedral between residues.

All MSM presented followed the same methodology, although number of microstates and macrostates varied depending on simulation data. All MSM were refined by GRAD and the transition pathways calculated from the updated transition matrix. In all plots, only 95% of the total flux is shown to filter out quick and rare transitions.

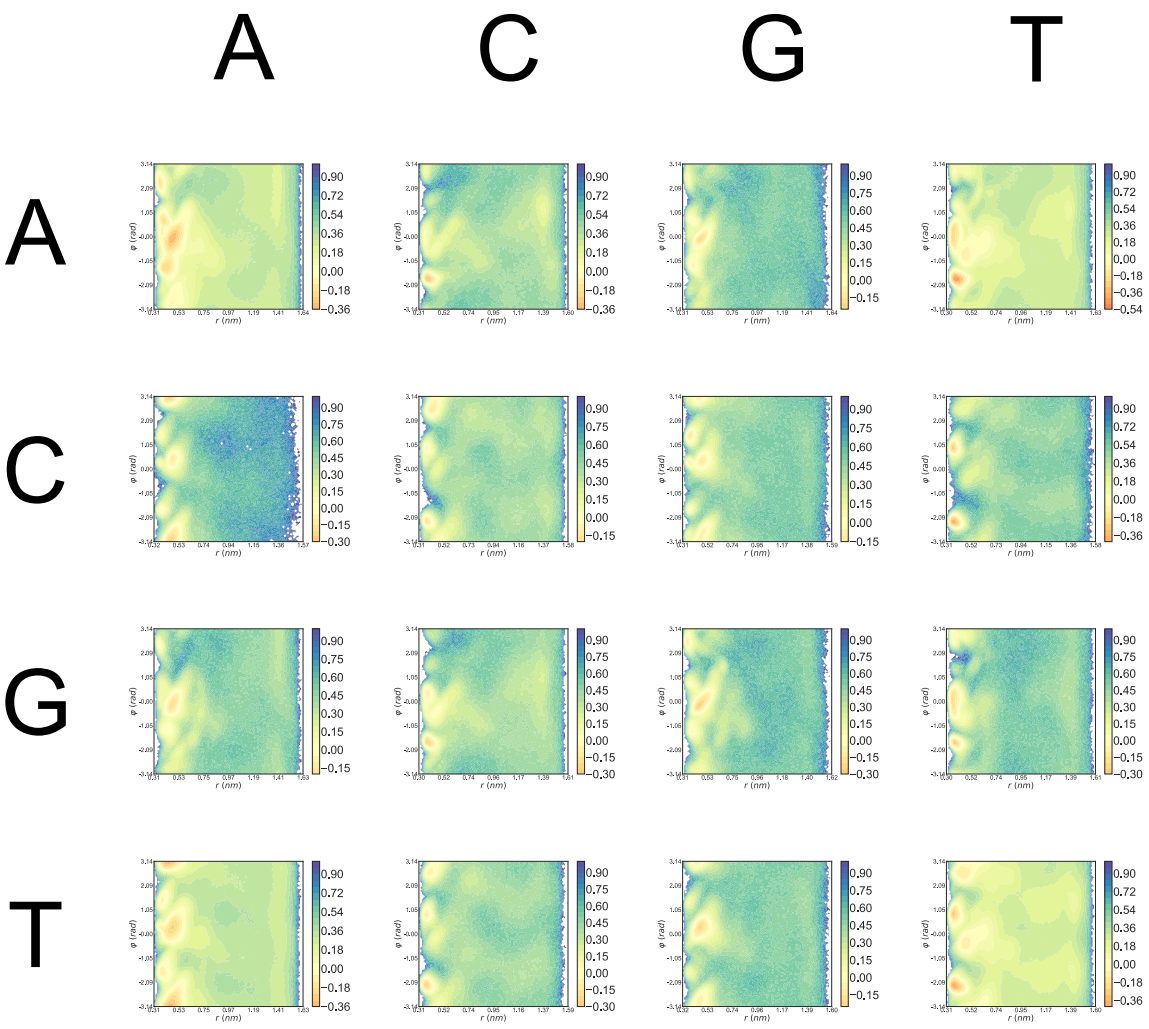

FIGURE 33. The free energy surface of all dinucleotides within the two-site per nucleotide
model. The dinucleotide 5'-AT-3' is given by the landscape in the first row and fourth column.

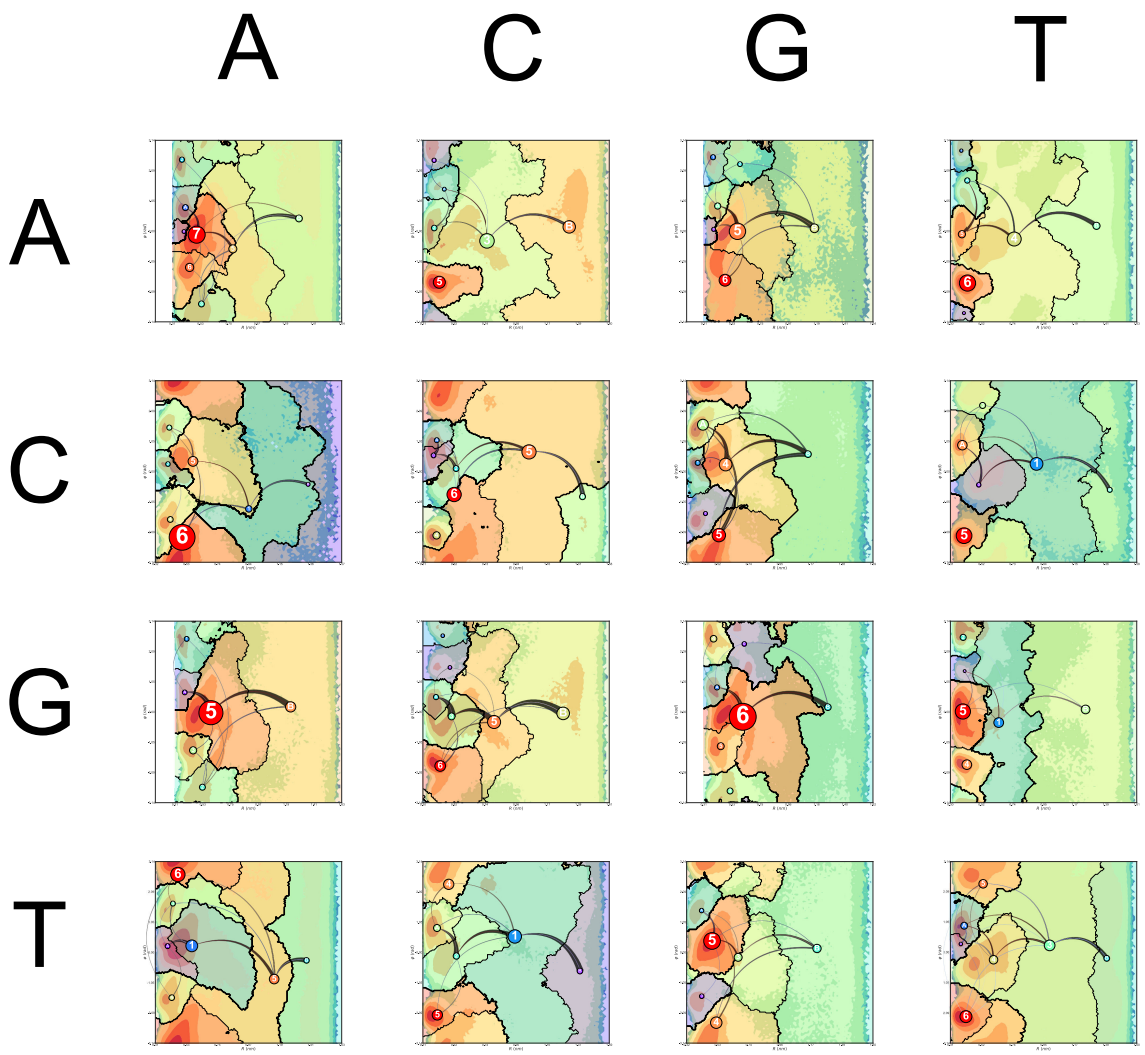

FIGURE 34. The state decomposition is projected along the free energy surface of all dinucleotides. The transition pathways are labeled for 95% of the total flux, where arrow width represents the flux probability, marker size represents equilibrium populations, and colors indicate state assignment.

REFERENCES CITED

[1] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[2] D. Arthur. K-means++: The advantages of careful seeding. In *In Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[3] D. Arthur and S. Vassilvitskii. How slow is the k-means method? In *Proceedings of the twenty-second annual symposium on Computational geometry*, pages 144–153. ACM, 2006.

[4] C. Bergonzo, R. Galindo-Murillo, and T. E. Cheatham. *Molecular Modeling of Nucleic Acid Structure: Electrostatics and Solvation.* John Wiley & Sons, Inc., 2001.

[5] C. Bergonzo, R. Galindo-Murillo, and T. E. Cheatham. *Molecular Modeling of Nucleic Acid Structure: Energy and Sampling.* John Wiley & Sons, Inc., 2001.

[6] V. Bloomfield, D. Crothers, and I. Tinoco. *Physical chemistry of nucleic acids.* HarperCollins Publishers, 1974.

[7] G. Bowman. Improved coarse-graining of markov state models via explicit consideration of statistical uncertainty. *The Journal of Chemical Physics*, 137(13):134111, 2012.

[8] G. Bowman, K. Beauchamp, G. Boxer, and V. Pande. Progress and challenges in the automated construction of markov state models for full protein systems. *The Journal of Chemical Physics*, 131(12):124101, 2009.

[9] G. Bowman, D. Ensign, and V. Pande. Enhanced modeling via network theory: Adaptive sampling of markov state models. *Journal of chemical theory and computation*, 6(3):787, 2010.

[10] G. Bowman, X. Huang, and V. Pande. Using generalized ensemble simulations and markov state models to identify conformational states. *Methods*, 49(2):197–201, 2009.

[11] G. Bowman, V. Pande, and F. Noé. *An introduction to markov state models and their application to long timescale molecular simulation*, volume 797. Springer Science & Business Media, 2013.

[12] G. R. Bowman, V. S. Pande, and F. Noé. *An introduction to Markov state models and their application to long timescale molecular simulation*, volume 797. Springer Science & Business Media, 2013.

[13] J. Bresenham. A linear algorithm for incremental digital display of circular arcs. *Communications of the ACM*, 20(2):100–106, 1977.

[14] R. Bridson. Fast poisson disk sampling in arbitrary dimensions. In *SIGGRAPH sketches*, page 22, 2007.

[15] C. Bush and I. Tinoco. Calculation of the optical rotatory dispersion of dinucleoside phosphates. *Journal of molecular biology*, 23(3):601–614, 1967.

[16] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, 2007.

[17] C. Cantor and P. Schimmel. *Biophysical chemistry. P. 3, The behavior of biological macromolecules*. Freeman, 1980.

[18] T. E. Cheatham II. *Molecular Modeling of Nucleic Acid Structure*. John Wiley & Sons, Inc., 2001.

[19] T. E. Cheatham III and P. A. Kollman. Molecular dynamics simulation of nucleic acids. *Annual Review of Physical Chemistry*, 51(1):435–471, 2000.

[20] J. Chodera and F. Noé. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25:135–144, 2014.

[21] J. Chodera, N. Singhal, V. Pande, K. Dill, and W. Swope. Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *The Journal of Chemical Physics*, 126(15):155101, 2007.

[22] P. Deuflhard and M. Weber. Robust perron cluster analysis in conformation dynamics. *Linear algebra and its applications*, 398:161–184, 2005.

[23] K. Dill and H. Chan. From levinthal to pathways to funnels. *Nature structural biology*, 4(1):10–19, 1997.

[24] N. Djurdjevac, M. Sarich, and C. Schütte. Estimating the eigenvalue error of markov state models. *Multiscale Modeling & Simulation*, 10(1):61–81, 2012.

[25] R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. Lee, et al. Ambertools 16, 2016.

[26] D. Dunbar and G. Humphreys. A spatial data structure for fast poisson-disk sample generation. *ACM Transactions on Graphics (TOG)*, 25(3):503–508, 2006.

[27] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije. Lincs: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[28] B. E. Husic and V. S. Pande. Ward clustering improves cross-validated markov state models of protein folding. *Journal of Chemical Theory and Computation*, 13(3):963–967, 2017.

[29] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[30] T. Lane, G. Bowman, K. Beauchamp, V. Voelz, and V. Pande. Markov state model reveals folding and functional dynamics in ultra-long md trajectories. *Journal of the American Chemical Society*, 133(45):18413, 2011.

[31] Y. Li and Z. Dong. Effect of clustering algorithm on establishing markov state model for molecular dynamics simulations. *Journal of chemical information and modeling*, 56(6):1205–1215, 2016.

[32] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[33] B. McConnell and P. H. von Hippel. Hydrogen exchange as a probe of the dynamic structure of dna: I. general acid-base catalysis. *Journal of molecular biology*, 50(2):297–316, 1970.

[34] R. McGibbon and V. Pande. Variational cross-validation of slow dynamical modes in molecular kinetics. *The Journal of Chemical Physics*, 142(12):124105, 2015.

[35] F. Noé and S. Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Current opinion in structural biology*, 18(2):154–162, 2008.

[36] F. Noé, I. Horenko, C. Schütte, and J. Smith. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *The Journal of Chemical Physics*, 126(15):155102, 2007.

[37] F. Noé and F. Nuske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation*, 11(2):635–655, 2013.

[38] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences*, 106(45):19011–19016, 2009.

[39] F. Noé, H. Wu, J. Prinz, and N. Plattner. Projected and hidden markov models for calculating kinetics and metastable states of complex molecules. *The Journal of Chemical Physics*, 139(18):11B609_1, 2013.

[40] S. Pall, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl. *Tackling exascale software challenges in molecular dynamics simulations with GROMACS*. Springer, 2014.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[42] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco. Refinement of the amber force field for nucleic acids: improving the description of α/γ conformers. *Biophysical journal*, 92(11):3817–3829, 2007.

[43] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé. Identification of slow molecular order parameters for markov model construction. *The Journal of chemical physics*, 139(1):07B604_1, 2013.

[44] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical recipes in fortran 77: the art of scientific computing, 1992.

[45] M. P. Printz and P. H. Von Hippel. Hydrogen exchange studies of dna structure. *Proceedings of the National Academy of Sciences*, 53(2):363–370, 1965.

[46] M. P. Printz and P. H. Von Hippel. Kinetics of hydrogen exchange in deoxyribonucleic acid. ph and salt effects. *Biochemistry*, 7(9):3194–3206, 1968.

[47] J. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134(17):174105, 2011.

[48] L. E. Reichl. *A modern course in statistical physics*. John Wiley & Sons, 2016.

[49] S. Röblitz and M. Weber. Fuzzy spectral clustering by pcca+: application to markov state models and data classification. *Advances in Data Analysis and Classification*, 7(2):147–179, 2013.

[50] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[51] A. Savitzky and M. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[52] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, 11:5525–5542, Oct. 2015.

[53] C. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. *Handbook of numerical analysis*, 10:699–744, 2003.

[54] C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden. Markov state models based on milestoning. *The Journal of Chemical Physics*, 134(20):05B609, 2011.

[55] C. Schütte and M. Sarich. A critical appraisal of markov state models. *The European Physical Journal Special Topics*, 12(224):2445–2462, 2015.

[56] C. R. Schwantes, R. T. McGibbon, and V. S. Pande. Perspective: Markov models for long-timescale biomolecular dynamics. *The Journal of chemical physics*, 141(9):09B201_1, 2014.

[57] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. Estimation and uncertainty of reversible markov models. *The Journal of Chemical Physics*, 143(17):174101, 2015.

[58] N. van Kampen. Stochastic processes in physics and chemistry, 1995.

[59] E. Vanden-Eijnden. Transition path theory. *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*, pages 453–493, 2006.

[60] P. H. von Hippel, N. P. Johnson, and A. H. Marcus. Fifty years of dna breathing: reflections on old and new approaches. *Biopolymers*, 99(12):923–954, 2013.

[61] D. Wales, M. Miller, and T. Walsh. Archetypal energy landscapes. *Nature*, 394(6695):758–760, 1998.

[62] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[63] J. D. Watson, F. H. Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[64] E. Weinan and E. Vanden-Eijnden. Towards a theory of transition paths. *Journal of statistical physics*, 123(3), 2006.

[65] E. Weinan and E. Vanden-Eijnden. Transition path theory and path-finding algorithms for the study of rare events. *Annual review of physical chemistry*, 61, 2010.

[66] Y. Yao, R. Z. Cui, G. R. Bowman, D.-A. Silva, J. Sun, and X. Huang. Hierarchical nyström methods for constructing markov state models for conformational dynamics. *The Journal of chemical physics*, 138(17):05B602_1, 2013.

[67] Y. Yao, J. Sun, X. Huang, G. Bowman, G. Singh, M. Lesnick, L. Guibas, V. Pande, and G. Carlsson. Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of Chemical Physics*, 130(14):04B614, 2009.